Spatial Model Training Program

Version 1.0

2020/10/31

Contents

Version Control

Version	Date	Description
0.1	2020/05/08	The first version, with Unit A (quiz, a glossary of terms and bibliography missing)
1.0	2020/10/31	First Version

Table 1: Version Control

Terms and abbreviations

Table 2: Terms and Abbreviations

Abbreviation/ Term	Meaning
CL	Chlorine, either Free or Total, depending on station. Typically measured in mg/L
TU	Turbidity measurement. Typically in NTU units
REDOX/ORP	Oxidation-Reduction Potential. Typically in mV units.
COND	Conductivity. Typically in $\mu S/cm$ units.
FL	Flow. Typically in m^3/S
Water quality measurements	Chemical and Physical measurements for water quality. Examples are the above measurements
Filtering	Removing irregular, extremely noisy or non-representative observations from the data.
SM	Spatial Model
Time series, Time Series Analysis	A series of measurements taken across different times. Time series analysis – the statistical practice of analyzing time series to gain insights.
EDS Detectors	A statistical module in the Detector EDS (Decision Makers' product) designed to detect a specific type of effect or abnormality in the data.
EDS	Event Detection System.
Correlation	A summary statistic quantifying how much does a change in one numeric measurement explain the change in a second numeric measurement.
Autocorrelation function (ACF)	A function quantifying how much is a time series of measurement correlated with a time-shifted version of itself. The input argument is the size of the time shift. The computed value is the quantity of auto-correlation, or self-correlation. See section 2.1.1.
Cross-correlation function (CCF)	A function quantifying how much is a time series of measurement correlated with a second time series, after inducing a time-shift to the second series. The input argument is the size of the time shift. The computed value is the quantity of cross-correlation, or correlation between series after a time shift in one of them. See section 2.1.2.
Robust correlation measures	A correlation measures whose value is relatively stable even if noisy measurements are found in the data.
Bootstrap	A statistical resampling procedure, where a set of samples are chosen from the original data. The
Prediction models	A statistical model predicting the value of a numeric variable (in our context), based on the value of another numeric value. The first variable is referred to as the dependent variable, and the latter as the independent variable. See Section 2.5 for discussion of prediction models.
Constant Shift model	A type of linear model between X and Y, of the form $Y = X + b$, explained in subsection 2.5
Linear Model	A type of linear model between X and Y, of the form $Y = aX + b$, explained in subsection 2.5
Kernel regression Model	A type of non-linear prediction model, explained in subsection 2.5. The model has no closed formula form, and is data adaptive.
Pair	SM model predicting the measured value of a station downstream based on a station upstream.

Triplet	SM model comparing a triplet of station for detecting abnormal combinations.
Flow time	The time required for a drop of water to travel between two points in the network.
Estimated flow time	An estimation for the real flow time, computed using the change in water quality measurements in the two stations. See Algorithm in Section 3.2
Flow regime	A set range of times, in certain days or hours, where the flow time is approximately fixed.
Flow-time adjustment/ Lag adjustment	**(Relevant when referring to water quality measurements in both a source and target station)** We explain using an example: Assume water takes 3 hours to flow between a source station and a target station. The flow-time adjusted measurement associated with the 8PM measurement at the target station, is the 5PM measurement at the source station. See Section 3.5.1. for a complete definition.
Window of measurements	In a time series of measurement, a subset of the series measured across some time interval, say , a window of three hours.
Score function / Objective function	A function giving for two time series (perhaps in a specific flow regime), and a flow-time value u, a score of well does a flow-time of u explain the correlation between the two time series.
Flow regime selection	A statistical analysis process in which different times of the day, and days of the week, are classified into different flow regimes based on auxiliary measurements such as flow meters, or prior knowledge about the operational nature of the water distribution network.
Prediction, Predicted Value	The statistical method of generating a prediction for numeric value (in our context), based on explanatory variables.
Error distribution	The empirical distribution of differences between measurements and the predicted values, i.e. the distribution of prediction errors.
Confidence Interval for Prediction/ CI	A range of values, around the predicted value, where the actual measurement will likely be found, with high probability.
Triplet types	Three possible types: Line, Fork-in, Fork-out triplets.
Spatial Model alarm types	The types of SM alarms are: Limits violation, Poor limits, Poor prediction, No data, Triplet Alarms. See Through discussion the at start of Unit C.
Network states	A water distribution network can be found in 1 of 4 states: State E: Water supplied from an external source State R: Water supplied from reservoirs State E→R: Network is transitioning from state E to R. State R→E: Network is transitioning from state R to E.
EDS alarms	Alarms raised by the EDS, due to one of its detectors. Operators are notified regarding alarms. One of the alarm generating detectors is a detector connected to SM events, see Section 4.7 for additional details.
Reports	HTML reports on SM statistical models, and historical alarms. See Section 4.12 for server-side reports, and Section 4.11 for reports on the client-side software.
Quizzes	Every one of the three units in this training program ends with a short quiz.

1. Introduction

This document serves as a training program for the Spatial Model (SM) module of the *Detector* EDS (Event Detection System) developed by Decision Makers. The training program is aimed for users operating the Detector software, together with the SM module, but it may also be useful to water utility personnel operating other SCADA/monitoring systems and seeking to improve their event detection capabilities. This document is also meant to be used as a hand-out field-manual, in a 1-2 day workshop introducing the SM.

This document is self-contained and can be read front to back without previous knowledge of statistics. We attempt to bring readers up-to-speed on all the statistical and data-analysis techniques needed in order to understand how the SM is able to estimate flow time between network locations and monitor water quality data for abnormal events.

The structure of the document is as follows. The training program is comprised of three units:

- Unit A – "Statistical Prerequisites"
- Unit B - "Algorithmic and Statistical Methodology"
- Unit C- "The Spatial Model Software"

Unit A discussed several background subjects relevant to the SM methodology, such as typical errors in data, measurement error filtration, Auto-correlation and Cross-correlation, Non-Parametric measures of association and prediction models.

Unit B discusses in-depth the algorithms used by the spatial model. The unit begins with a "bird's eye view" of the SM, and the problems it is meant to address. Specifically, we provide motivation for comparing pairs and triplets of sensors across a network in order to pinpoint the location of water quality events better. The unit introduces the SM score function, a non-parametric equivalent of the cross-cross correlation aimed at estimating the time required for water to flow between network measurement locations (henceforth known as "flow time", "lag time" or "delay time"). Afterwards, the unit proceeds to describe how to make use of the flow time estimate in order to train prediction models for water quality at different network locations. Models for pairs and triplets of sensors are described.

Unit C introduces the SM software module of the EDS. The unit reviews the SM module screen-by-screen and describes: how the water quality can be monitored using the SM; how the SM connects to the exiting EDS alarm management system; how models for pairs or triplets can be built; and how their performance can be reviewed. The different methods and algorithms presented in unit B are implemented in the software presented in Unit C. Therefore, at the start of each lesson in unit C, we describe which lessons in Unit B are most relevant to the current discussed screen/model.

Each lesson in Units A and B introduces a few concepts and should take no more than 10-15 minutes to go over. After a short intro, several examples based on real data would be given, usually in the form of graphs and figures.

Each unit concludes with a short quiz, 30 minutes in length, aimed at testing if the main concepts of the Unit were understood. Quizzes are usually comprised of 15-20 questions, most of which are multiple-choice. Answers to the quizzes are given after each quiz.

2. Unit A - Statistical Prerequisites

2.1 Introduction to Time Series Analysis

A time series is a series of repeated measurements from the same source of data, e.g., a water quality measurement device, across time, usually at fixed and predetermined intervals. The statistical study of time series was developed significantly in the 1980s, in order to provide analysis methods for economical, meteorological, and physical science datasets. Unit A will focus on preliminary concepts in the analysis of time-series data, that will be relevant the theory and application of the Spatial Model (SM) module of the Detector EDS.

Time series analysis methods include both descriptive statistics of time series data, along with advanced modeling methodologies. The key assumption in analyzing this data is that data points are measured over fixed time intervals. Many times, this assumption is not met by the data. For example, for water quality measurements, e.g., turbidity and conductivity, data is collected at varying time resolutions at a different location at the network. Moreover, field analyzers may have some jitter in their measurement times: while measurements may ideally be taken every minute or 10 minutes, some measurements may be unavailable or may be delayed due to technical difficulties at field analyzers. Problems that may include this type of jitter include faulty measurements that are discarded or redone or temporary lack of communications. The next example will demonstrate differences in time resolutions across different sampling locations in a water distribution network.

Example 2.1A – varying time resolution: (Figure 1 shows a time series of conductivity measurements, across a month of data. Measurements are taken every minute (60 seconds), forming 1440 data points per day. A key property of water quality time series data is observed in the graph: consecutive

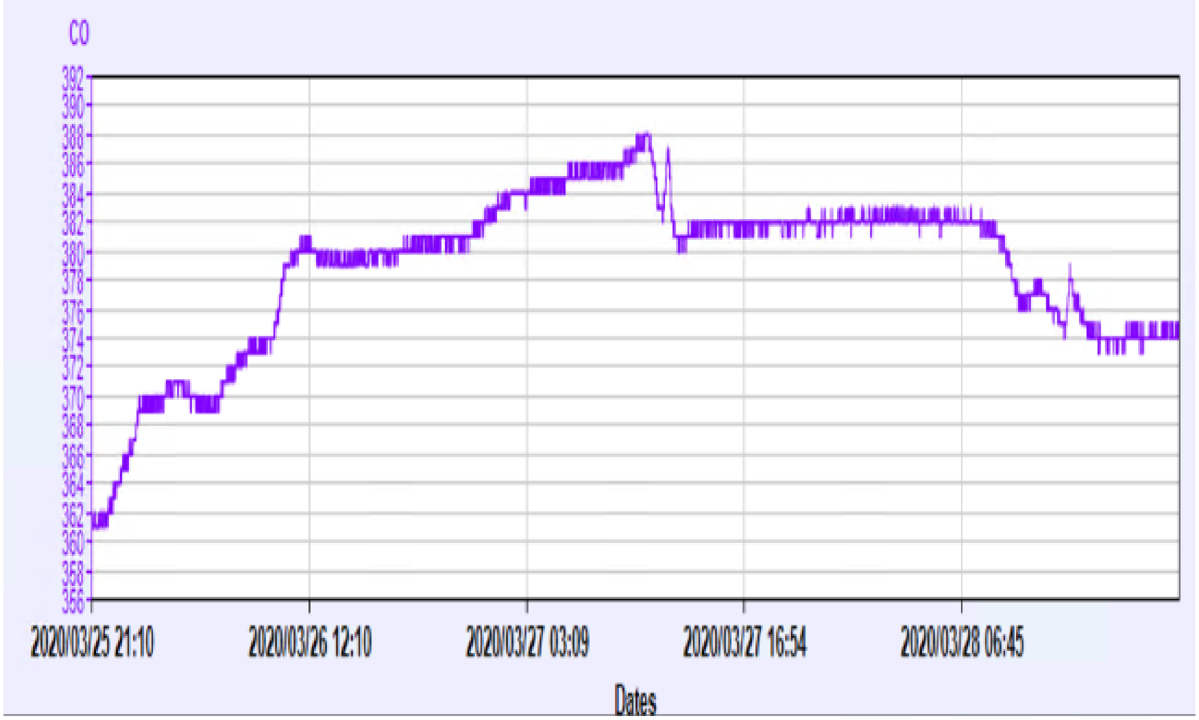


Figure 1: Example for varying time resolution - conductivity measured at upstream station at 1 minute intervals,

measurements are highly correlated. This feature of water quality time series data is called auto-correlation ("the correlation of a series with itself") and will be discussed shortly. Water quality time series data is expected to be "smoother" across time as it is measured deeper, or downstream, into the water distribution network. The larger the volume of water producing the measurements, the more time and water volume were available to attenuate changes in the water quality originating at the water source for the water distribution network. In fact, abrupt changes in water quality that cannot be explained, e.g., are not maintenance work, may be indicative of a water quality event.

Figure 2 shows time series data for conductivity from a nearby station in the network, for the same dates. Characteristic water flow in this network involves water measured in the station depicted in Figure 2, typically 2-3 hours after being measured in the station depicted in Figure 1. This can be verified by the same "curve" of conductivity measurements appearing almost identically at the two stations. However, there is a key difference between the two water quality analyzers in Figures 1 and 2: while the analyzer in Figure 1 measures conductivity once every minute, the analyzer in Figure 2 measures water quality only once every 5 minutes. The reason for the difference is that the first analyzer is connected to power in a water monitoring facility, and therefore can gather measurements continuously. The second series of measurements were collected by a battery-powered water quality analyzer. This form of low energy measurement device collects information once every 5 minutes and transmits collected data once every two hours. Lower sampling rates allow for longer battery life (possibly spanning years), and for measurements to be gathered at locations were power or facilities are not easily available.

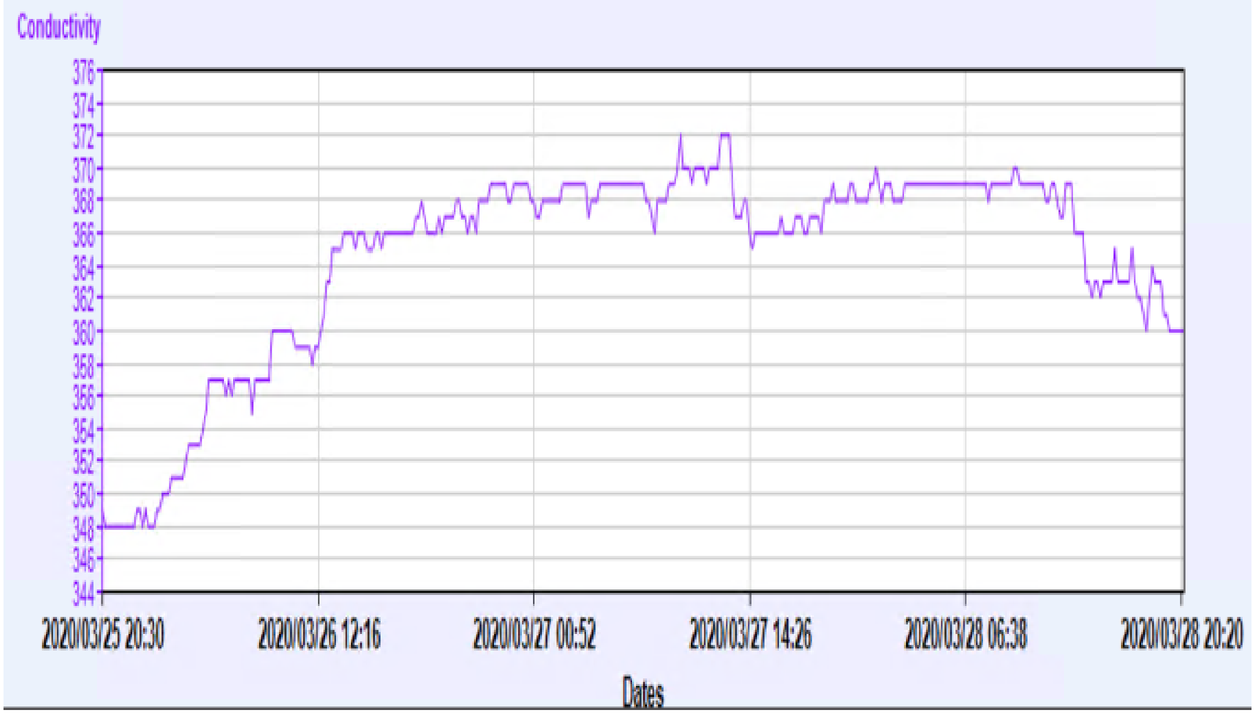


Figure 2: Example for varying time resolution - conductivity measured at downstream station at 10 minute intervals,

While the data series in Figure 1 contains more data points, it does not contain additional information on the underlying water quality. Due to high autocorrelation ("smoothness") of conductivity data, a single measurement once every 10 minutes is sufficient in order to describe water conductivity up to a change in 1-2 uS/cm.

Analysis of time series data usually involves treating data as a series of equally spaced points in time. Moreover, for different time series to be compared, the temporal sampling resolution of different series needs to be identical. With equal temporal resolution, if the first measurements of two series are aligned in time, so are all succeeding measurements. Therefore, a mandatory preprocessing step for time series analysis is to coerce data to a fixed grid of measurements in time.

When working with time-series data in the SM model, one method for coercing time-series data to predetermined, fixed resolution grids of measurements would be to consider the **median value in fixed time intervals bins**. For example, the two data series shown in Figures 1 and 2 could be preprocessed to two series of measurements, with median taken over 30 minutes intervals, starting from 2020/03/25 21:30. After aggregation by medians, the two time-series could be written as :

$$X_1, X_2, X_3, \dots, X_n$$

$$Y_1, Y_2, Y_3, \dots, Y_n,$$

with different X's and Y's corresponding to the median values of conductivity in 30-minute intervals in the first and second time-series, respectively. X_1 would correspond to the median value of conductivity measured in station 1, with the median taken across measurements from 2020/03/25 21:30 to 2020/03/25 21:59 (30 measurements total). X_2 would correspond to the median value of conductivity measured in station 1 across measurements from 2020/03/25 22:00 to 2020/03/25 21:29. Y_1 would be computed by taking the median of all conductivity measurements between 2020/03/25 21:30 to 2020/03/25 21:59 (5 measurements total, since station 2 measures once every 5 minutes). The slots in the temporal grid collecting measurements will be referred to as **bins**.

This method has the benefits of coercing two data series to the same temporal resolution while avoiding jitter in measurement times. A favorable side effect of this method is that extremely noisy measurements ("spikes") in the data, unrepresentative of the actual water quality, are filtered out from the data if adjacent measurements in time are without extreme noise (for example, given five values, one of which is an outlier in measurement, the median value of the five measurements is not affected by the outlier). Types of errors and noise in data, ways to counter them, will be the topic of Section 2.2.

The above method is fully valid when analyzing historical data. However, when analyzing on-line data, data becomes valid as time progresses. Consider a case where we wish to assess the conductivity value in station 2, as of 21:04. Taking the median value of all measurements collected between 21:00 and 21:30 as of 21:04 would result in taking the median of a **single measurement** collected at 21:00. The next measurement for this bin in the time grid will be available (hopefully) only at 21:05. Since this single measurement may be faulty, the representative median value may still be unreliable. A solution to this problem is to partition measurements into 15-minute bins in time (instead of 30-minute bins), and take the representative values in each time point to be the **median over the two most recent time interval bins**. The key idea is that while a single bin might contain only a single measurement (making the median vulnerable to noisy data points), two bins will always contain more than a single data point. For example, using this approach, the representative value at 21:04 will be computed by taking the median value of the measurements collected at 20:45,20:50,20:55, and 21:00.

Whenever computing a median value for an even number of values, The median will be taken to be the average of the two most central measurements. If we ever encounter a setting where we need to compute the median of 2 measurements, the median will simply be the average of the two measurements.

2.1.1 Autocorrelation

Previously, we have observed water quality time-series data to be relatively smooth, with measurements in high proximity in time also being in high proximity in value. Without delving into definitions, we called such time series highly autocorrelated. In this subsection, we properly define the autocorrelation of a time-series and present several properties of this autocorrelation with respect to time series data.

Before we define autocorrelation, we recall the definition of the Pearson correlation coefficient. Section 2.3 will discuss several types of correlation in-depth, but for now, recalling the Pearson correlation will suffice.

Consider two vectors of observations:

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_n$$

It is assumed that X_1 and Y_1 were measured jointly, X_2 and Y_2 were measured jointly, and so on. For example, X and Y measurements could represent pH and Conductivity measurements taken at the same station, jointly, at 8 am, with different pairs of measurements (corresponding to index subscripts) representing different days. A possible statistical question of interest is how values of X and Y vary together: When X increases, is it likely to see Y increase/ decrease/ remain constant? Correlation analysis quantifies the strength of the statistical relationship between two quantities, both in terms of directionality (increase/ decrease) and magnitude (a slight increase compared with a large increase).

In terms of notations, the notations $X_{1:n}$ and $Y_{1:n}$ will be used to denote the two series of measurements, with n measurements in each series. When referring to a sub-series (part of a series) of measurements, we will use the notation, $X_{start:end}$, with $start$ and end denoting two indices, to denote the sub-series

$$X_{start}, X_{start+1}, \dots, X_{end-1}, X_{end},$$

a sub-series including $end - start + 1$ values total. Let $mean(X_{start:end})$ and $sd(X_{start:end})$ denote the mean and standard deviation of a series of measurements, respectively. The sample Pearson correlation coefficient between two series of joint measurements is defined to be:

$$r(X_{1:n}, Y_{1:n}) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - mean(X_{1:n})}{sd(X_{1:n})} \right) \left(\frac{Y_i - mean(Y_{1:n})}{sd(Y_{1:n})} \right).$$

This value can be thought of as the mean product of time series values, with each series normalized to have a mean value of 0 and a standard deviation of 1.

The Pearson correlation coefficient has value within the range -1 to 1 (including -1 and 1), with negative values describing a negative association (when X is smaller than it's the mean value, Y tends to ba larger than it's mean value) and positive values describing a positive association (when X is smaller than it's the mean value, Y tens to be smaller than it's mean value as well). A similar value called the **sample covariance** is given by:

$$cov(X_{1:n}, Y_{1:n}) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - mean(X_{1:n})\big)\big(Y_i - mean(Y_{1:n})\big).$$

This value is not constrained to the range $[-1,1]$, since the two series of measurements were not normalized.

We proceed to describe the sample autocorrelation function of a series of measurements. The autocorrelation function of a series of measurements ($X_{1:n}$), is given by:

$$\begin{aligned} ACF(X_{1:n}, \delta) &= r\big(X_{(\delta+1):n}, X_{1:(n-\delta)}\big) \\ &= \frac{1}{n-\delta}\sum_{i=1}^{n-\delta-1}\left(\frac{X_{\delta+i} - mean\big(X_{(\delta+1):n}\big)}{sd(X_{(\delta+1):n})}\right)\left(\frac{X_i - mean\big(X_{1:(n-\delta)}\big)}{sd(X_{1:(n-\delta)})}\right). \end{aligned}$$

For each value of δ, the value of $ACF(X_{1:n}, \delta)$ gives the Pearson correlation value between $X_{1:n}$ with the first δ measurements removed and $X_{1:n}$ with the last δ measurements removed. Intuitively, $ACF(X_{1:n}, \delta)$ gives the correlation of $X's$ with a lagged version of the X's, with the lag being δ time units behind.

The value of $ACF(X_{1:n}, 0)$ is always 1 since the correlation of the data with itself is always 1. As delta increases, the value of $ACF(X_{1:n}, \delta)$ tends to decrease. If the series of X's is relatively smooth across changes in time, the decrease of $ACF(X_{1:n}, \delta)$ across values of δ is slow, since the series X looks almost the same, after shifting it δ measurements in time. The following example demonstrates the relationship between time series smoothness and how the ACF changes by δ.

Example 2.1B – Time-series smoothness and the ACF: Figures 3 and 4 show Total Chlorine as measured across a month in two sampling locations across a water distribution network. The sampling location depicted in Figure 3 is placed a few hours of water flow away from the main entry point of water to the city. The city is supplied with chlorinated water from an intercity water distribution system. Figure 4 shows total chlorine at another sampling location, however, this sampling location is in a different DMA, from the location depicted in Figure 3. All water passing into the DMA must pass through a large reservoir. We can observe changes in chlorination levels across the graphed time period, as chlorination levels in Figure 3 vary in a daily manner. These changes in chlorination are unobserved in Figure 4, since chlorination is almost constant for water coming from the reservoir. The time series observed in Figure 4 is smoother than the one observed in Figure 3 and has less changes across time. After Figures 3 and 4, we continue the discussion, and show how this form of "smoothness" is translated into the change rate of the ACF across δ.

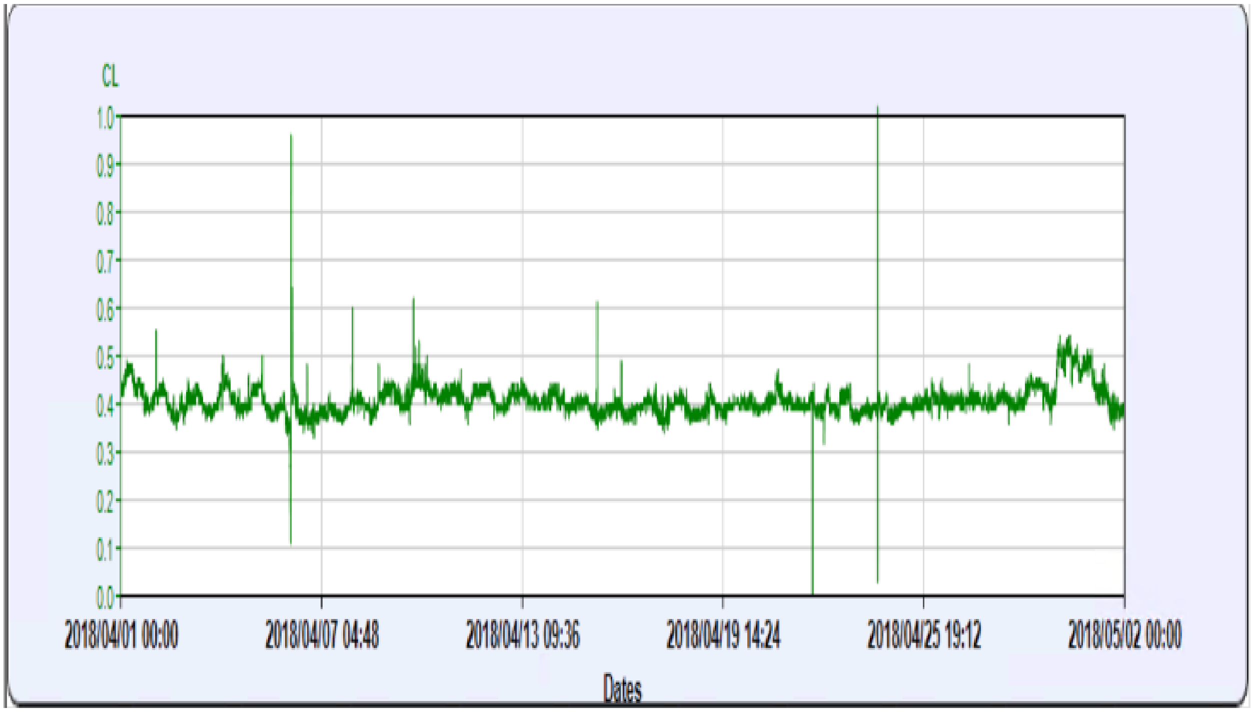


Figure 3: Total CL measured during a month, in data analyzer connecter directly to an intercity water source.

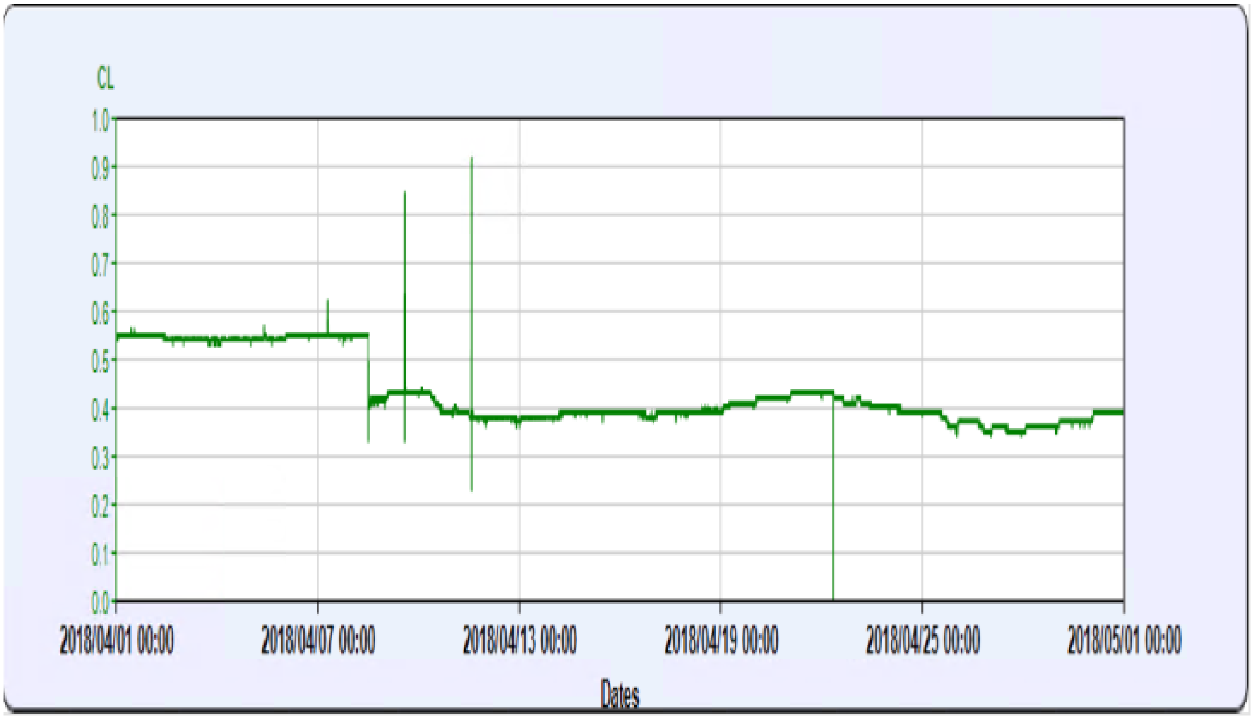


*Figure 4:Total CL measured during a month, in data analyzer downstream from a large water reservoir.*The water reservoir itself is filled from the same source as the sampling station in Figure 3. Note the drop at 2018/04/08 in baseline values due to sensor calibration.

Figure 5 shows the ACFs for the total CL time series in the two sampling locations depicted in Figures 3 and 4. The ACFs were evaluated over three months of data, starting from 2018/04/10 (removing the calibration event), with data aggregated to hourly median values. Observations larger than 0.5 mg/L or smaller than 0.345 mg/L were removed prior to analysis. As expected, both ACFs have a value of 1 for a lag (δ) of 0 hours. As the lag increases, the ACFs decrease in value; however, their decrease occurs at different rates. We observe the ACF for the second sampling station (Figure 4) in **red** to decrease at a slower rate, than the **black** ACF, describing the first sampling station (Figure 3). The slower decay of the ACF indicates a higher correlation of the data time series compared with time-dilated versions of itself. This results from the higher smoothness of the time series.

Interestingly, the **black** ACF reaches negative values, indicating negative autocorrelation for some lag values. Negative ACF values are frequent in periodic time series data, were certain lags (δs) may cause the maxima and minima of the data, and its respective time dilated version to overlap.
On the next page, we discuss why it was crucial to remove the calibration event from the data before computing the ACF.

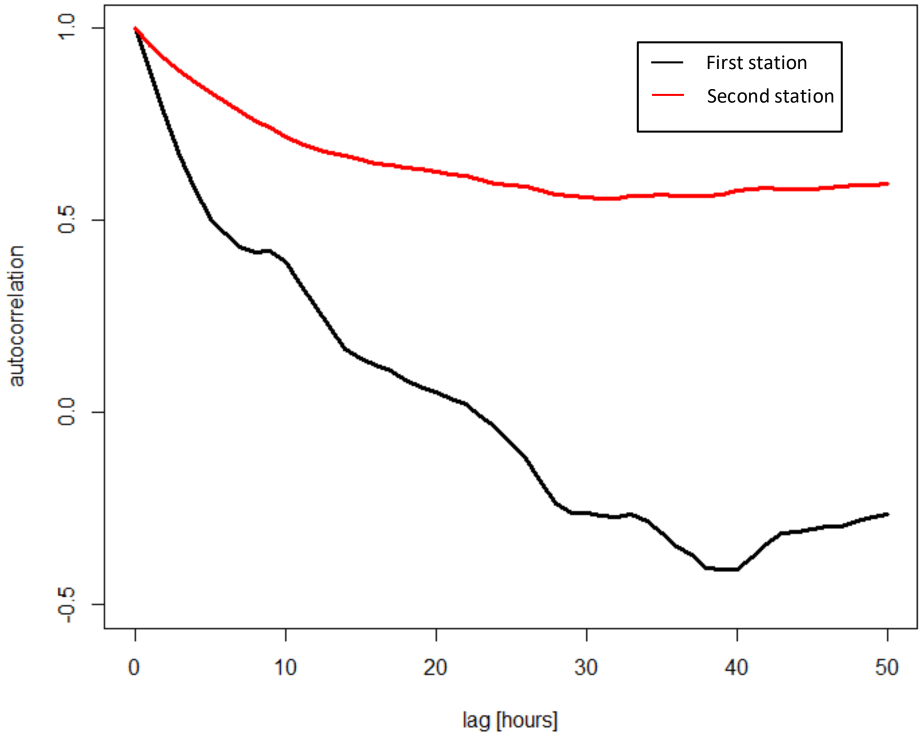


Figure 5: ACFs functions for total CL time series in the sampling locations shown in Figures 3 and 4. ACFs were computed for three months of data, starting from 2018/04/10 (removing the calibration event). Data was aggregated to hourly median values across bins before computing ACFs. Observations larger than 0.5 mg/L or smaller than 0.345 mg/L were removed prior to analysis.

Figure 6 shows the ACFs for analysis similar to the one presented in Figure 5, with one crucial difference: analysis was performed for data starting from 2018/04/01 and up to 2020/04/14, so the calibration event of 2018/04/08 in the second time series was included. With the calibration event included, we observe a peculiar behavior: despite the second station having locally (in time) smoother measurements, it's ACF not descends substantially faster and reaches and ACF value of 0 at $\delta = 50\ hours$. The reason for the for this linear descent[1] in the **red** graph originates from the calibration event at 2018/04/02. Having a calibration event in the data changes the "data series mean" substantially for different values of δ. Since the baseline measurement value changes, the subtraction of data means (in the ACF definition) makes the ACF incomparable for different values of δ. Section 3.2 in Unit B will discuss how calibrations are accounted for in the SM.

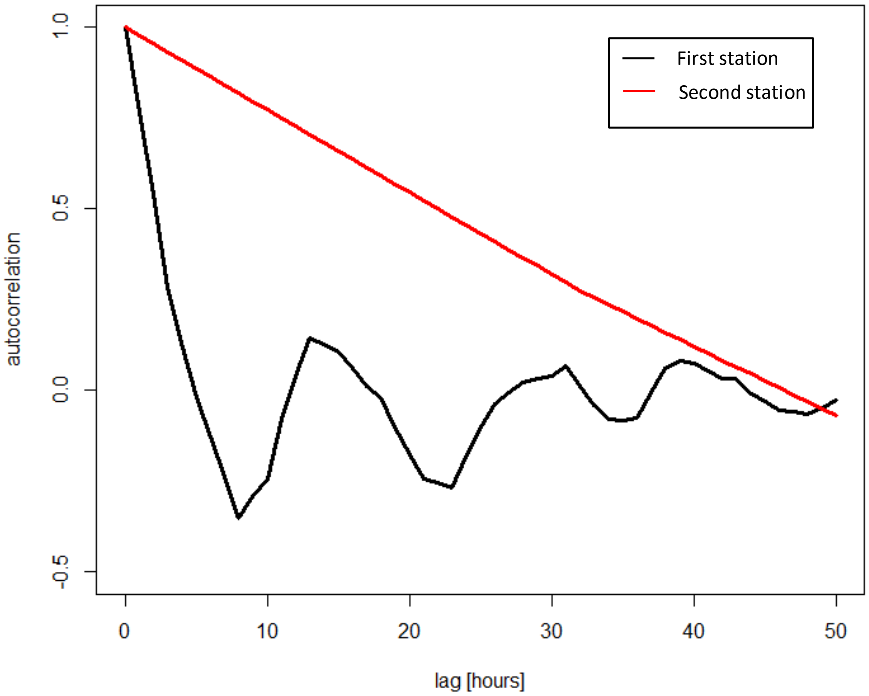


Figure 6: ACFs similar to the ones presented in Figure 5 (see figure description for details); however, the analysis was performed for data starting from 2020/04/01 and up to 2020/04/14, and with the calibration event in the second station included.

[1] To those recalling functional analysis, the ACFs mimics the convolution transform of a step function with another (shifted) step function.

2.1.2 Cross-correlation

We proceed to define the cross-correlation function (CCF). Suppose we observe to series of measurements:

$$X_1, X_2, \dots, X_n,$$

$$Y_1, Y_2, \dots, Y_n,$$

then the cross-correlation function for a positive lag, $\delta > 0$, is defined to be:

$$\begin{aligned} CCF(X_{1:n}, Y_{1:n}, \delta) &= r\left(X_{(\delta+1):n}, Y_{1:(n-\delta)}\right) \\ &= \frac{1}{n-\delta} \sum_{i=1}^{n-\delta-1} \left(\frac{X_{\delta+i} - mean\left(X_{(\delta+1):n}\right)}{sd(X_{(\delta+1):n})}\right)\left(\frac{Y_i - mean\left(Y_{1:(n-\delta)}\right)}{sd(Y_{1:(n-\delta)})}\right). \end{aligned}$$

For a negative lag, $\delta < 0$, we can compute the CCF value using $CCF(Y_{1:n}, X_{1:n}, -\delta)$. If the ACF, at some value δ, describe the correlation of the X's with a time dilated (δ units in time) version of themselves, then the CCF at value δ describes the correlation between X's, and a time dilated version of the $Y's$, for some value of δ. The next example will demonstrate why the CCF is useful in our analysis of water quality data.

Similar to the autocovariance function defined in Section 2.1.1, one may also define the cross-covariance function by avoiding the division by $sd(X_{(\delta+1):n})$ and $sd(Y_{1:(n-\delta)})$ in the above expression.

Example 2.1B – The CCF and it relation to spatial water quality patterns in the water distribution network: Figures 7 and 8, describe two time series of conductivity measurements taken at a Source and Target station, respectively. The data displayed in the graphs is from Match-April 2017. Water flows in this network from the Source station to the target station. We can see the changing pattern of conductivity is similar in both stations. Looking at Figure 9, showing a scatterplot of median hourly values across a 45 day period, we see the measurements in both stations are highly correlated. We continue discussing the CCF for this data after Figure 9.

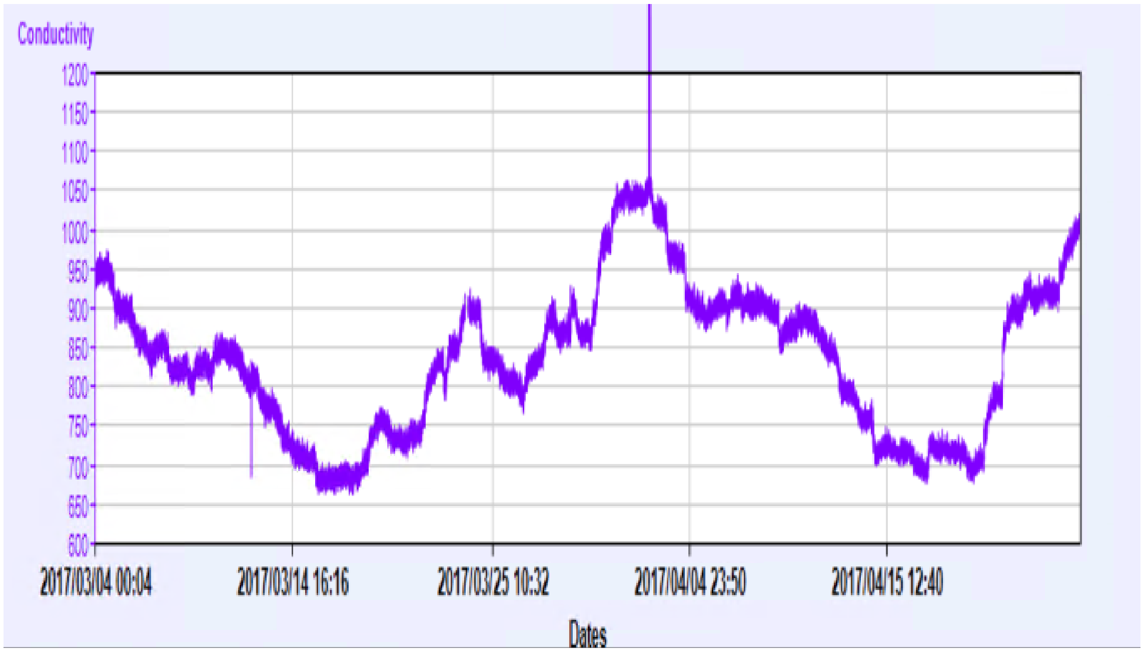


Figure 7: Conductivity measurements for a time period of 40 days in a Source station. Water from this station flows to the station described in Figure 8 (the Target station).

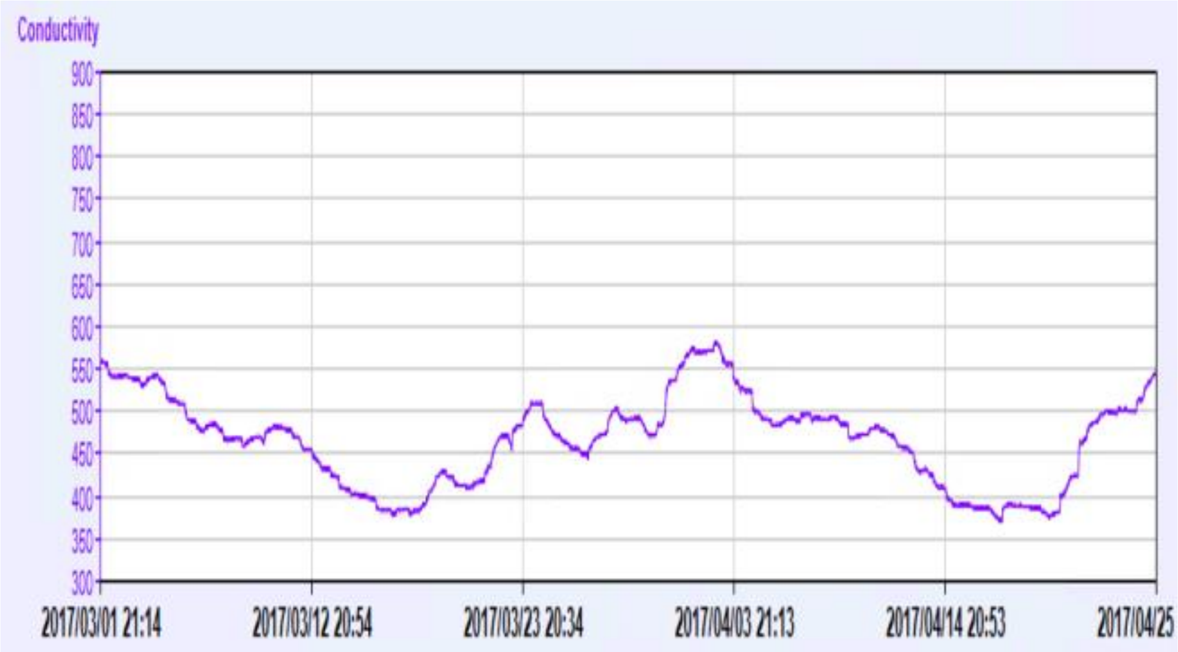


Figure 8: Conductivity measurements for a 55 day period in the Target station, receiving water from the Source station depicted in Figure 7.

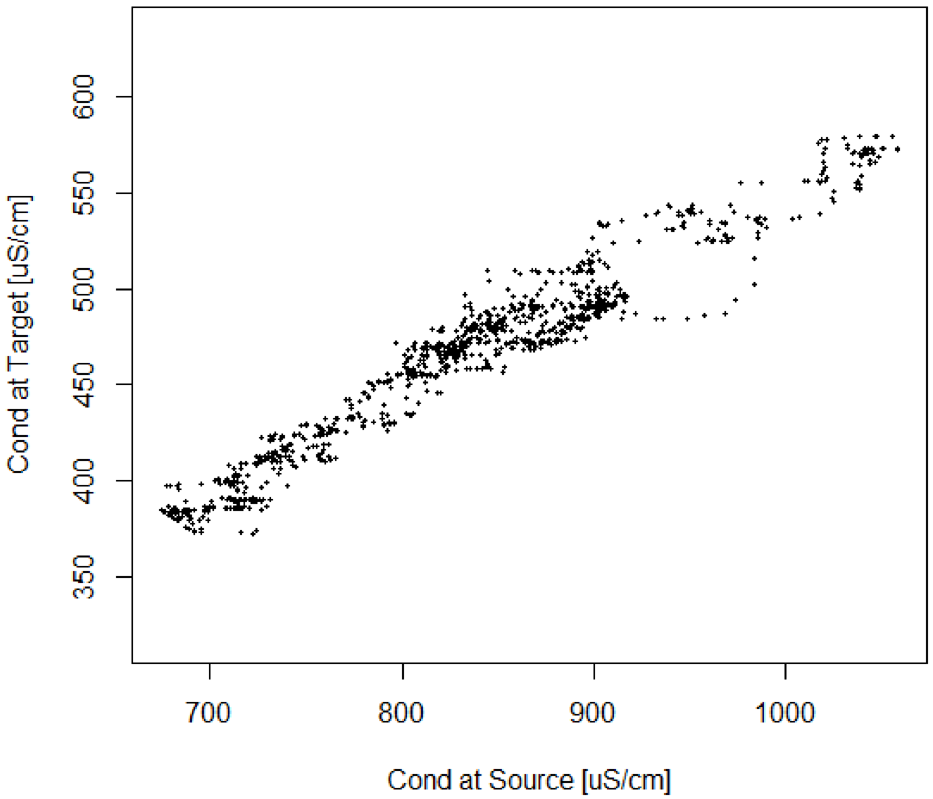


*Figure 9: Scatterplot for conductivity measurements from the Source and Target stations, depicted in Figures 7 and 8, respectively.*Each point represents a joint measurement taken from the two stations at the same hour, with X and Y values representing hourly median values.

Figure 10 shows the Cross-correlation Function (CCF) for the two conductivity time series, computed over 45 days of data, after aggregating the datapoints to hourly medians. We note that unlike the ACF, which considered only positive lag values, the CCF considers both positive and negative lag values. Both positive and negative lags can be considered, since the correlation of series of X's can be computed with both a right-shifted or left shifted version of the series of Y's.

The CCF shows a distinct peak at a lag of three hours, suggesting **the time required for water flow between the two stations is roughly 3 hours.** This is a **very coarse** estimate since:
I) There is no reason to assume the time required for water to flow is constant at all times.
II) There is no error estimation for the lag of three hours: are 45 days enough to provide an estimate? Will a different time period produce a different number? And if so, by how much?
III) We have no certainty the outliers in the data may bias the lag value at which the CCF receives its maximum value.

Throughout Unit A, we will describe tools that allow us to address issues I-III. Unit B will implement these tools altogether in order to address the above challenges.

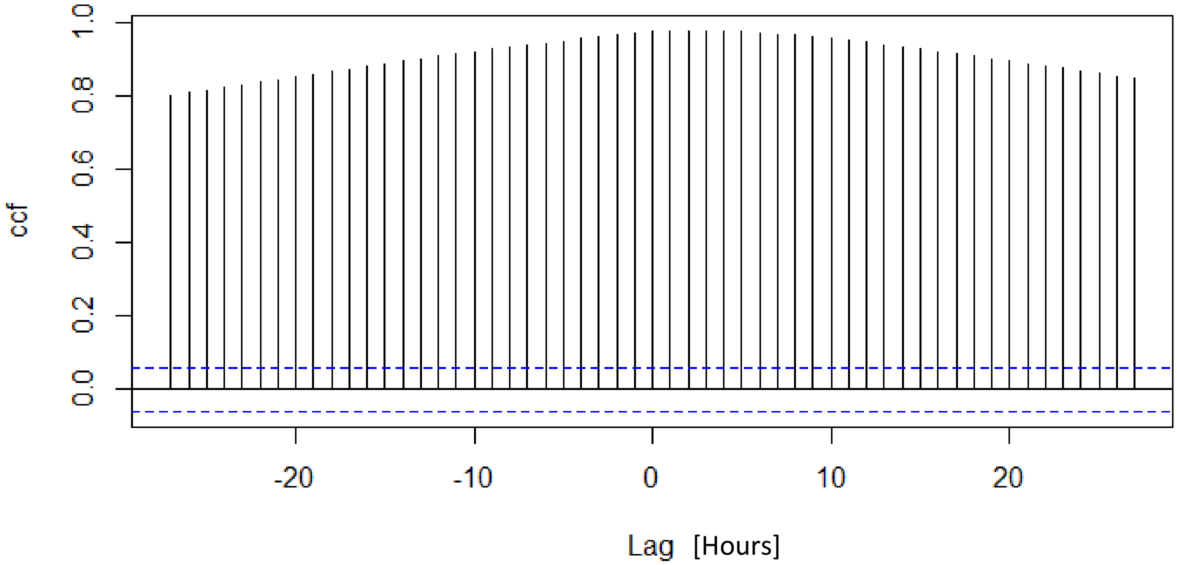


Figure 10: CCF for the conductivity data from the Source and Target stations, computed for for a 45 day period starting from 2017/03/04. Data points were aggregated to bin, and hourly median values were used to represent measurements before analysis.

Figure 11 shows concurrent, hourly median, conductivity measurements for the two measurement sites depicted in Figure 7-9, for a time span of 300 days. The data includes several calibrations of conductivity sensors throughout the time period, as observed by the 'jumps' between several linear relations in the data, featuring several different slopes. Figure 12, shows the CCF for the data depicted in Figure 11. While the CCF has a maximum present at LAG=3 Hours, it is much less distinct than the maximum presented in Figure 10 (avoiding a technical definition for 'maximum distinctiveness' at the moment). The flattening of the CCF is mostly due to normalization being done with means and standard deviations computed over a longer time period, containing several sensor calibration configurations. Unit B will address flow time estimations for time periods with several calibrations in detail.

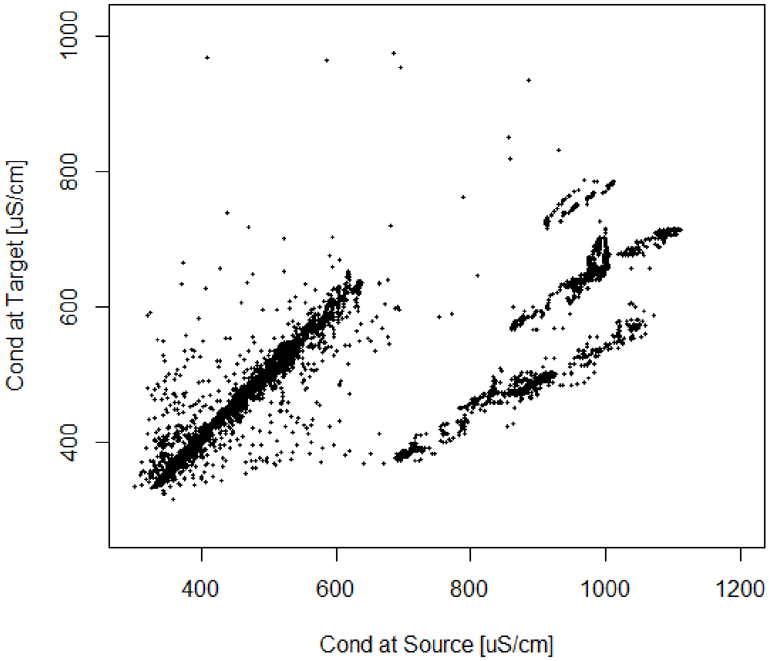


Figure 11: Conductivity measurements, for the two stations described in Figure 9, across a timespan of 300 days.

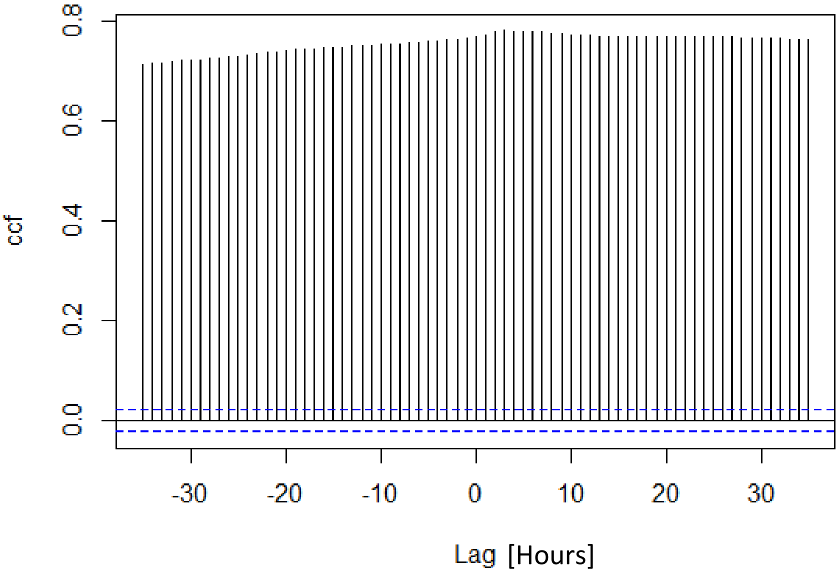


Figure 12: CCF for the data presented in Figure 11.

2.2 Irregular measurements and filtration methods in Time Series Analysis.

In Section 2.1, we have seen in Example B how a change in the baseline of sensor measurements, due to sensor calibration can bias analysis results. Careful analysis of the data and filtration of the calibration event, allowed the analysis in Example B to reach the required result (relate between the ACF decay rate by lag and data smoothness). Generally, data preprocessing and filtering must be performed before any statistical analysis. However, certain types of outliers and noise patterns are unique to water quality data and require specialized preprocessing. In Section 2.2.1 we detail several known forms of outliers common to water quality data. In Section 2.2.2, we describe possible solutions to the faulty data forms presented in Section 2.2.1. The solutions presented in Section 2.2.2 will be implemented throughout the Spatial Model methodology described in Unit B.

2.2.1. Some common forms of outliers in water quality data.

Fixed value – Figure 13 shows pH, Free-CL, and Turbidity measurements across a time span of two days. The turbidity measurements are fixed across the two days. This result is very unlikely since a constant level of turbidity may be experienced at levels close to zero (increased variance for turbidity is associated with an increased the baseline value), but not at values around 0.4 NTU. This type of phenomenon most likely indicates a problem at the analyzer level, since other measurements show valid values.

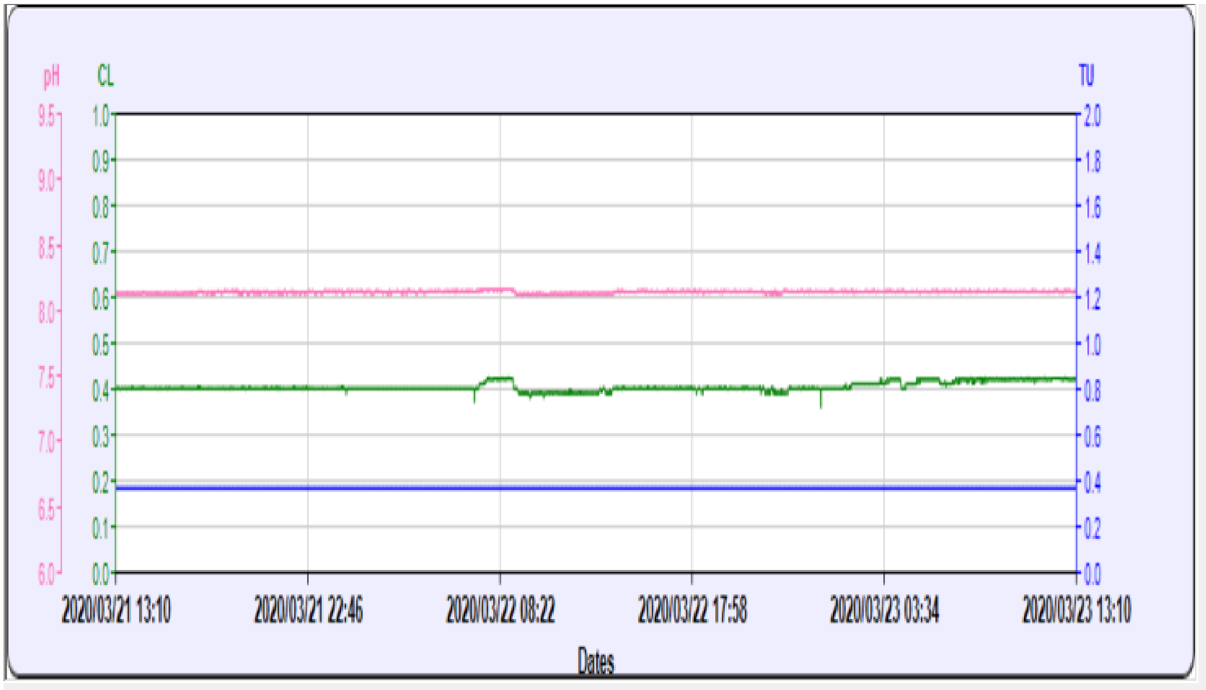


Figure 13: pH, Free CL, and TU measurements for a time span of two days, as measured in a water quality analyzer. Turbidity measurements are fixed across the two days, but pH and Free CL measurements vary, indicating a problem with the turbidity measurement, most likely at the analyzer level.

Communication problem – Figure 14 shows water quality data for three measurements, pH, Turbidity, and Free CL, for a time span of 24 hours as measured by a water quality analyzer. We observe a drop of all three measured values to zero, around 07:44 AM (marked by a red rectangle). This concurrent drop to zero can be due to a communication failure between the analyzer to the SCADA system, or due to a hardware failure at the analyzer level. A drop to zero in all three measurements indicates this data point should be removed from all future analysis. Typical values for communication faults or hardware failures include the value 0, the value 65535, the value 999 or negative values.

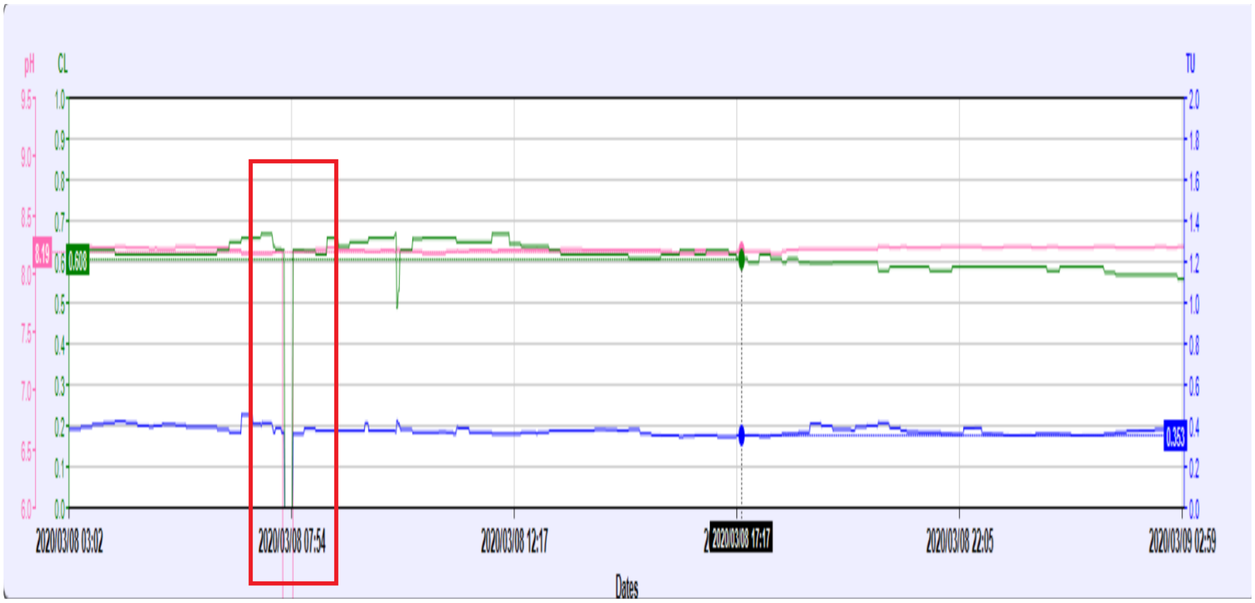


Figure 14:A drop to zero for three conqurenet measurements at water analyzer, at the same time. This phenomenon in the data indicates a communucation failure.

IO jump – Figure 15 shows a spike, for a single measurement, in three different measurements of the same water quality analyzer at the same time. Such jumps can be due to a voltage spike or a hardware fault in analog IO channels used to connect the water quality analyzer to a controller.

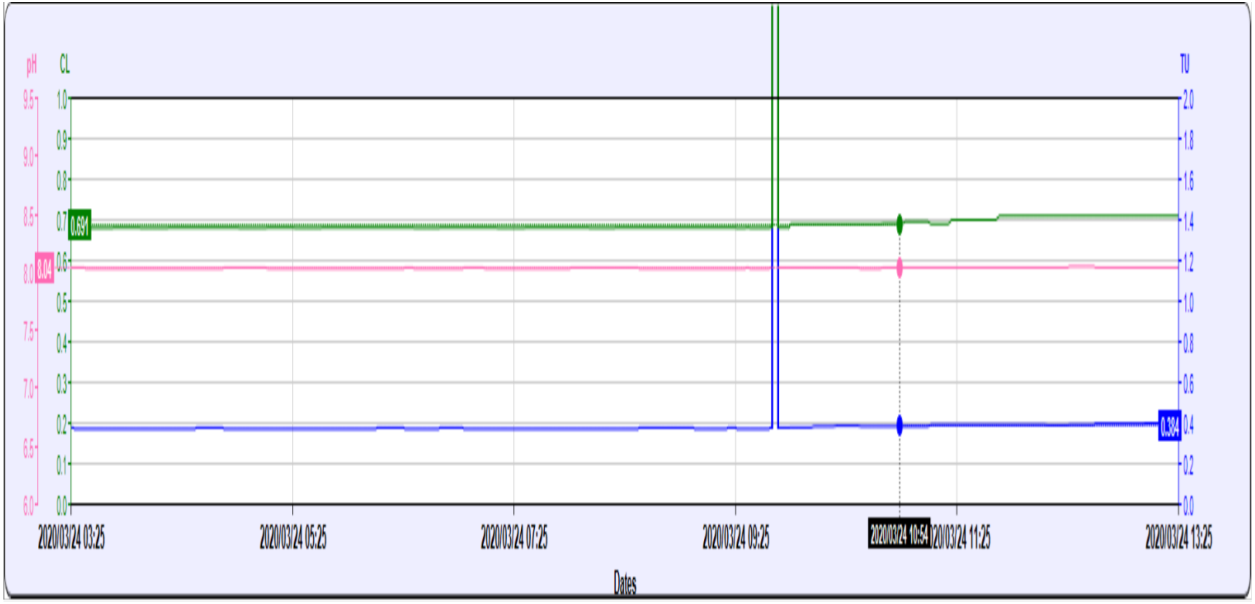


Figure 15: A concurrent jump in all three measurements due to a spike in IO analog channels.

CL drop – Figure 16 shows a drop in measured Free-CL values, at three single measurements (marked by red arrows). These drops in Free CL measurements are due to a faulty measurement, and do not represent the underlying levels of Free CL in the water.

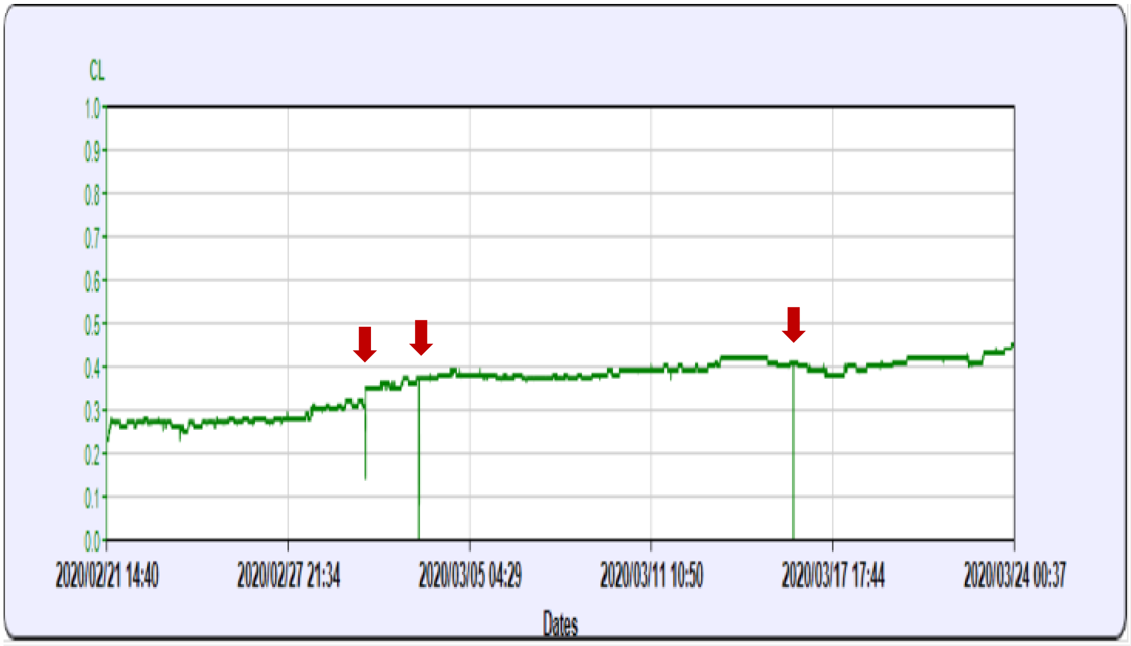


Figure 16: A time series for Free CL measurements. Single measurements are "negative spikes", o drop from the base line value.

Calibration – Figure 17 shows a calibration event in the data sample. The drop in turbidity values is due to scheduled maintenance and cleaning of the measurement compartment of the analyzer. Free CL measurement experience extreme high and low values due to the use of buffers in calibration.

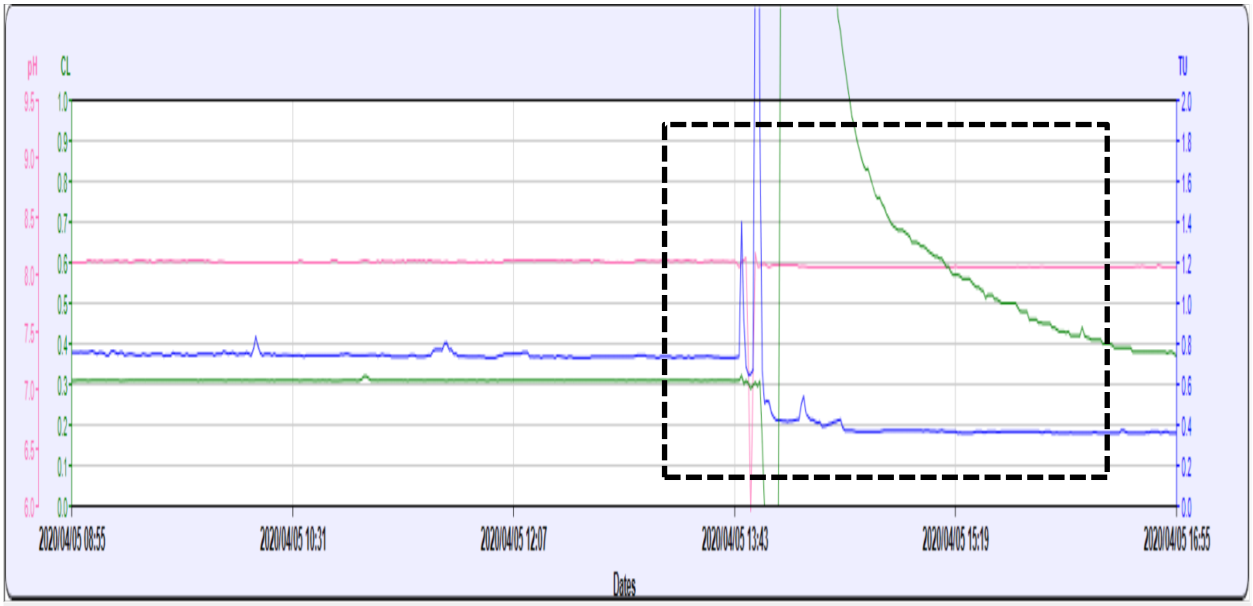


Figure 17: an event of sensor calibration and maintenance in the data in the data. A drop in turbidity is due to scheduled maintenance and cleaning of the measurement compartment of the analyzer. The change in CL is due to use of buffers in calibration.

2.2.2. Some simple, yet useful, ways to filter out faulty data measurements in water quality data.

Temporal binning and use of representative values – outliers and faulty measurements can be removed from the data, by binning the data points by their timestamps and taking the median of measurements within each bin as the representative value for measurements. Time series analysis and modeling can proceed by performing analysis on bin representative values instead of raw measurements. This approach was discussed in detail in Section 2.1.

Filter out by variable specific limits – Variables can be given upper and lower thresholds, limiting the measurements considered valid for analysis. For example, pH values above 8.5 or below 6.5 can be regarded as sensor malfunctions and can be removed from the data. Setting limits for valid variable measurements is location and time-specific, and involves looking at the distribution (histogram) of measurements, taken over an extended period of time.

Removing measurements that are constant across long periods of time – Water quality data may include measurements that are constant for extended periods of time, either to communication problems with analyzers or due to hardware faults. While a few consecutive measurements carrying the same value may not be due to an error, measurements that are constant across long periods of time are improbable. These fixed measurements should be removed from the data before performing analyses of the data, and even before the temporal binning of data, since fixed measurements misrepresent the underlying water quality values, and thus may bias analyses' results. Choosing the cutoff time period, after which a time interval with fixed measurement values should be removed, is location and measurement specific, as we show in the next example.

Example 2.2 A – removing data from extended time periods with constant measurements – For this example, we consider the Flow values as measured at the exit point from a water reservoir, providing water to a large residential district. Flow measurements are given in m^3/h, with typical values of 500-700. Analyses of flow measurements will be the main focus of Section 3.3 in Unit B, discussing the partitioning of water quality data to different flow regimes based on the distribution of water flow measurements. Therefore, preliminary preprocessing is required in order to ensure faulty flow measurements do not compromise the results of the analysis.

Throughout the two years of flow data gathered from the selected station, there are several time durations with values of zeros. These zeroes represent a technical error rather than actual zero flow for continuous times. Figure 18 gives the distribution of the lengths of time, in minutes, for the different periods where no flow were reported. Bins for the histogram are 20 minutes long. For example, if we observe the **red arrow** in Figure 18, we have a single "no-flow" event which lasted over 800 minutes, 5 events which lasted between 600 to 800 minutes (**green arrow**) and over 150 "no flow" events which lasted no more than 20 minutes (most of which lasted for a single measurement, **blue arrow**). All time periods with zero flow measurements that are above 30 minutes in length sum up to several days worth of data that should be removed from all further water flow analyses.

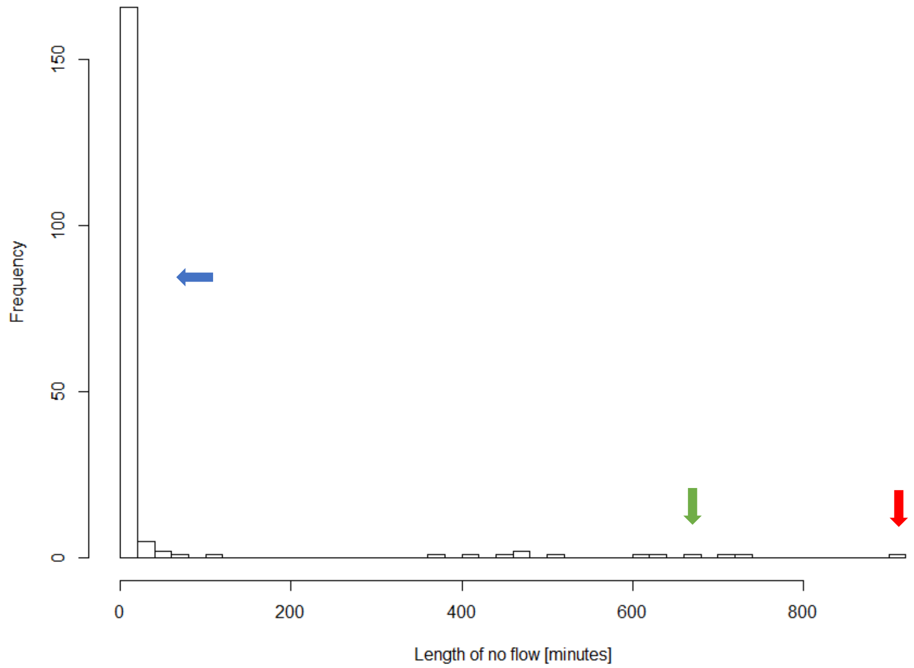


Figure 18: Histogram showing the distribution of of time period length having fixed flow values at a flow meter located at a water reservoir exit. The blue arrow shows most time period with fixed measurements are under 20 minutes, and are mostlikely a few consecutive measurements with the same values. The green and red arrows show 5 events with no flow measurements for at least 10 hours (600 minutes), and single time period with fixed flow measurements for over 800 minutes, respectively.

Remove calibration and water quality events – Some water quality events and time periods, including maintenance, are better removed from the data, before performing certain analyses. Scheduled maintenance and sensor calibration usually require some manual interference with the water measured by the sensor, e.g., disconnecting from the water line in order to clean measurement compartments, and using premade water samples for calibration. Sampled values do not represent the underlying water properties in the network, and should be removed from analyses.

Abnormal water quality events inserted into the data used to determine the threshold for declaring future water quality events may bias to thresholds, and make them more conservative, at the danger of a "false negative": missing future water quality events. Therefore, abnormal water quality events need to be removed from any analysis used to determine alarm thresholds. However, analyses used to determine flow time between stations may benefit from using data from abnormal water quality events that observed traveling through the water distribution network at different times at locations.

Ranking – Ranking is a data transformation replacing a series of measurements by their respective ranks, from 1 to N. For example, ranking the observations 5,8,6.5,32,9 would lead to the ranks 1,3,2,5,4. The purpose of ranking is to confine the possible values the data may take to a predetermined set of values that A) is known in advance, and is scale agnostic B) is unaffected by outliers in the data and C) maintains their relative ordering.

A data transformation that can be given as a counterexample, for upholding criteria A-C is normalizing data sets by subtracting the mean of sample values from the data. The sample mean may be highly sensitive to noise in the data, and hence may "mean subtraction" may compromise further analyses.

We note that rankings are only interpreted within a sample. If a measurement was given the rank 46, out of 52 measurements, there is no way of knowing what exact rank it will obtain in another sample. Moreover, ranking a single sample is irrelevant – it will always receive the rank of 1.

While ranking upholds properties A-C described above; some analyses are cannot be performed on ranks, setting thresholds for irregular events or finding the baseline measurement value for some sensor.

A form of analysis which can benefit from ranking observations before analyses, is computing correlations. If two series of measurements are positively correlated, measurements with high values in the first series are likely to be associated with measurements with high values in the second series. This association can still be observed in ranked datasets: if both data series are ranked, measurements with high ranks in the first series are likely to be associated with measurements with high ranks in the second series. Analyzing correlations over ranks, rather than over the raw data, ensures that correlation estimates do not hinge on outliers in the data. Robust measures of correlations for data analysis will be the our next topic of discussion in Section 2.3.

2.3 Correlation and Non-Parametric Correlations

In Section 2.1, we showed how correlation coefficients may provide insights when discussing time series smoothness, or the relationship between one time series of measurements to another. In Section 2.2, we presented several types of measurement errors and outliers in the data. At the end of Section 2.2, we briefly introduced the idea of transforming data measurements to ranks prior to analyses. The purpose of this subsection is to extend the discussion to other correlation measures (other than Pearson), that are robust to outliers in the data . The next example will serve as a working example for most of the subsection.

Example 2.3A – simulated dataset for discussing the effect of outliers on correlation estimation: Figure 19 shows a simulated dataset with 50 data points. The X,Y values were simulated from a data generating process, causing them to have a mild but distinct positive correlation, as seen in the graph. We are interested in testing how the Pearson correlation coefficient will respond to noise added to the first (leftmost) data point in the chart.

(Example cont. on next page)

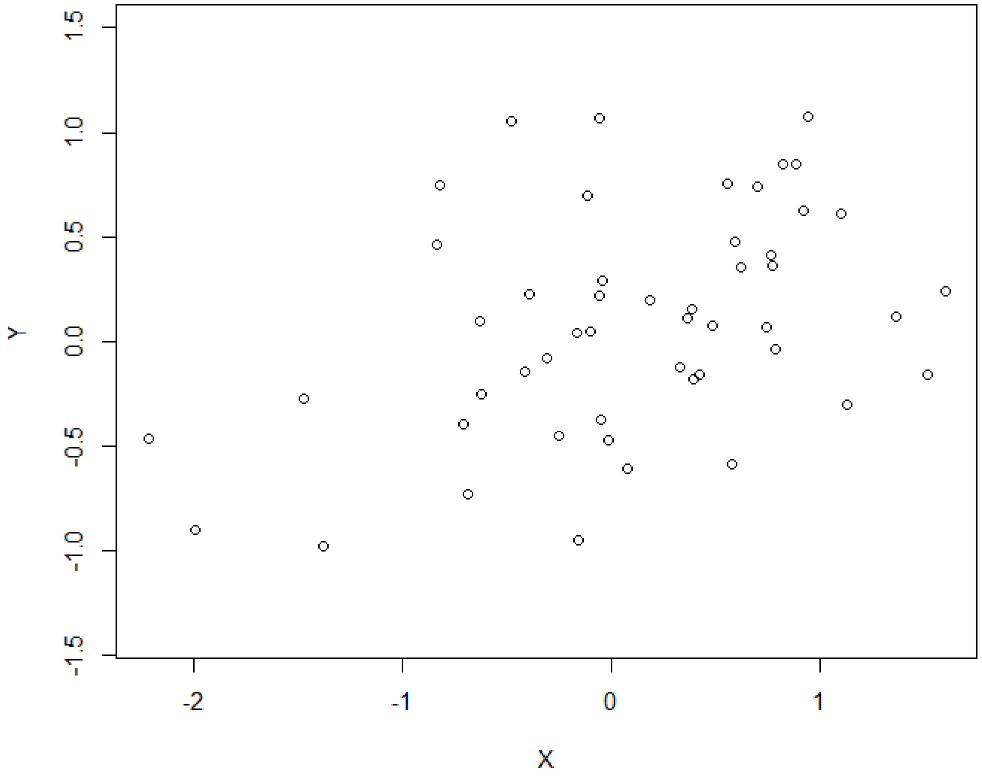
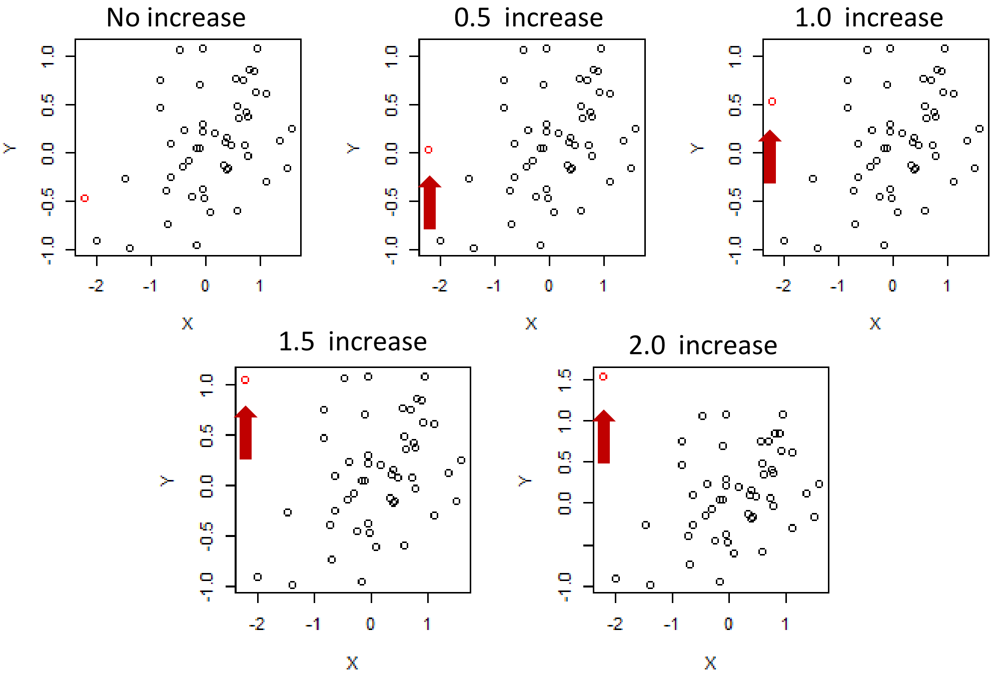


Figure 19: A simulated data sets with 50 data points, showing a positive correlation.

Figure 20 shows 5 augmented version of the data presented in Figure 19. In each version of the data, the Y value of the first measurement was increased by either 0 (no change in value), 0.5,1.0,1.5 or 2.0. The first point is marked in red to show it's movement across different versions of the data.



No increase

0.5 increase

1.0 increase

1.5 increase

2.0 increase

Figure 20: Five augmented versions of the data presented in Figure 19, with varying levels of noise added to the first measurement. The single point moved between different versions of the data is marked in red.

We proceed and calculate the Pearson correlation coefficient for the five datasets presented in Figure 20. Figure 21 (on the next page) shows the Pearson correlation coefficient as a function of the level of noise inserted, for the five datasets displayed in Figure 20. We observe the Pearson correlation coefficient decreasing as a function of the level of noise added. We conclude the Pearson correlation coefficient is not a robust measure of correlation: even a single measurement may affect the estimated correlation. The problem this subsection will address is how to define correlations that are insensitive to outliers in the data.

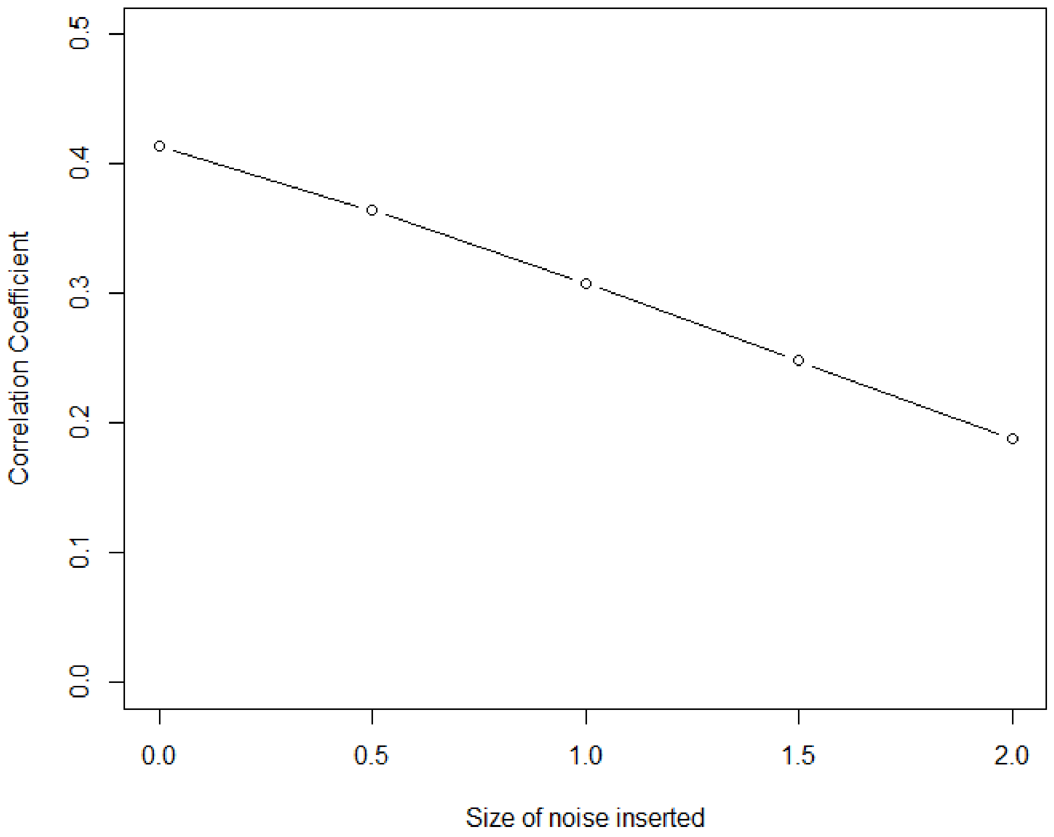


Figure 21: The Pearson correlation coefficient as a function of the amount of noise added, for the five augmented datasets presented in Figure 20.

Robust correlation measures

We present two robust measures of correlation: Spearman's correlation coefficient, and Kendalls' tau.

Spearman's correlation coefficient is defined to be the Pearson correlation coefficient of the ranks of X and Y. In order to compute the Spearman correlation coefficient, both sets of coordinates, X and Y, are replaced by their respective ranks, and the Pearson correlation coefficient is computed as-is over the rank transformed data set.

In order to define **Kendall's tau**, we will need to define concordant and discordant pairs of data points.

Figure 22 shows a sample data set, with concordant and discordant pairs of points in the data. **Concordant** pairs are pairs of data points were an increase in X values also shows an increase in Y values. **Discordant** pairs are pairs of data points were an increase in Y values shows a decrease in Y values. For a dataset with n points, there are $n \cdot (n-1)/2$ possible pairs of points.

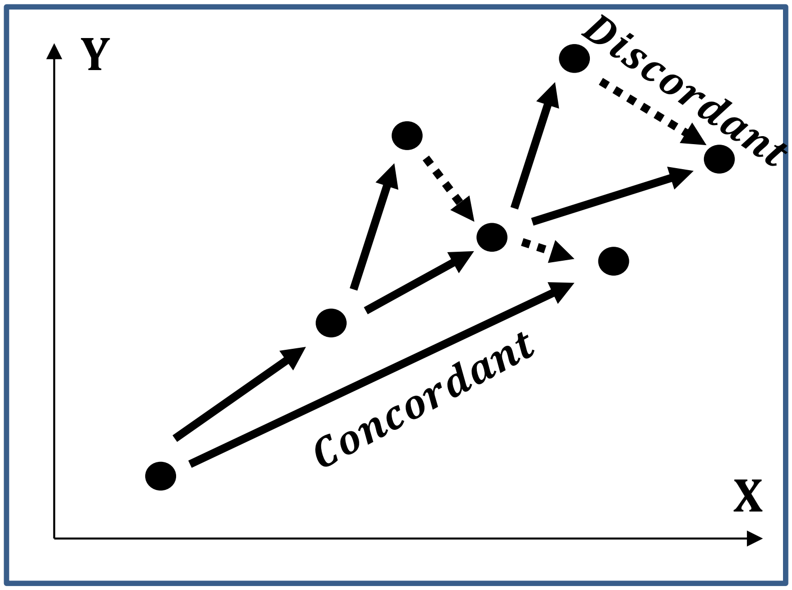


Figure 22: Concordant and discordant pairs of data points in an example data set. Concordant pairs are marked by a solid-like. Discordant pairs are marked with a dashed line. Not all pairs in the data are marked.

Kendall's tau is a correlation coefficient using the concept of concordant and discordant pairs, in order to assess the directionality and magnitude of the correlation between the sample values of X and Y. Kendall's tau is defined to be the difference between the number of concordant pairs and discordant pairs, out of the total number of possible data point pairs:

$$\tau\,(X_{1:n}, Y_{1:n}) = \frac{\#\boldsymbol{Concordant\ pairs}\ - \#\boldsymbol{Discordant\ pairs}}{n \cdot (n-1)/2}$$

Similar to Pearson's and Spearman's correlation coefficients, τ also obtains values in the range $[-1,1]$.

We now return to our working example, and show the benefit of using these two robust measures of correlation.

Example 2.3A (continued) – Figure 23 shows the Pearson correlation coefficient, as well as the Spearman correlation coefficient and Kendall's tau for the five data sets presented in Figure 20. This figure is the equivalent of Figure21, with two additional measures of correlation added. We observe the relative decrease in Pearson correlation to be substantially greater than the relative decrease in the other two correlation coefficients, as the inserted noise in measurement increases. Next, we extend this example and insert noise into several measurements.

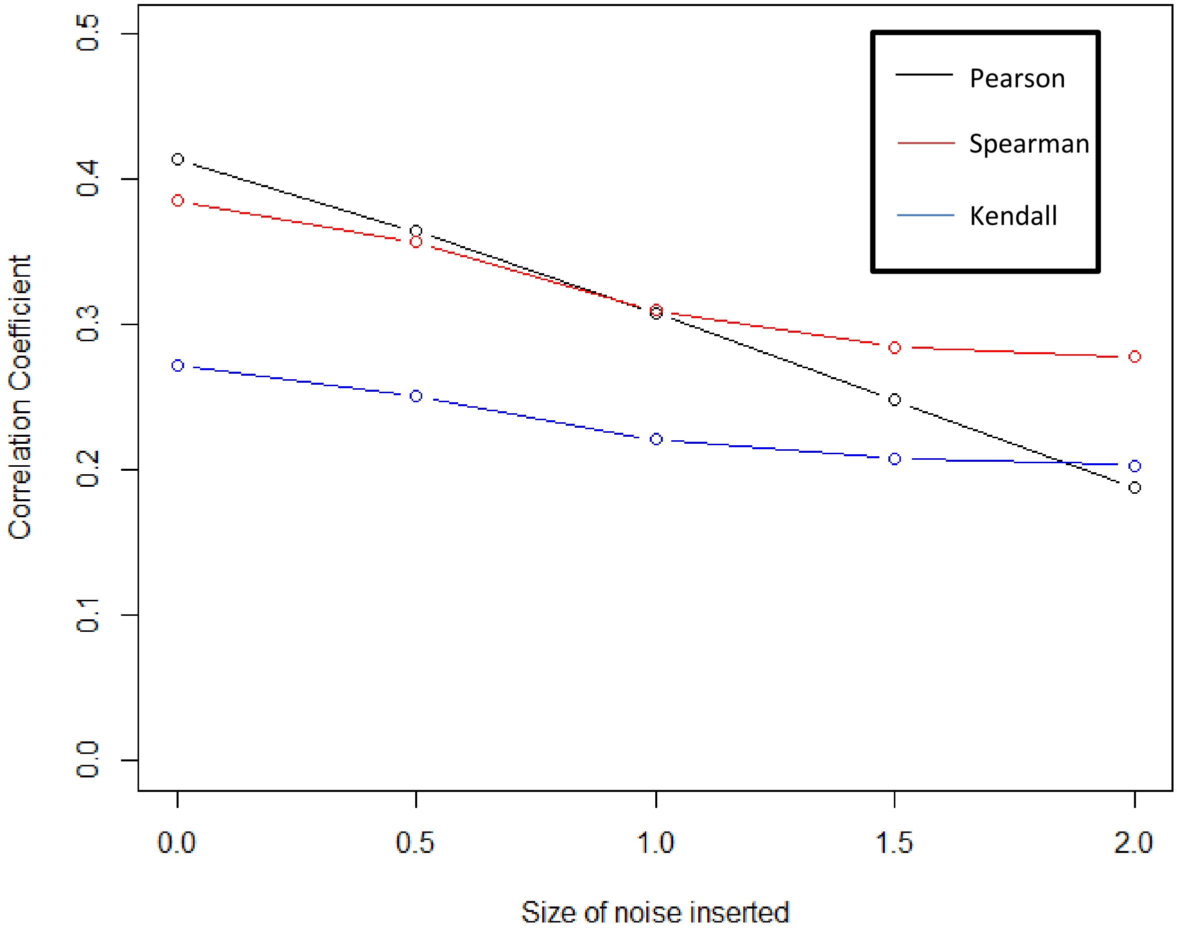


Figure 23: Pearson's, Spearman's, and Kendall's correlation coefficients as a function of the level of noise inserted to a single outlier, computed for the five data sets described in Figure 20.

Example 2.3A (continued – noise inserted at five measurements) – Figure 24 shows five data sets, similar to the one presented in Figure 19 (the original data), but with noise inserted into the first five measurements (as opposed to Figure 20 that showed noise inserted into a single measurement). In the next page, we present the three different correlations computed for this data as a function of the level of the noise added.

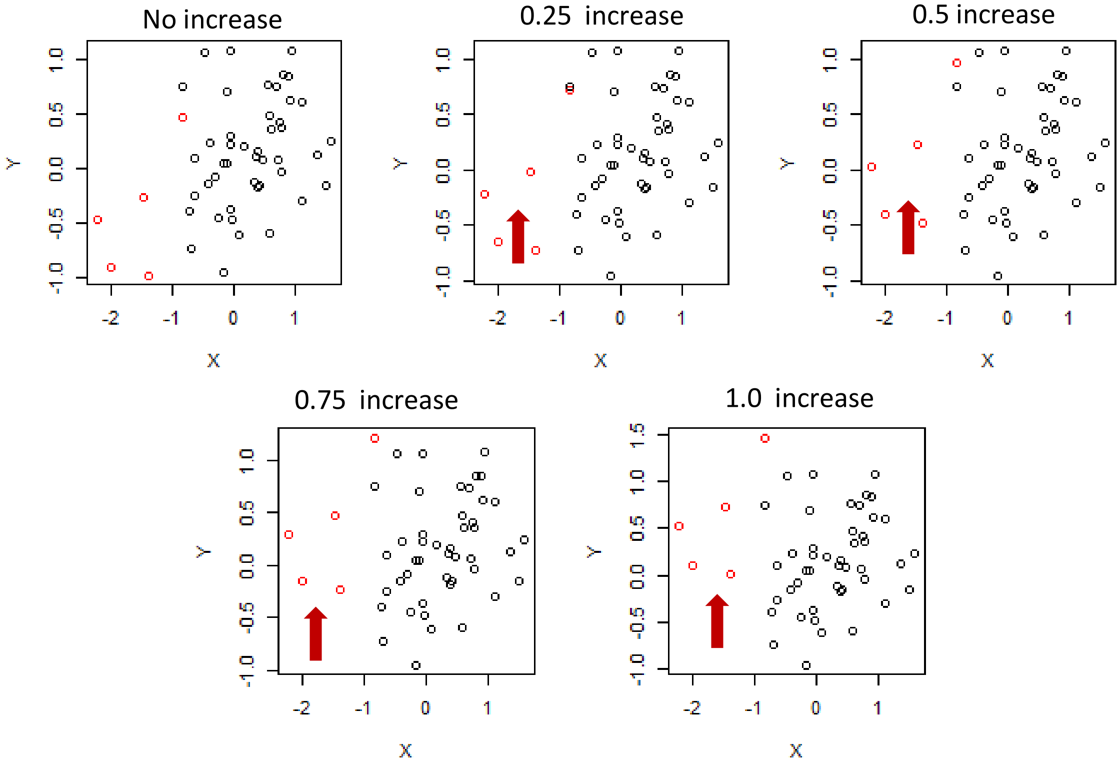


Figure 24: Five datasets generated from inserting noise to five leftmost measurements in the data described in Figure 19.

Figure 25 the three correlation coefficients as a function of the level of noise added. Again, the relative decrease in Spearman's and Kendall's coefficients is smaller than the one observed in the Pearson correlation coefficient as a function of the noise added. The effect observed in Figure 23 is even more severe in this figure: Pearson's correlation coefficient may reach zero (no correlation) when only 10% of the data are affected by the added noise. Next, we present a real data example comparing the three different correlation coefficients.

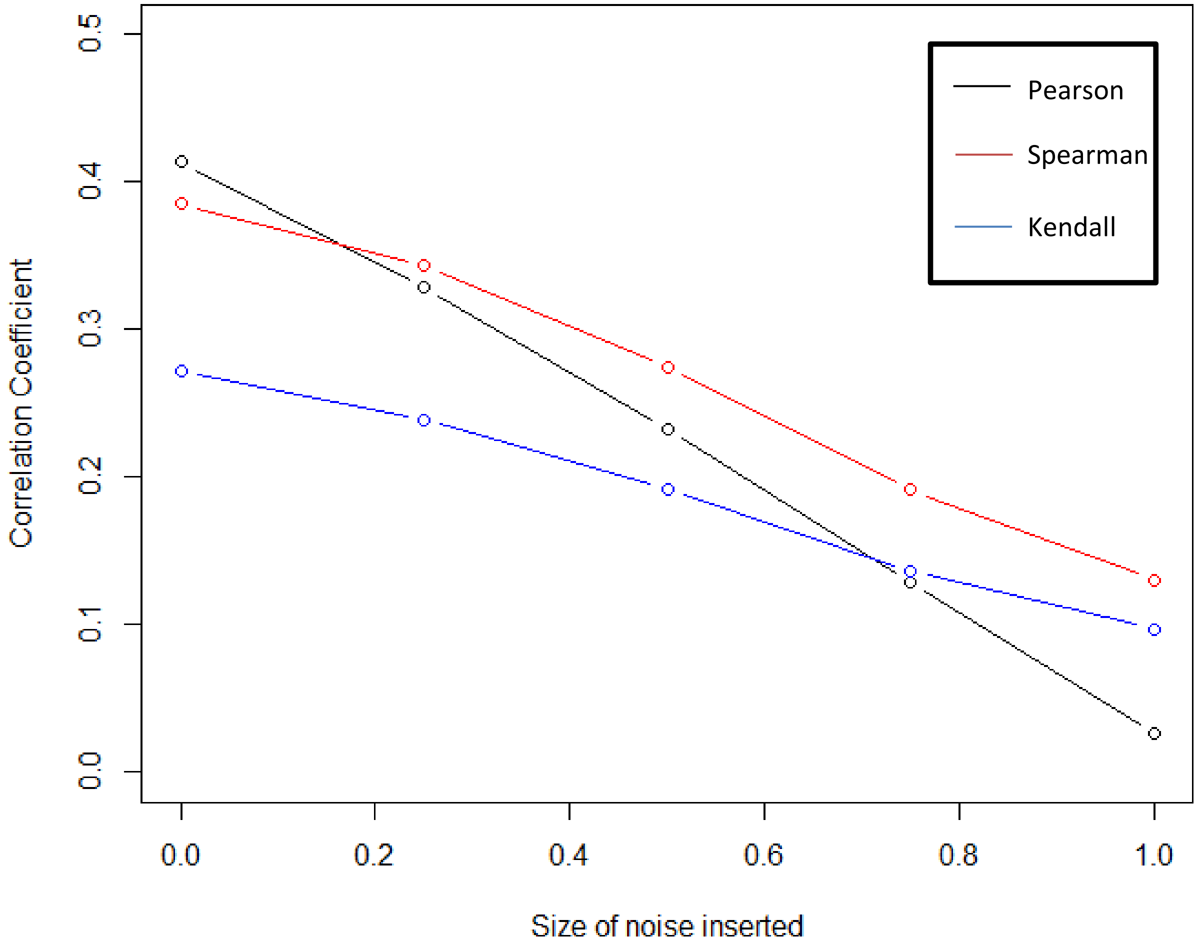


Figure 25: Pearson, Spearman and Kendall correlation for the five datasets displayed in Figure 24. Correlations are given as a function of the noise added.

Example 2.3B – Comparing correlation coefficients for data measured in two stations: Figure 26 shows Turbidity measurements for a time span of two years, as measured in two neighboring water quality analyzers in a water distribution network. Water flows from the first station, called "Source", to the second station, called "Target." Data coordinates represent the hourly medians of turbidity measurements at the two stations, for the same hour. No adjustment was made for the time required for water to flow between the two network locations.

Data in Figure 26 is characterized by several "clusters" of points, with each cluster having a positive correlation (slope). However, different clusters show different slopes. This grouping of observations may be due to either a change in the water source, measured compartments of one (or both) sensors being affected by dirt or residue, or periodic calibration of the sensors, leading to different "response curves" to the actual turbidity values in the water. Nevertheless, within each cluster, we observe a strong positive correlation. Looking at the data point cloud as a whole, it is difficult to assess if a correlation coefficient computed over the entire dataset will be positive.

When computing the three correlation coefficients for the data in Figure 26, Pearson's coefficient has the value of 0.044, Kendall's tau has a value of 0.157, and Spearman's coefficient has a value of 0.221. We observe the two robust measures of correlation presented in subsection to have higher coefficients for this data.

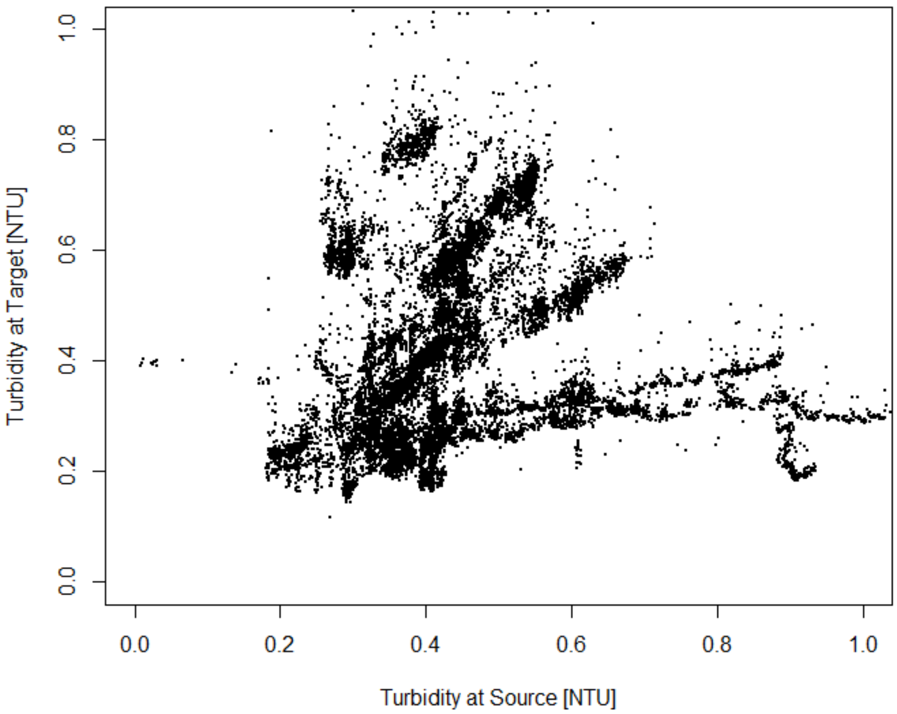


Figure 26: Turbidity measurements for a time span of two years, across two neighboring stations. Water flows from the Source station (X-axis) to the Target station (Y-axis). Data points represent hourly medians for the same hour at the two stations.

To conclude, in this subsection, we have shown Pearson's correlation coefficient to be highly sensitive to outliers in the data. We presented two correlation coefficients based on the ordering of observations, which proved robust to outliers or seasonal effects in the data.

2.4 Bootstrap sampling and Bootstrap sampling for time-series.

The bootstrap is a statistical sampling procedure used to estimate the variance of possible error of an estimation procedure. The idea behind the bootstrap is to imitate the variance in the original sample caused by sampling, by generating 'possible datasets' that could have been observed, from the same data generating process. In it's simplest form, the bootstrap creates a large number of possible datasets, e.g., 100, each with size identical to the original data, by sampling observations without replacement from the original data. The estimation procedure is run independently on the different bootstrap samples, and the variance of the different "bootstrapped" estimates is computed, in order to estimate the variance of the estimator for the original data. This principle is best demonstrated with an example.

Example 2.4A – Estimating the variance of the sample mean, when estimating the population means. Consider an even 6-sided die. Each of the numbers 1-6 may be sampled with an equal probability of 1/6. The population mean or the mean of an infinite number of dice throws, is $\frac{1+2+\cdots+6}{6} = 3.5$. For this example, we will assume the population mean is unknown, and we wish to estimate it using a sample mean. The sample mean will provide a point estimate, a single number trying to estimate the population mean of 3.5. Therefore, we will also be interested in assessing how accurate the sample mean is in this estimation, or, what is the variance of the sample mean. Bias estimation will not be discussed in this example.

For our example, we measure 10 independent dice throw results:

2 3 4 6 2 6 6 4 4 1

For this data, the sample mean is 3.8. Since we are familiar with the generating process for this data, we know the population mean to be 3.5. We are interested in estimating the variance of the sample mean (as a procedure), when trying to estimate the population mean. In order to do that, we generate 100 "bootstrap datasets" from the original data. This is done by sampling 10 observations for each generated dataset, with each measured result obtained by sampling one of the original data points with equal probability. We obtained the following bootstrap samples:

Bootstrap sample 1: 1 6 2 1 2 4 4 4 3 4
Bootstrap sample 2: 2 4 3 4 3 4 6 4 6 4
Bootstrap sample 3: 6 2 6 1 6 3 6 6 4 3
Bootstrap sample 4: 4 3 6 6 3 2 6 6 2 1
⋮
Bootstrap sample 100: 2 6 2 1 4 3 4 2 6 2

Each of the 100 bootstrap samples could have been observed instead of the original data, with equal probability. We note that the number 5 was not sampled in the original dataset. Therefore, none of the bootstrap samples include the number 5.

We compute the sample means for the 100 bootstrap samples. Figure 27 shows the distribution of the 100 bootstrap samples.

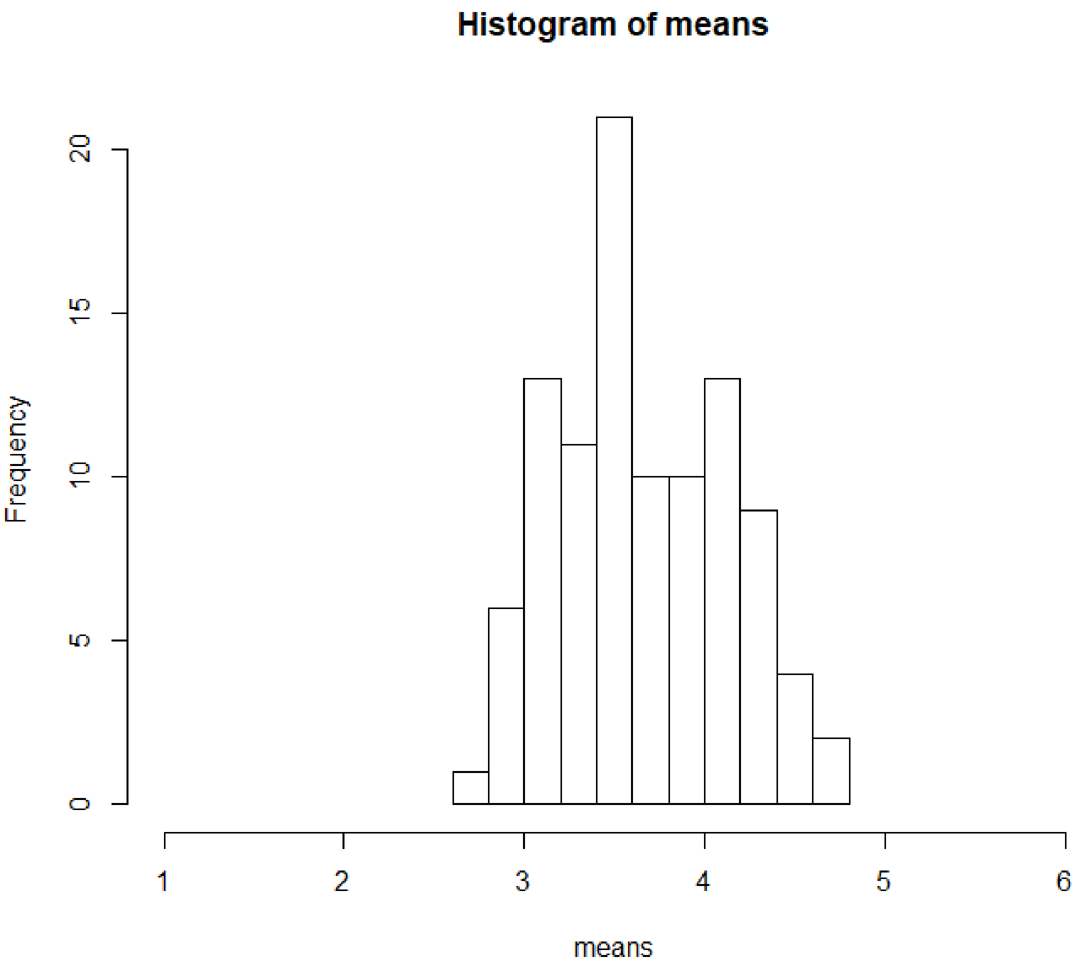


Figure 27: Distribution of bootstrap sample means for the bootstrap samples of Example 2.4A

The 0.05 and 0.95 percentiles of the distribution of samples means are 3.0 and 4.5, respectively. The bootstrap procedure concludes that the range $[3.0,4.5]$ is a 90% confidence interval for the population mean. This is quite a large range: we would like the decrease our uncertaintanty in estimating the population mean. However, we cannot do so by simply increasing the number of bootstrap samples. The Bootstrap procedure simply imitates the variance in estimation when observing a sample of 10 observations from our data source. To produce tighter bounds on the population mean – our original sample must include a larger amount of samples. We show that larger datasets produce shorter confidence intervals in the following example.

Example 2.4B – Estimating the variance of the sample mean, when estimating the population mean. (continued - with larger samples):

We consider a case where we observe 50 measurements from an even die. The observed sample is

2 3 4 6 2 6 6 4 4 1 2 2 5 3 5 3 5 6 3 5 6 2 4 1 2 3 1 3 6 3 3 4 3 2 5 5 5 1 5 3 5 4 5 4 4 5 1 3 5 5

The observed sample mean for the 50 observations is 3.7. We produce 100 bootstrap samples, 50 observations each:

Bootstrap sample 1: 1 4 2 5 6 2 3 3 2 6 5 5 4 5 3 5 1 5 2 4 5 4 6 5 1 4 4 5 5 5 2 5 5 5 1 2 5 6 5 4 2 4 3 4 5 3 4 6 5 3

Bootstrap sample 2: 3 6 3 5 4 2 6 1 1 3 5 5 6 2 4 2 5 6 4 3 5 2 2 4 1 4 3 2 2 3 6 6 3 3 1 3 6 1 5 3 5 3 2 5 5 4 4 1 6 4

Bootstrap sample 3: 3 3 5 4 3 1 1 6 3 2 5 4 4 5 4 6 6 5 5 5 4 5 5 5 5 4 3 1 3 5 5 1 4 3 4 1 1 5 3 3 5 5 6 1 5 3 6 5 3 3

Bootstrap sample 4: 3 2 3 3 1 3 6 6 5 5 5 6 5 3 5 3 3 3 3 5 3 4 5 3 1 3 5 4 6 5 6 6 3 4 5 5 5 3 2 4 3 6 5 4 6 2 2 6 1 3

⋮

Bootstrap sample 100: 5 1 1 3 2 6 5 2 3 1 1 3 3 3 5 4 6 1 4 4 6 6 5 4 5 1 5 1 1 5 3 1 4 3 4 6 5 4 6 2 2 6 5 5 3 5 2 6 4 1

Figure 28 shows the distribution of sample means, for the 100 bootstrap samples. The 90% central values in the distribution are found in the range $[3.319,4.02]$, giving a shorter confidence interval than the one obtained in Example 2.4A.

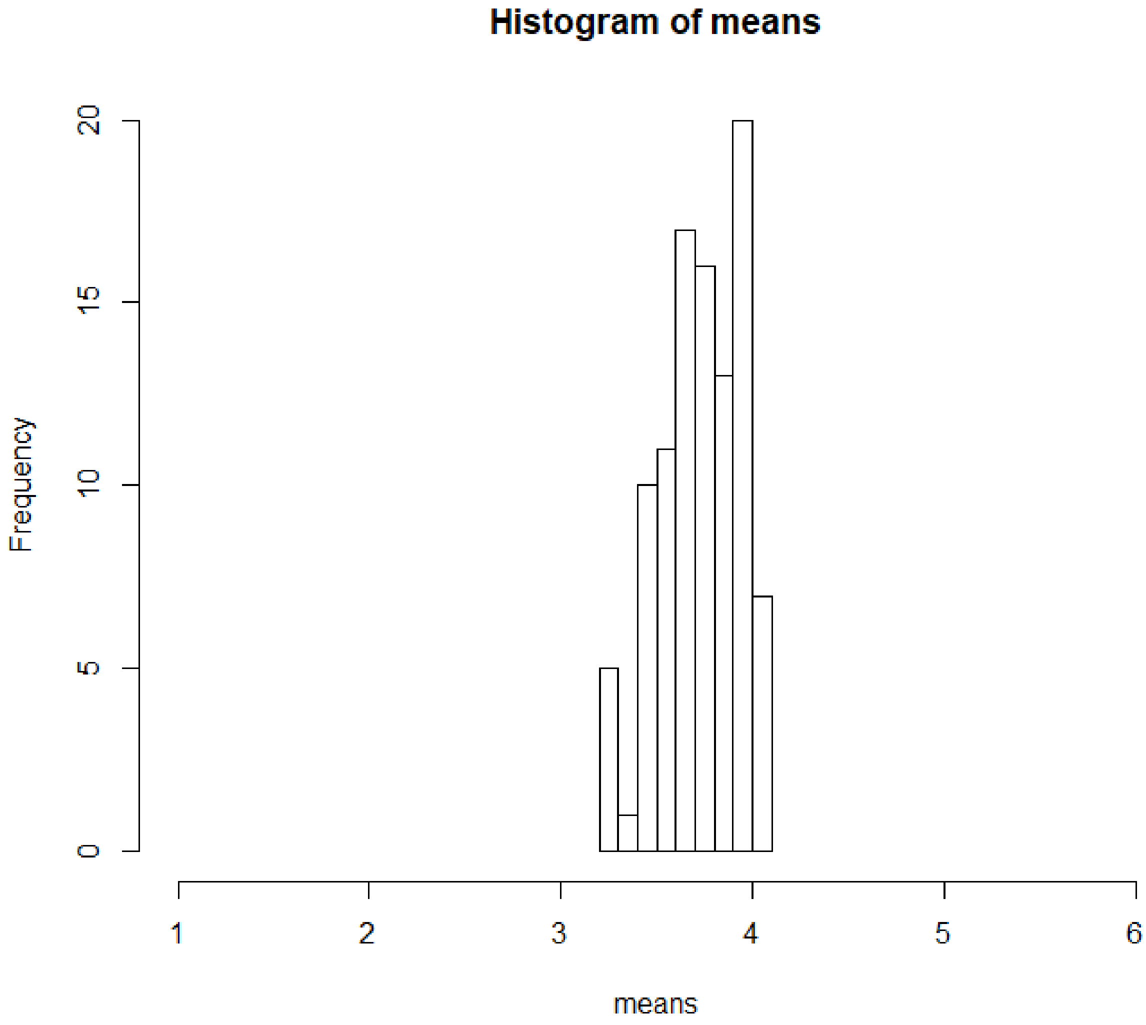


Figure 28: Distribution of bootstrap sample means for the bootstrap samples of Example 2.4B

The previous example demonstrated the bootstrap sampling procedure for data with independent samples. However, measurements in water quality time-series data are not independent across time, as shown in Section 2.1. We proceed to demonstrate a bootstrap sampling procedure for time series data.

When trying to estimate population parameters by bootstrap, one cannot sample individual observations. However, different time periods, which are not overlapping or in high proximity in time, can be considered as independent and be sampled with replacement from the data to form makeshift time-series imitating the original data. For example – from a year with 52 weeks, sample a new 52-week year by sampling one week after another from the original data. A drawback of this approach is that by stitching together weeks which are not adjacent from the original data, the bootstrapped time series will be noncontinuous once every seven days.

If the quantity of interest is local in time, e.g., estimate the distribution for the rate of change in the data, estimates can be computed over many short terms windows in time and bootstrapped estimates by computing samples estimates obtained from the different windows. We show a practical example for a time-series bootstrap procedure making use of local estimates across windows.

Example 2.4C – a bootstrap procedure for estimating the distribution of hourly changes in conductivity, and testing if the distribution of hourly changes in conductivity has changed across different years: The data for our example is a time series of conductivity measurements, measured throughout the years 2018-2019 at a sampling location located in an urban water distribution network. We split the dataset into two periods – February to December 2018 and February to December 2019. We are interested in testing whether the distribution of the hourly rate of change of conductivity differs between the two years: was the hourly rate of change of conductivity larger or smaller in 2019 compared to 2018?

The month of January was excluded from the data due to a different operation scheme in January 2018. Figure 29 on the next page, visualizes the data.

Figure 29 shows histograms for the distributions of the maximum daily change across the two years. The histograms were formed by computing the difference between the maximal and minimal daily measurements for each calendrical day in the data.

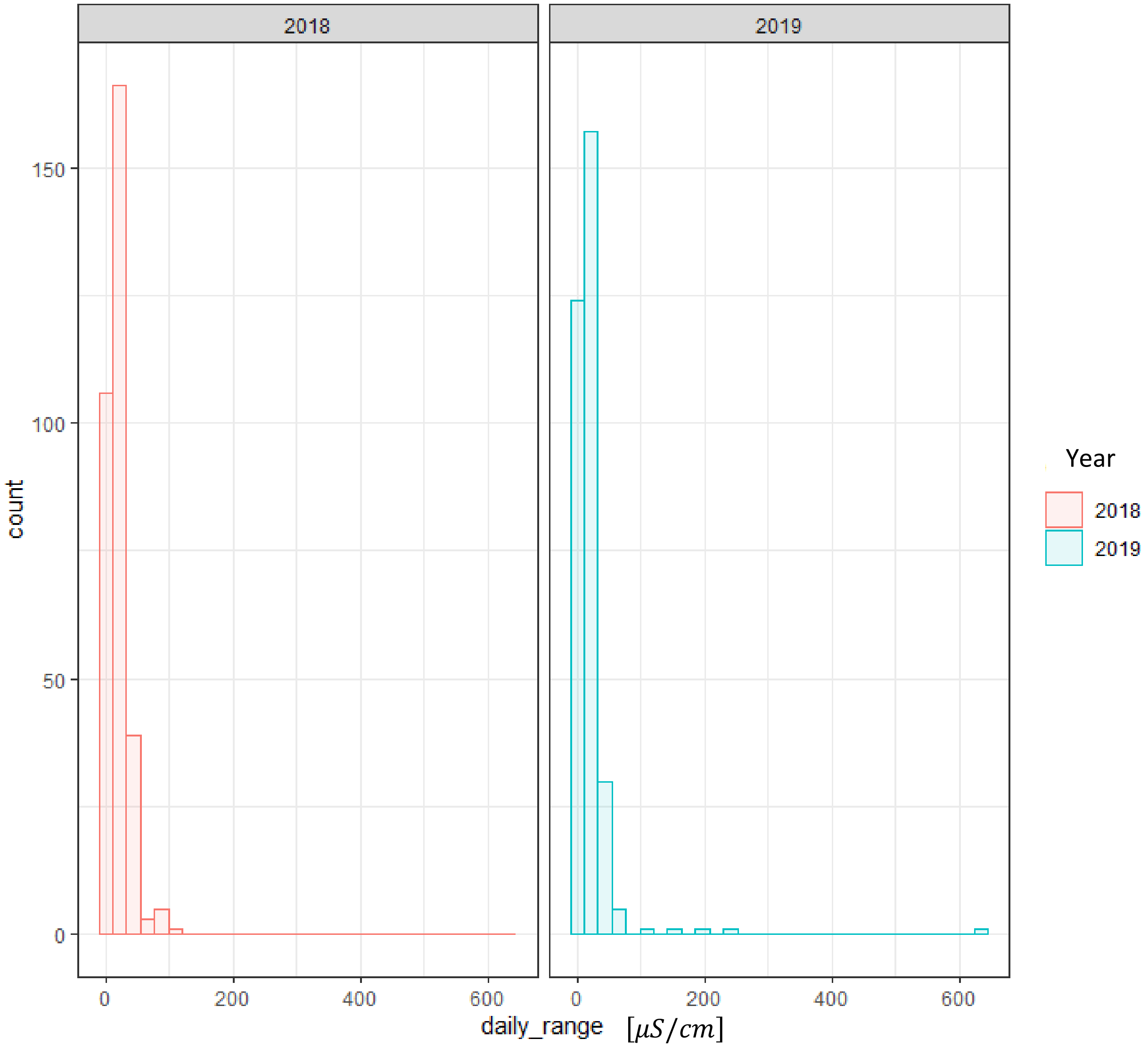


Figure 29: Histogram for the maximum change of conductivity measurements in a day across the two years in the data.

Our task is to find out if the differences between the two distributions can be explained by sampling error alone (i.e., we were likely to see the same two graphs, with years for labels switched) or not. Our approach will consist of sampling 2018-like and 2019-like bootstrap years of data, and comparing the bootstrap samples generated from both years.

From each of the two years, we sample 1000 bootstrap datasets (years), by sampling 11 months of 14-day intervals from the data, with replacement (an interval may be sampled several times). Figure 30 shows the Cumulative Distribution Function (CDF) for the two sets of bootstrap samples by year. Each graph shows the 0.05, median, and 0.95 quantiles, for each value of "maximal daily change in conductivity". **The figure is interpreted as follows: each time we generate a bootstrap sample for a year, its CDF has a 90% chance of passing between the two furthermost curves on each graph.**

On the next page – we compare the two sets of bootstrap samples

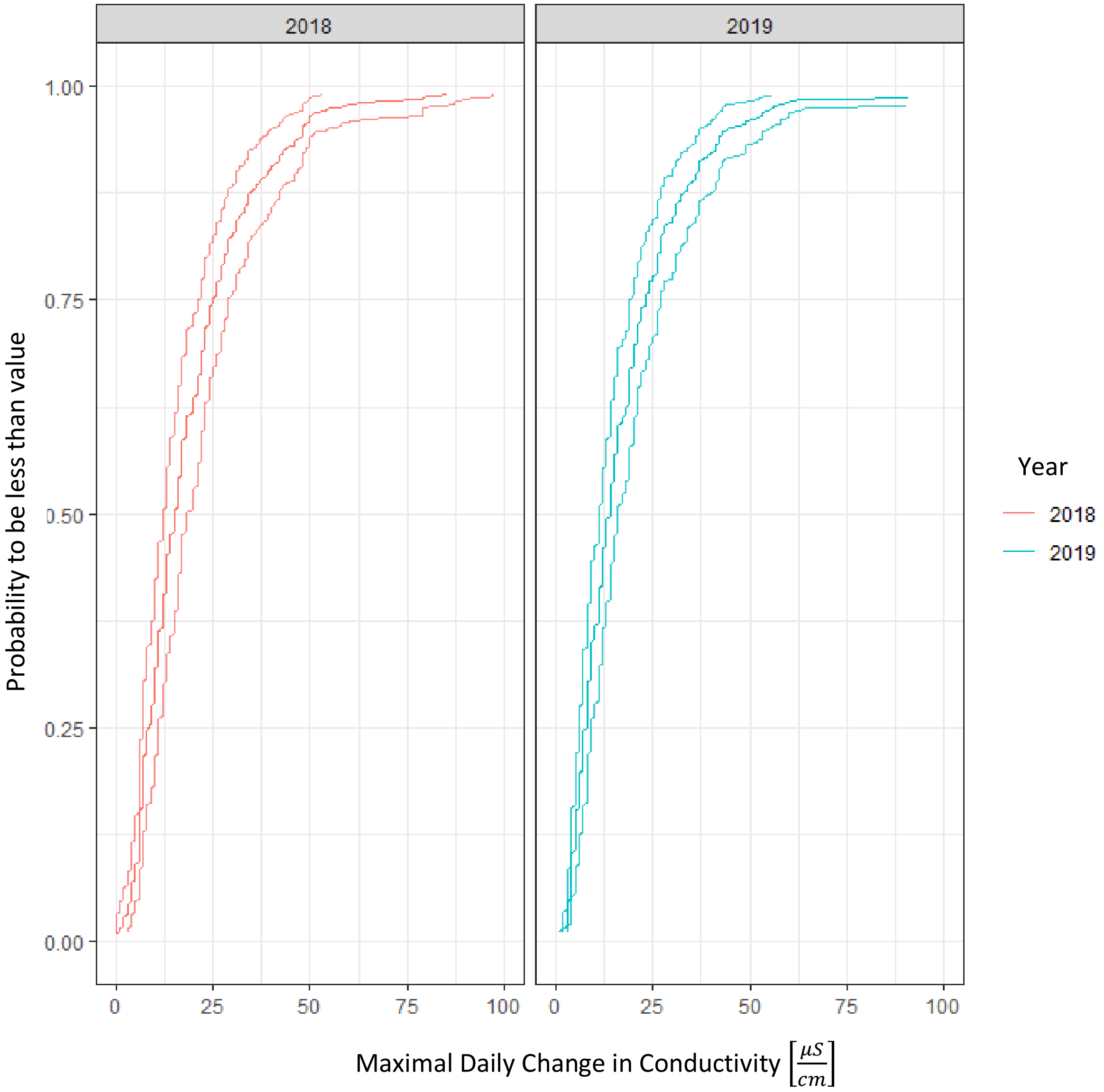


Figure 30: Cumulative distributions functions for the Maximal daily change in Conductivity, across 1000 time-series bootstrap samples generated from the datasets shown in Figure 29. The three lines in each year represent the 0.05, median, and 0.95 quantiles, for the distribution of the daily range of Conductivity, across the bootstrap estimates.

Figure 31 plots the two sets of bootstrap distributions, shown in Figure 30, together on the same graph. We see a substantial overlap between the two sets of bootstrap samples. Therefore we conclude that all differences in the distribution of the "maximal daily change in conductivity" between the two years **may be the result of sampling error alone**.

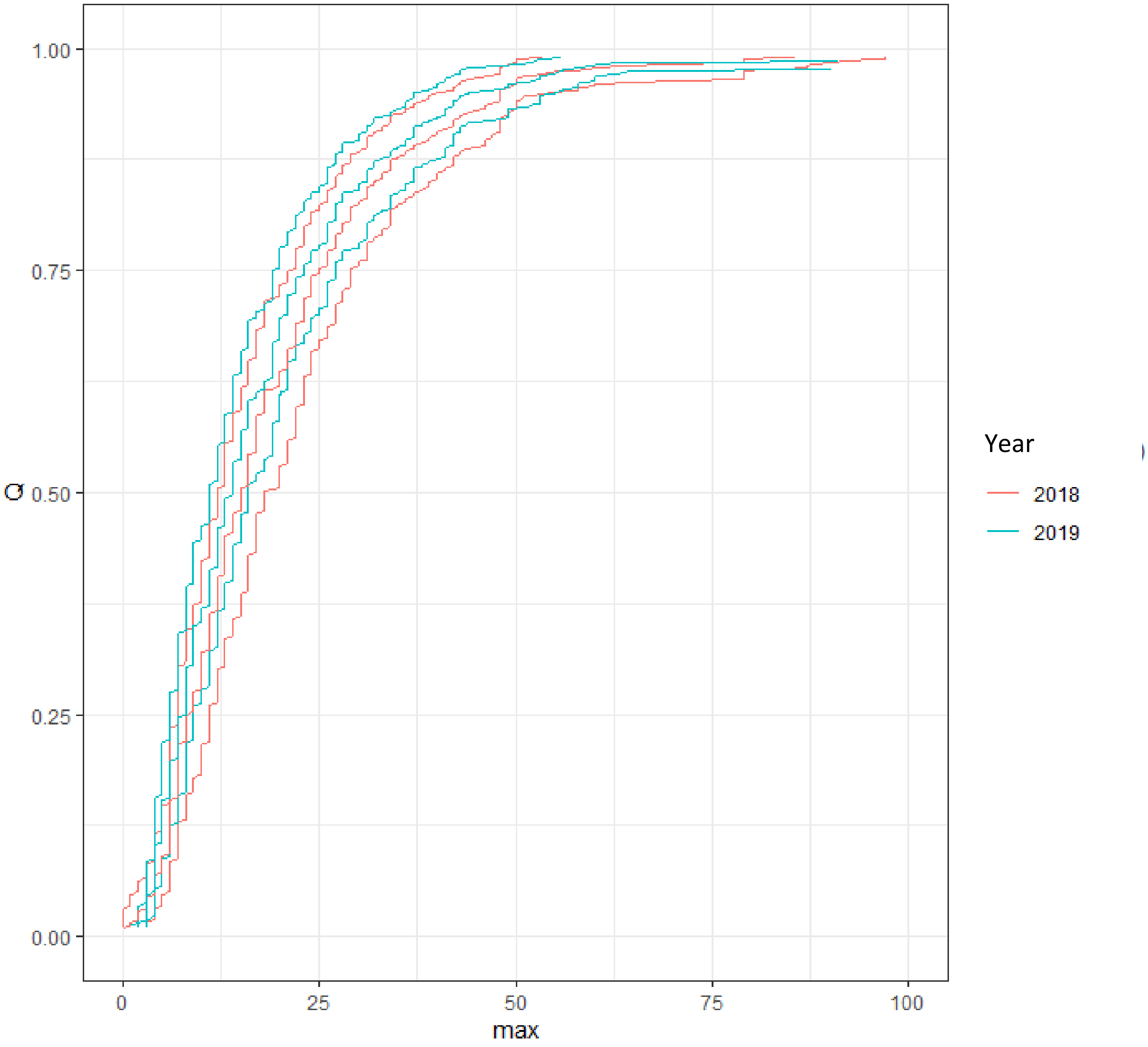


Figure 31: Comparison of the two CDF distributions of Figure 30. We see a large overlap between the two sets of bootstrap samples, indicating all differences in "maximal daily change in conductivity" between the two years may be the result of sampling error alone.

2.5 Prediction Models and model error distributions

Prediction models are used in statistics to associate changes in a dependent variable, usually denoted by Y, with a dependent variable, usually denoted by X. Different statistical models have different pros and cons – some of the models are very simplistic, and therefore allow higher interpretability of model behavior, while others are more complex, at the price of reduced interpretability.

The Spatial Model methodology, presented in Unit B of this document, will make use of Prediction models to predict water quality measurements at a target station, based on the equivalent water quality measurements at a source station found upstream. The models presented in unit B will account for the (varying) flow time between stations, along with adjustments that need to be made for specific water quality data. Since Section 3.5 in unit B will focus on adapting existing prediction methodologies for use in the Spatial Model framework, we found it best to include a subsection reviewing these prediction methods.

In this chapter, we will review three models:

- The "Constant Shift" model – assuming measurement in Y is similar in distribution to the measurements of X, with a shift in distribution values.
- The linear regression model – assuming a linear relation between X and Y.
- The kernel regression model – predicting Y given X by observing the behavior of Y for data points with similar values of X.

For each model, we will provide the model equation, explain how prediction is calculated, and briefly discuss the pros and cons for each model. At first, the models will be presented on a toy dataset, for better clarity of model characteristics. Next, we will review 4 data examples taken from water quality analyzers, placed at different locations in a water distribution network. For these examples, we will compare the prediction given by the three different models, along with a graphical representation of the distribution of prediction errors. Eventually, using the error distribution as a yardstick for abnormality of measurements will allow us to detect irregular water quality events in Section 3.5 in Unit B.

Example 2.5A – toy dataset for demonsrating different prediction models: Figure 32 shows the dataset used to exemplify the three types of models mentioned before. The dataset consists of 2000 points.

On the next pages, we will look at the three different models fitted to this data.

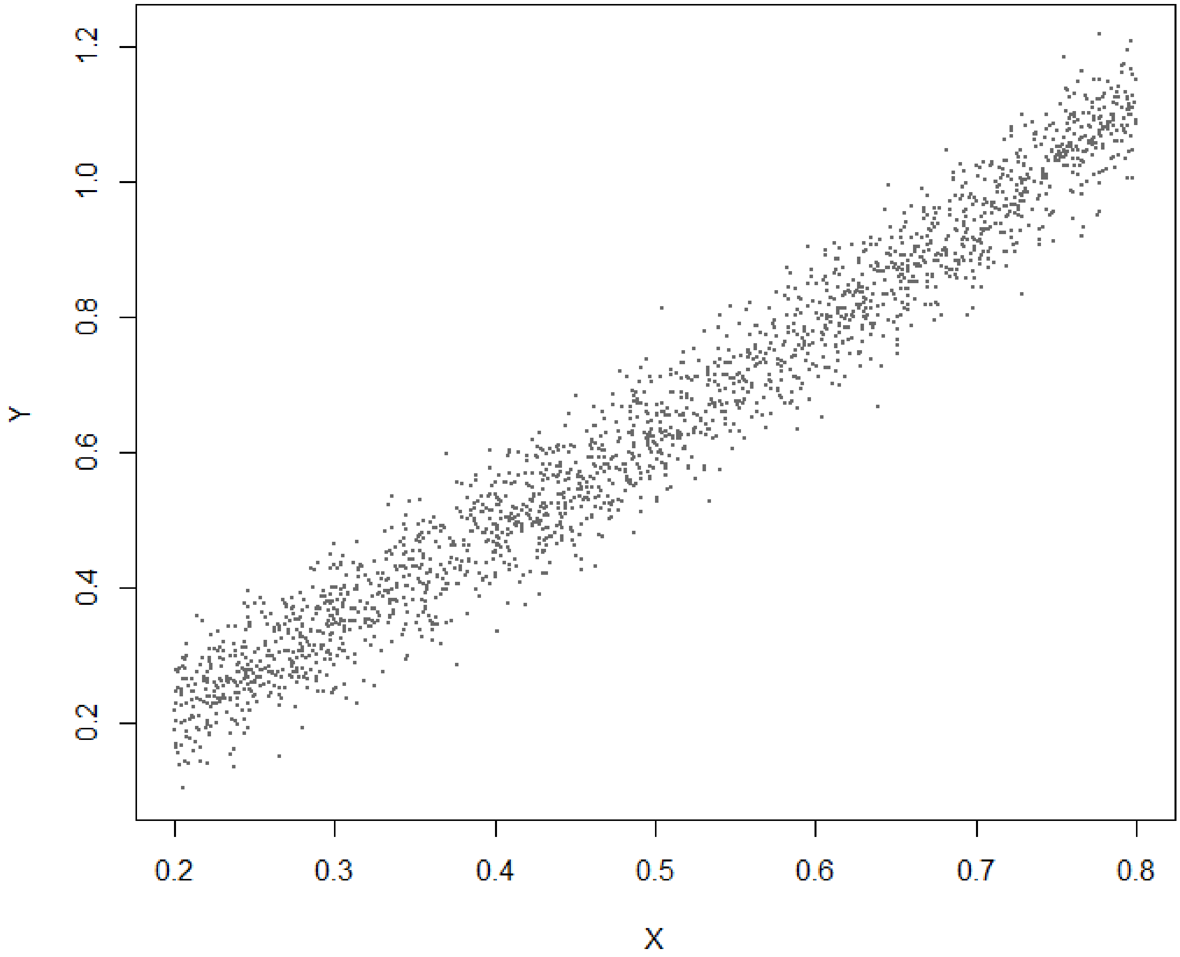


Figure 32: Toy dataset for demonstrating different prediction models.

The constant shift model: Figure 33 shows the constant shift model fitted to this data. The model equation is assumed to be:

$$Y = a + X,$$

i.e., the distribution of Y's is a shifted version of the distribution of the X's. Such models may be appropriate for conductivity measurements between two measurement locations when there is no attenuation of measurement values across the water distribution network. For measurements like Total and Free CL, or turbidity, it is not reasonable to assume a slope of **1** between two measurement locations. Moreover, sensors for these measurements must be calibrated on-site, leading to different "response slopes" for the same values of T-CL, F-CL, and Turbidity in the water. The parameter a is fitted in a robust manner, using

$$estimated\ a = median(Y - X),$$

i.e. taking the median value of $Y - X$ across all data points. In the example below, we reach $a = 0.121$. We see a very poor fit between the model given by the red line and the data.

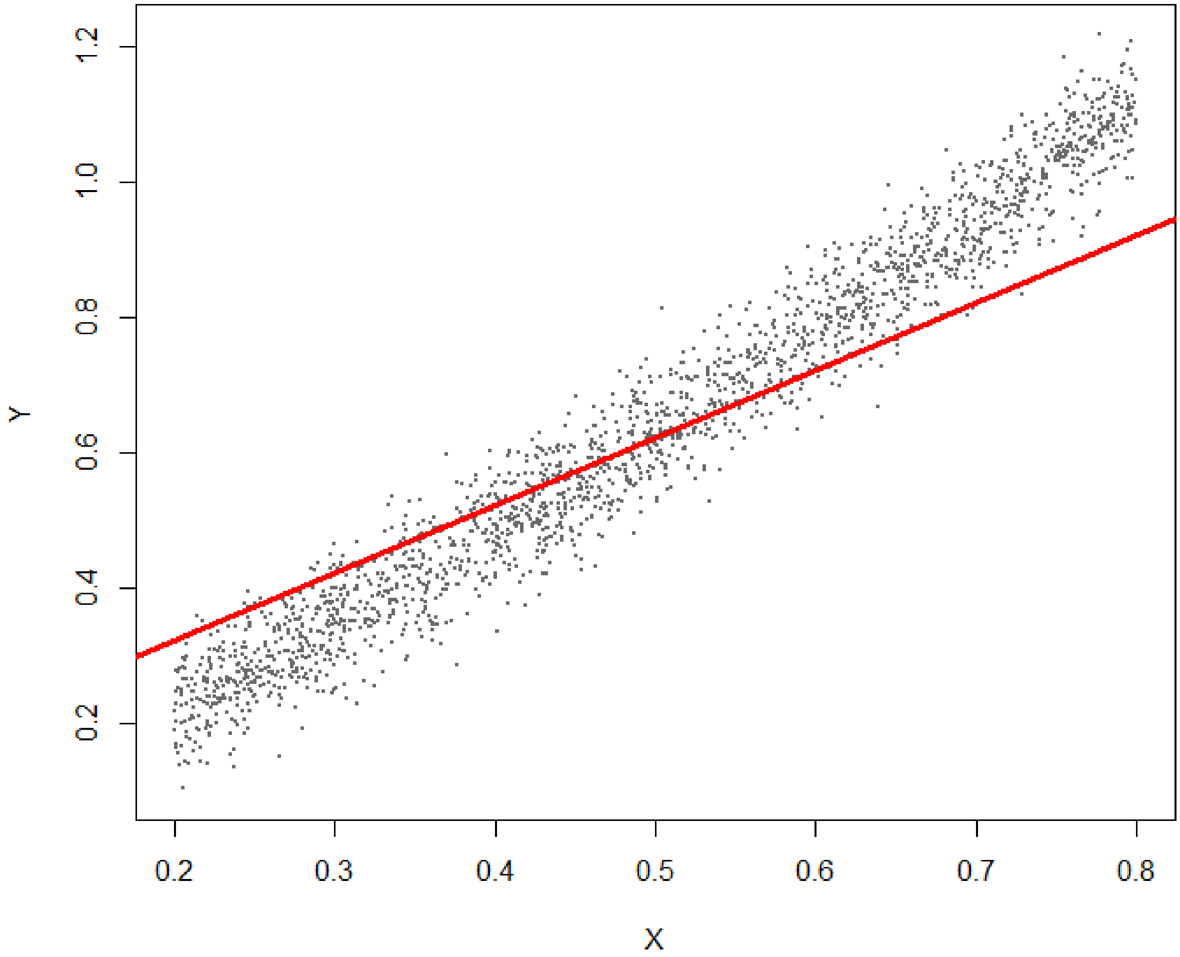


Figure 33: The "constant shift" model exemplified on the toy data example

Figure 34 shows a linear regression model fitted to our example dataset. The linear regression model equation is given by:

$$Y = a + bX.$$

The parameters a and b are estimated by minimizing the sum of squared distances between points and their predictions on the green line. In other words, out of all possible green lines that could be fitted to the data, the green line presented in Figure 34 is the one where the sum of square vertical distances between data points and their corresponding points one the line (for the same value of X), is the smallest. The Fitted line for this data is $Y = -0.1065 \; + \; 1.4938X$. Graphically, the model fits the data much better, and we observe a slipe closer to 1.5 is more appropriate than 1. However, we can still see a slight non-linearity in the data – for the highest and lowest values of X, the model usually predicts lower values than expected.

There are several ways to extend the linear model to complex relations in data: some possible estimation parameters may be robust to outliers, considering the absolute value of errors, instead of square errors; other models may consider polynomial model equations such as $Y = a + bX + cX^2$. To allow modeling of complex relationships in data, the SM uses non-parametric kernel regression, explained on the next page

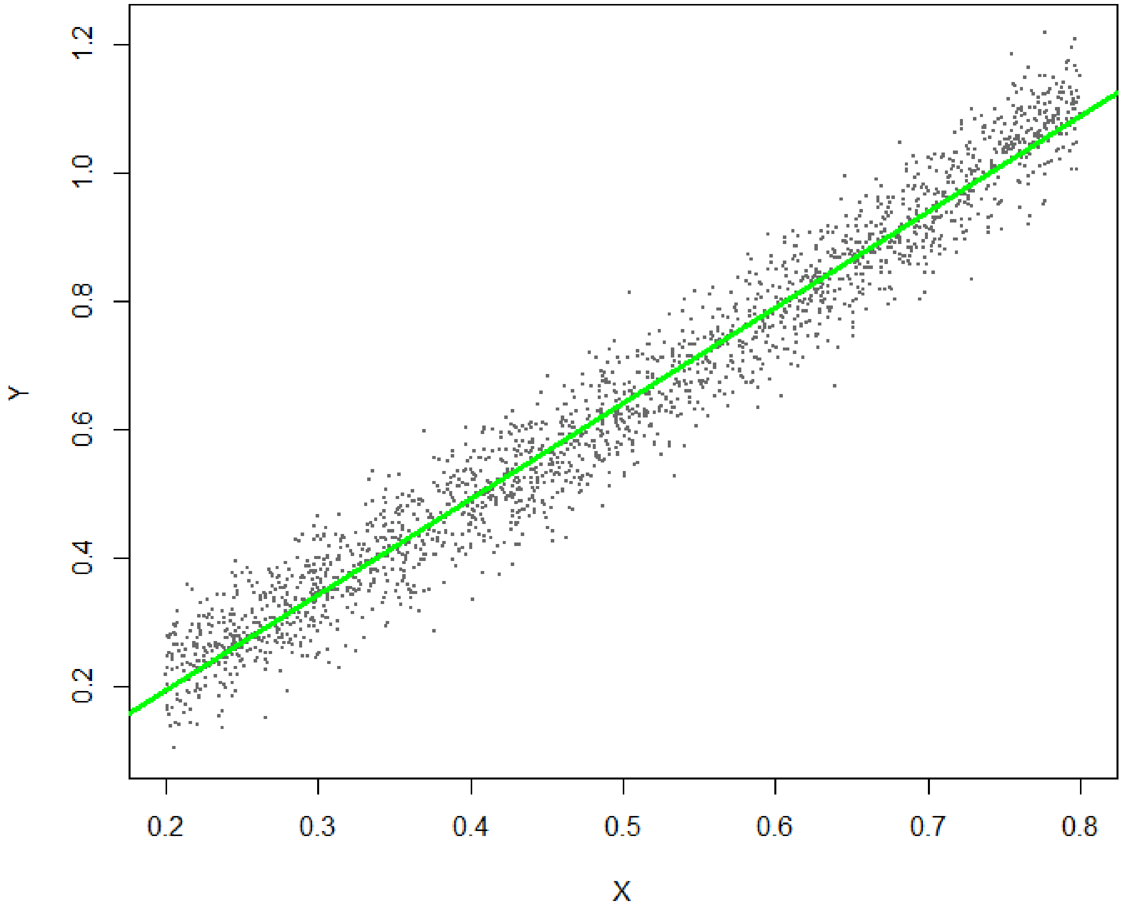


Figure 34: The linear regression model exemplified on the toy data example

A non-parametric model is a statistical model that cannot be put in a closed equation form, with a limited number of parameters. For comparison, the last two models we reviewed were parametric, since their model equation could be summarized by either one or two parameters. **A kernel regression** is a nonparametric regression model, predicting a value for Y, the dependent variable, using the value X, and a selected set of historical data points with X values similar to the one used for prediction.

Consider the setting where we are given observations $(X_i, Y_i), i = 1, \dots, N$ and want to predict Y' for a new observation with a given $X = X'$. We select some domain $D(X')$, a valid range over the X's, and compute:

$$Prediction\ of\ Y' = \frac{\sum_{i=1}^{N} I(X_i \in D(X')Y_i}{\sum_{i=1}^{N} I(X_i \in D(X')},$$

where $I(X_i \in D(X')$ is an indicator function receiving the value 1 if X_i is in the domain selected around X', and 0 otherwise. The numerator for the above expression sums overall Y's that are associated with Xs in the domain around X'. The denominator counts how many observations are summed over in the numerator. Therefore, the above expression is just a local average of Y's, for all point with X's in the domain around X' (as defined by the domain $D(X')$). Defining the domain $D(X')$ is application and data specific. For rich data sets, with complex signals, the domain $D(X')$ might be very local in terms of X's. For small datasets, $D(X')$ might consider up to 20-30% of the data, at the cost of simplistic modeling.

The indicator function $I\big(X_i \in D(X')\big)$ can also be replaced by a kernel function, $K(X_i, X')$, receiving values greater or equal to zero. The model equation then becomes:

$$Prediction\ of\ Y' = \frac{\sum_{i=1}^{N} K(X_i, X')Y_i}{\sum_{i=1}^{N} K(X_i, X')}.$$

The idea is to replace the unweighted average over the domain including $D(X')$ by a distance specific weight function. Again, selecting the right kernel function for a model is data and domain-specific. Some application may even require that the kernel function be non-homogenous in space, i.e., rely on the values of X' and X_i explicitly, and not just through their distance, $|X_i - X'|$.

In the setting encountered by the Spatial Model, we are interested in constructing many models with thousands of observations in each. Storing all tens of thousands of data points, for hundreds of models is inefficient. Therefore, before model construction, we aggregate observations to bins by X coordinate, as explained on the next page.

SM Kernel regression, Step I: Kernel regression in the Spatial Model begins by binning observations to bins, by X values, as exemplified in Figure 35. Different horizontal bins are represented by the white spaces between gray vertical lines, with data points being divided across the different bins. For each bin, we compute the median Y value for all observations in the bin. The median values of Y in each bin represented by red lines in Figure 35.

Let $\left(\tilde{X}_j, \tilde{Y}_j, \tilde{N}_j\right), j = 1, \ldots, B$ represent the X value of the jth bin (by the center point falling in the pan of the bin), the median value of Y's for observations in the bin, and the number of observations per bin, for bins 1,…,B. In Figure 35, B is set to 60. Step II, on the next page, step II of the algorithm will aggregate over bin medians to form predictions.

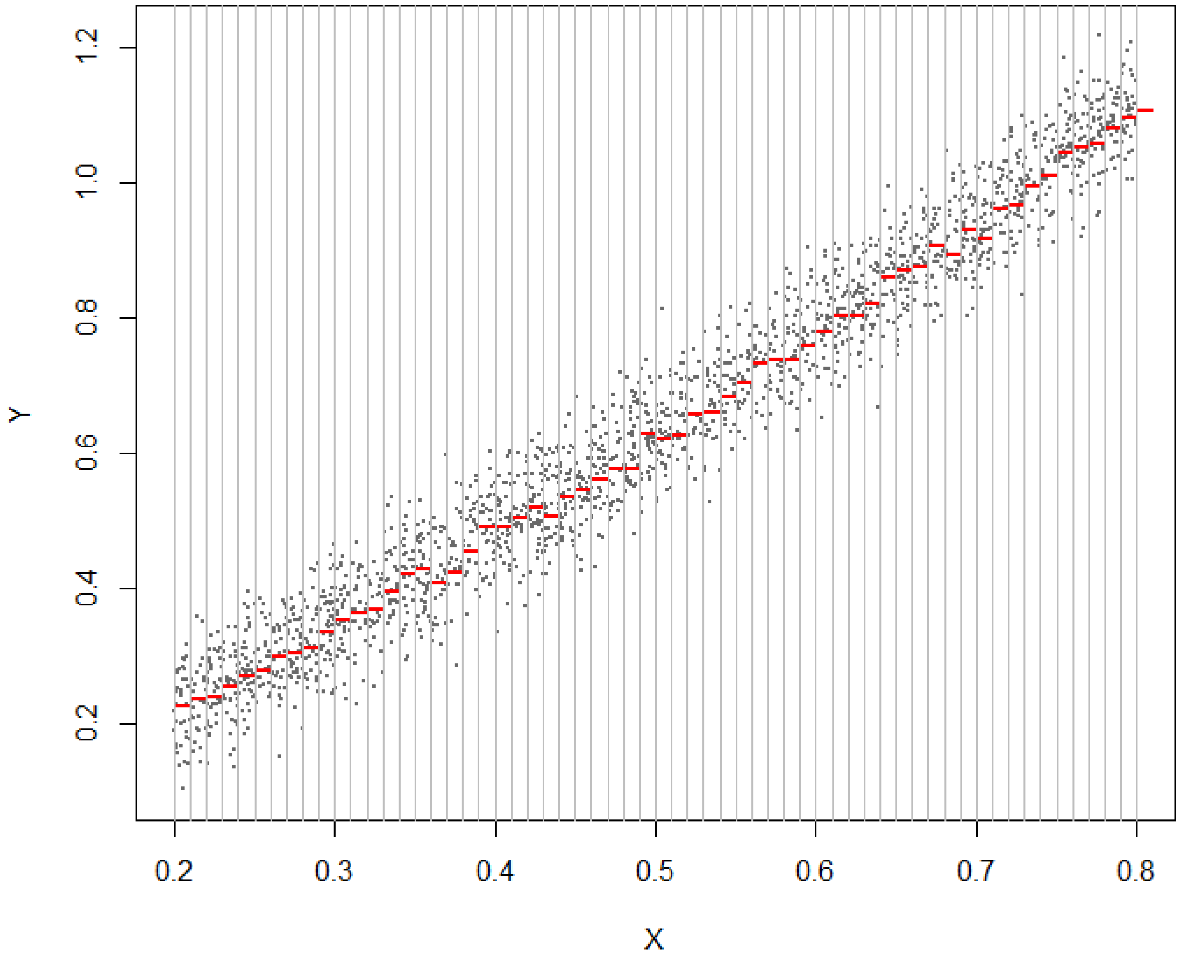


Figure 35: Kernel regression, with binning over X's, step I. Bin median values are represented by red lines in each bin.

SM Kernel regression, Step II: The kernel regression model equation is given by the formula:

$$Prediction\ of\ Y' = \frac{\sum_{j=1}^{N} K(\tilde{X}_j, X')\tilde{Y}_j}{\sum_{j=1}^{N} K(\tilde{X}_j, X')},$$

$$where\ K(\tilde{X}_j, X') = \begin{cases} N_j, & if\ \tilde{X}_j\ \mathbf{is}\ in\ the\ 5\%\ bins\ closest\ to\ X' \\ 0, & if\ \tilde{X}_j\ \mathbf{isn't}\ in\ the\ 5\%\ bins\ closest\ to\ X' \end{cases}.$$

This model equation predicts Y' by taking a weighted average of the bin representative Y values, of the 5% of bins closest to X' . The number of bins, set to 60, and the threshold for proximity, set to 5%, are model parameters to finetuned to the data at hand. Figure 36 shows model predictions for the data. The model captures the data well, and predictions are not biased for extreme values of X, as in Figure 34.

Sections 2.5.1-2.5.4 provide interesting case studies comparing the three model types for real data. In Figures 33-36, we maintain a color-coding where red curves are for Constant Shift models, green curves are for linear regression, and blue lines are for kernel regression. This color-coding is kept throughout the next subsections.

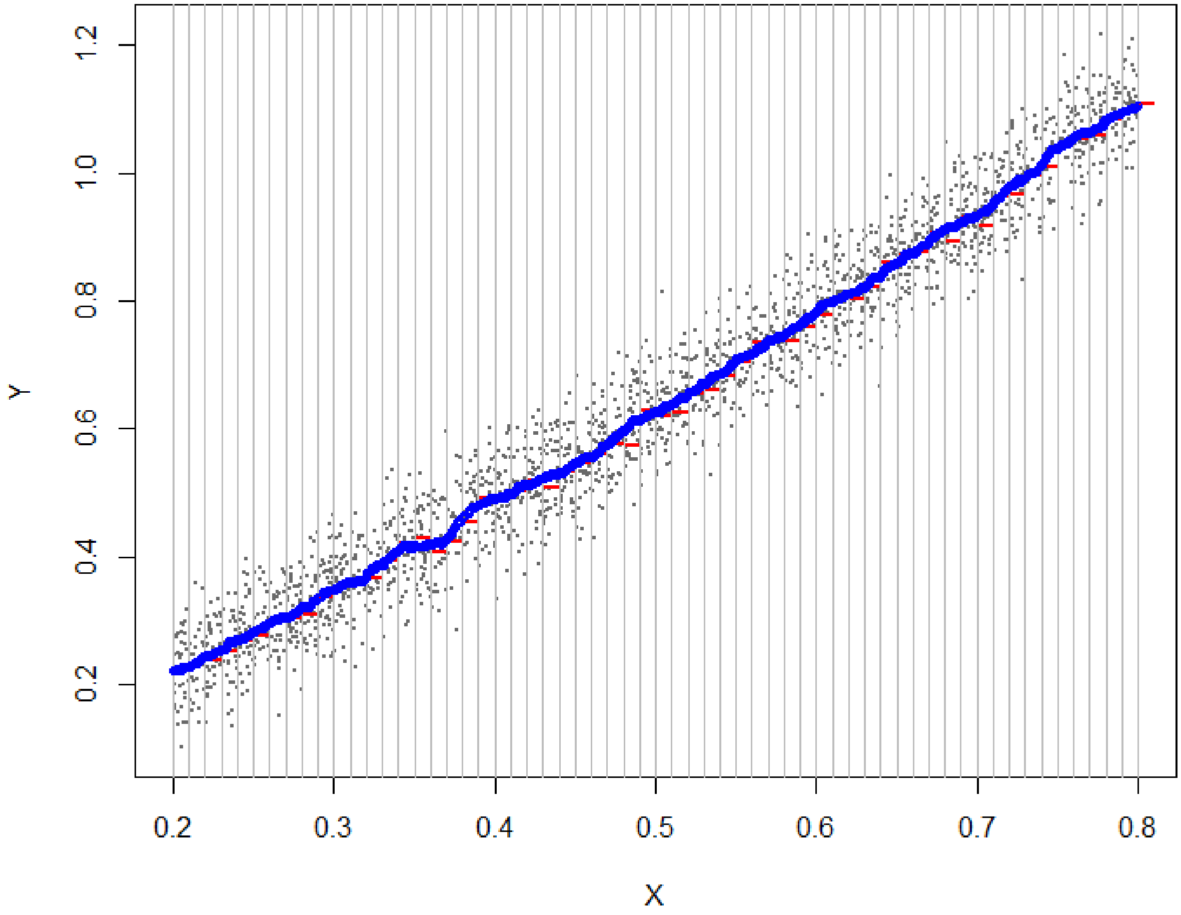


Figure 36: Kernel regression, with binning over X's, step II. Model predictions by weighted averaging of the 5% closest bins are given by the blue curve.

2.5.1 Example for good model fit, some outliers

In this example, we fit the Constant Shift, linear regression, and kernel regression models to a dataset comparing Turbidity measurements at a source and target station, with water flowing between the two stations. The data used for analysis is for a time span of 20 days, with one measurement every 10 minutes. Datapoint coordinates in Figure 37 are the hourly median values in the two stations for the same calendrical hour. Figure 37 shows the three models presented in this section, fitted to the data. For this data, since data points are themselves hourly medians, no additional binning at the kernel regression level was performed. We observe all three models to fit the data well, with a slope of 1 (in the constant shift model) describing the ratio of turbidity measurements between the two stations. Some of the measurements at $X = 0.41$ and $X = 0.36$ (marked by dashed ellipses) are outliers in Y values, representing increased turbidity measurements at the target station. However, despite being outliers, these points do not jeopardize model validity, as seen by the green line for the linear model. On the next page, we discuss the error distributions for these three models.

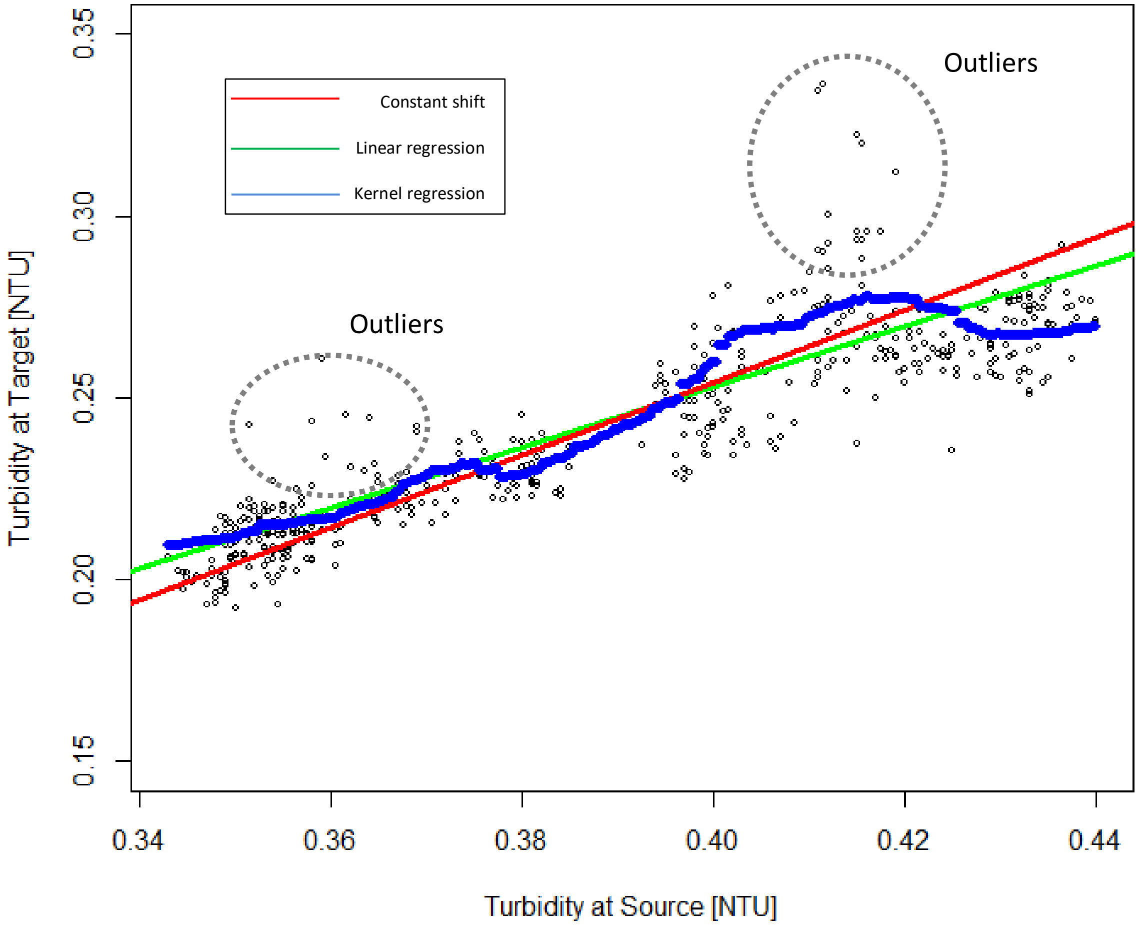


Figure 37: Three valid models for Turbidity at a source and target station.

Figure 38 shows the (smoothed) error distribution for the three models presented in Figure 37. All three models have error distributions in the range $[-0.03, 0.03]$ NTU, excluding some outliers. Of the three models, The constant shift model fates slightly worse than the other two.

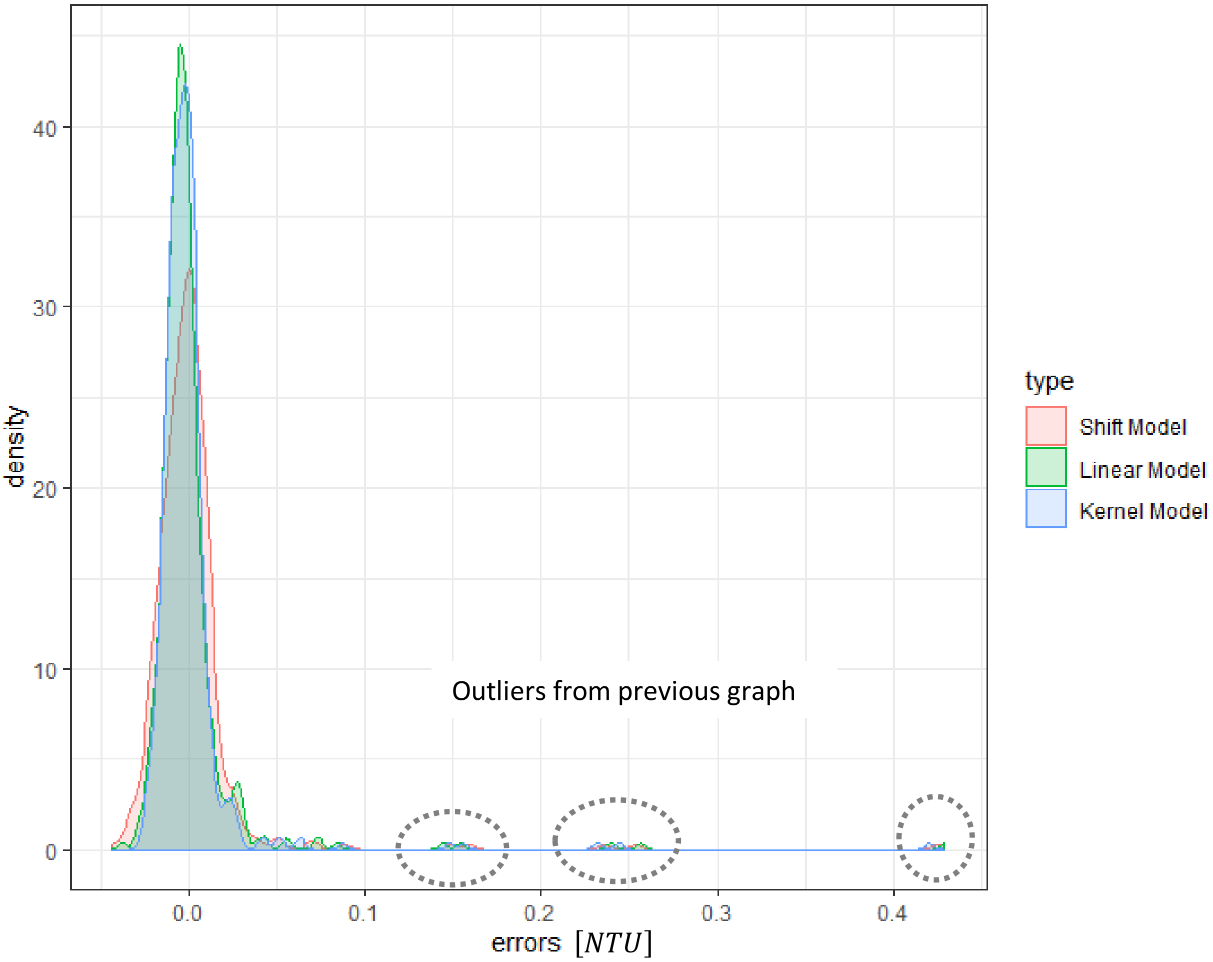


Figure 38: Error distributions for the three models of Figure 37.

2.5.2 Example for a non-linear relationship in the data

Figure 39 shows data for turbidity measurements over a timespan of 20 days, with a similar sampling resolution and preprocessing as in Section 2.5.1. Again, the data points compare turbidity measurements from a source and target location, with water flowing between the two locations. Unlike Figure 37, which features a linear relationship between the two stations, Figure 39 shows a highly non-linear relation for concurrent turbidity measurements between the two stations. The red line, showing the fitted "Constants Shift model" misses the data almost entirely. The green line, showing the fitted linear regression model, gives biased predictions for the low and high values of turbidity at the source station, as exemplified in the toy example of Figure 34. Only the kernel regression can fit the data correctly. The error distributions are compared and discussed on the next page.

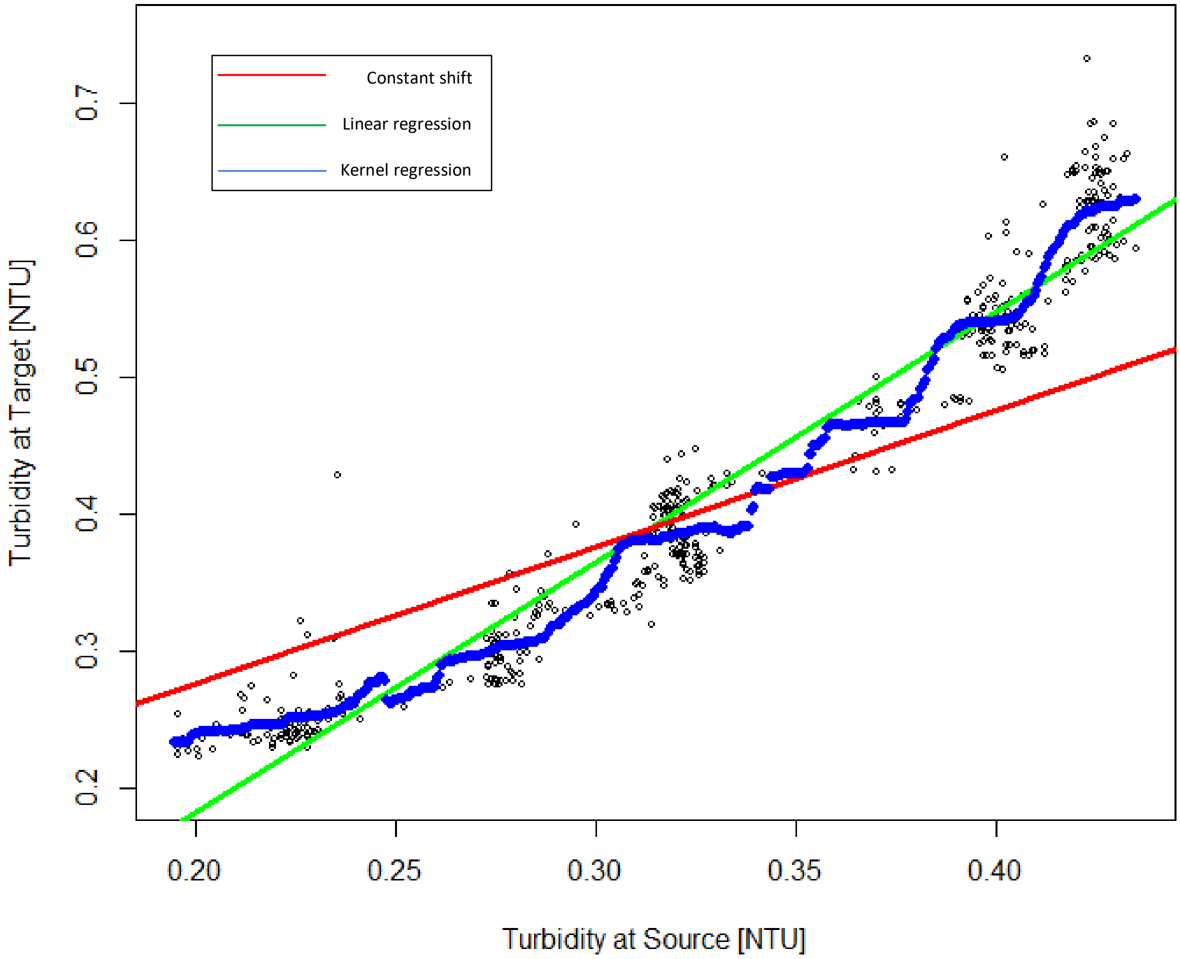


Figure 39: Three models for turbidity, with data showing a non-linear relation between source and target stations.

Figure 40 shows the error distributions for the three models. The majority of the mass for the error of the kernel regression model is found in the $[-0.03, 0.03]$ range. Other models exhibit wider distributions (larger prediction errors), which are also highly skewed, with a right tail. Only the kernel regression model provides an adequate fit to the data.

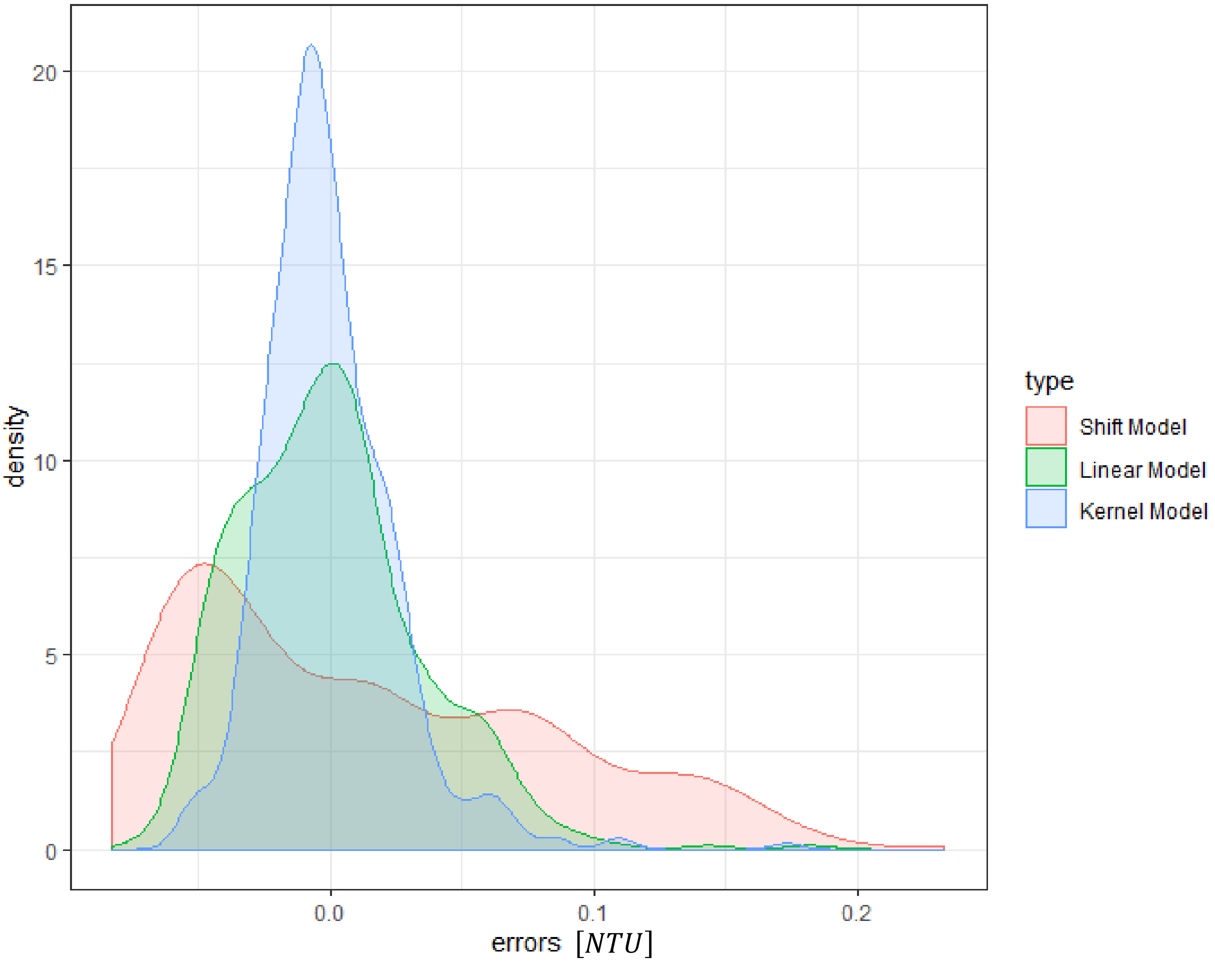


Figure 40: Error distributions for the three models of Figure 39.

2.5.3 Examples for bad model fit, when using data gathered over a long period

We examine Free CL measurements data collected over 604 days, at 10 minutes intervals, for two stations called Source and Target. In this setting, water flows from the source station to the Target station. Point Coordinates in Figure 41 represent hourly medians of measurements for the same hour at both stations. The data presented in Figure 41 does not show a clear correlation pattern between the two stations. Attempts to fit the three previously discussed models to the data lead to significantly different results. The constant shift model, forced to use a slope of "1" passes through the center of the data point cloud; however, there is little evidence in the data to suggest a positive correlation between measurements gathered at the two sites. The linear regression and kernel regression methods show a slight negative correlation between the two sites: this is not a likely result, given the layout of the water distribution network. We proceed to examine the error distributions of the different distributions on the next page.

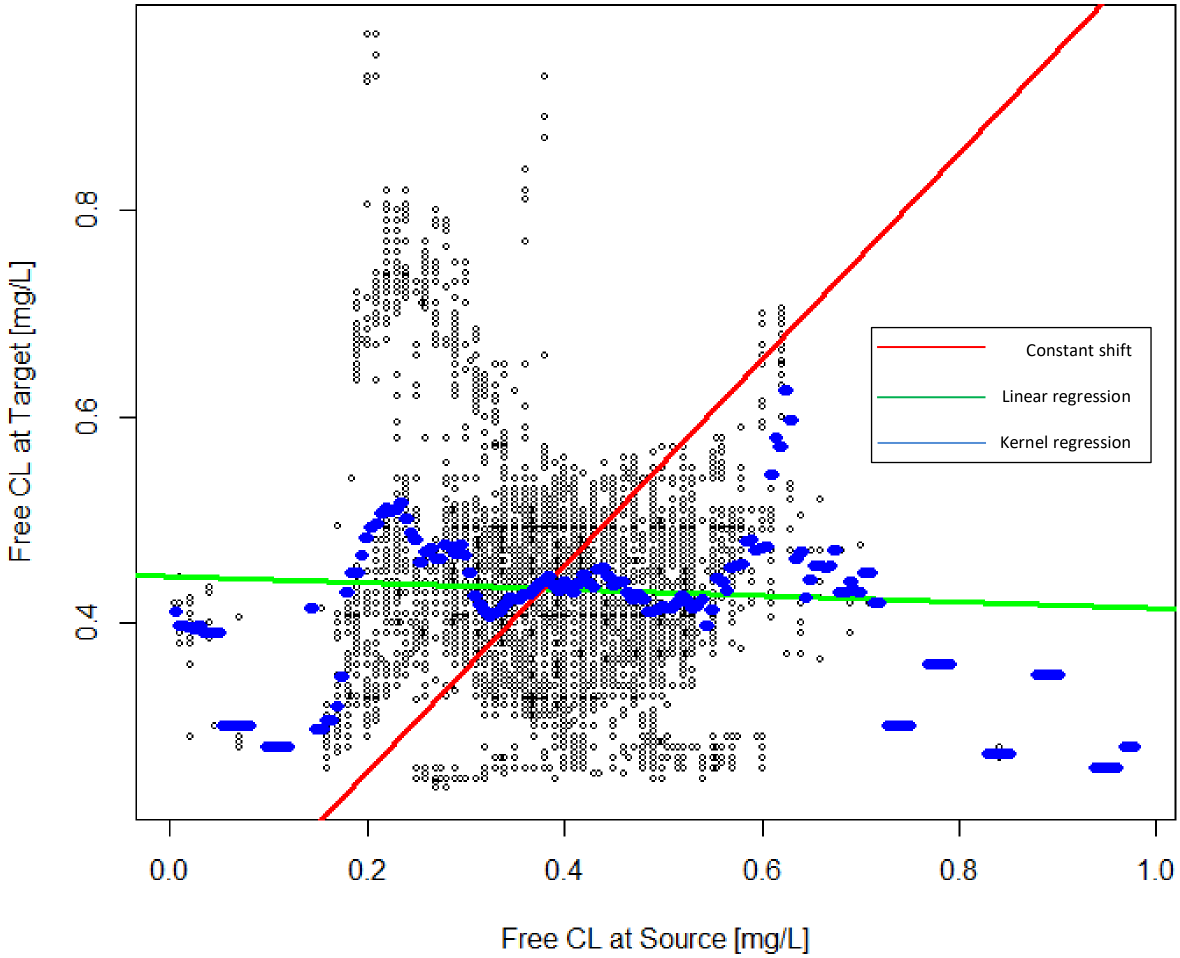


Figure 41: Comparing prediction models for Free CL measurements at a source and target station, over 604 days.

Figure 42 shows the error distributions for the three models. All three models have the majority of prediction errors in the range $[-0.2,0.2]$ mg/L. Using Figure 42, one may erroneously conclude that all three models perform "equally well", when in fact the three models perform "equally worse": a span of 0.4 mg/L encompasses almost the entire range of valid measurements on the Y axis of Figure 41, meaning the models presented in both figures perform no better than a "guess" of the Y value.

The prediction model using a large time period for training rely on data collected with differing source water quality, sensor calibrations, and seasonal effects. Simplistic models, such as predicting the target site water quality measurement value by the corresponding source site value, cannot account for these effects without additional effort in modeling. While one may seek to extend the simplistic models presented in this chapter to account for additional variables, we prefer instead to rely on models trained on datasets with a shorter time period, and fewer covariates. These models are easier to analyze and diagnose. The next subsection will review these two sites again, with models trained using 60 days of data alone.

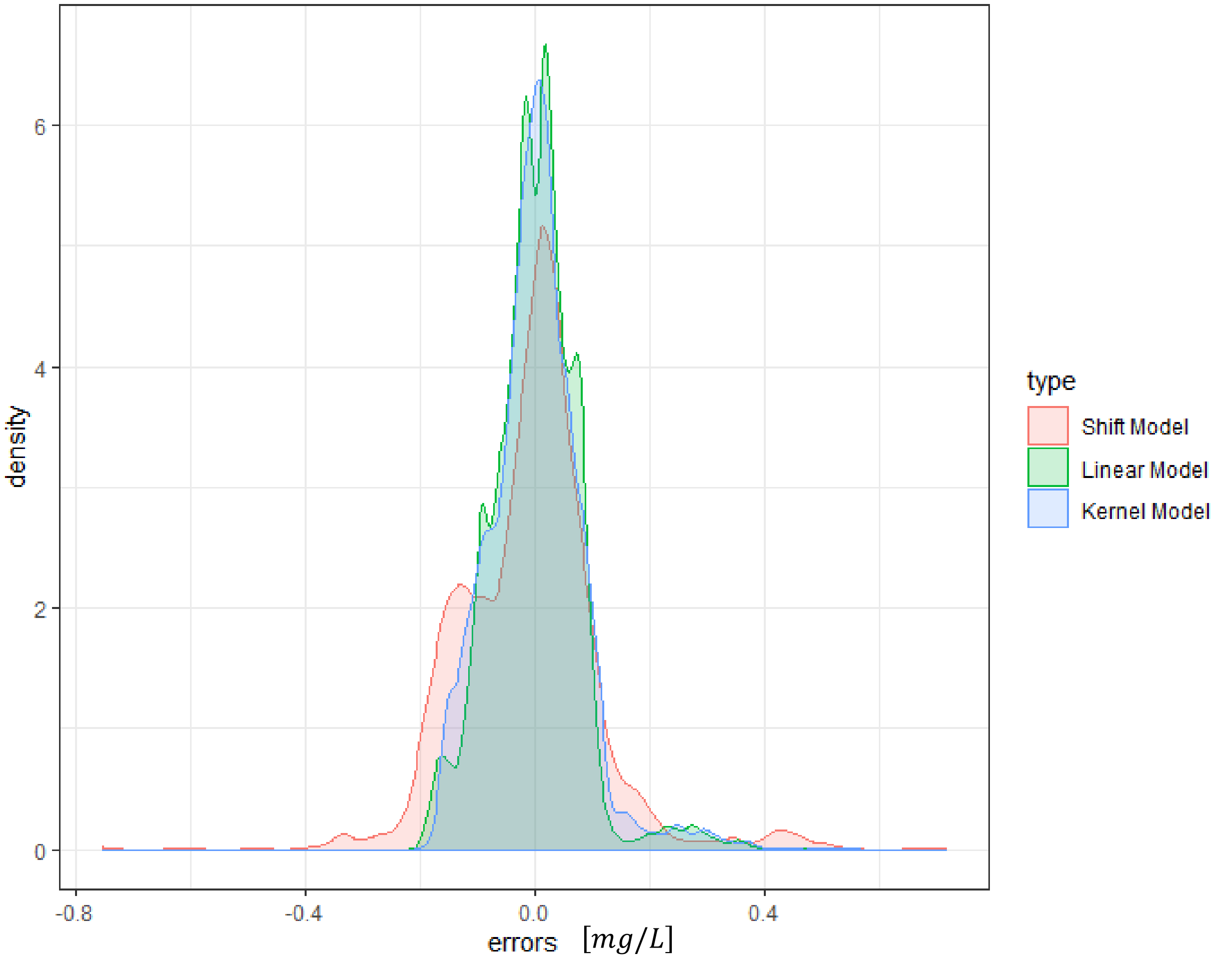


Figure 42: Error distributions for the three models of Figure 41.

2.5.4 An example resolving the issue raised in 2.5.3

Figure 43 shows Free-CL measurements for a period of 60 days, for the two stations described in Section 2.5.3. We observe a clear positive association in the data captured by all three models. The linear and kernel regression models relatively agree on the predicted Y values, across the range of possible X values. On the other hand, the 'Constant Shift' forced to use a slope of 1 model does not fit the data well: chlorine dissipation is linear, when comparing the two stations, but not with a 1:1 ratio. Figure 44 on the next page shows the 'Constant Shift' model to have biased predictions, with an error distribution that is not centered around 0.

The problem presented in Section 2.5.3 has been partially solved by reducing the time period of data available for model training. In Unit's B and C we will show better prediction results can be achieved with training datasets spanning 14-30 days.

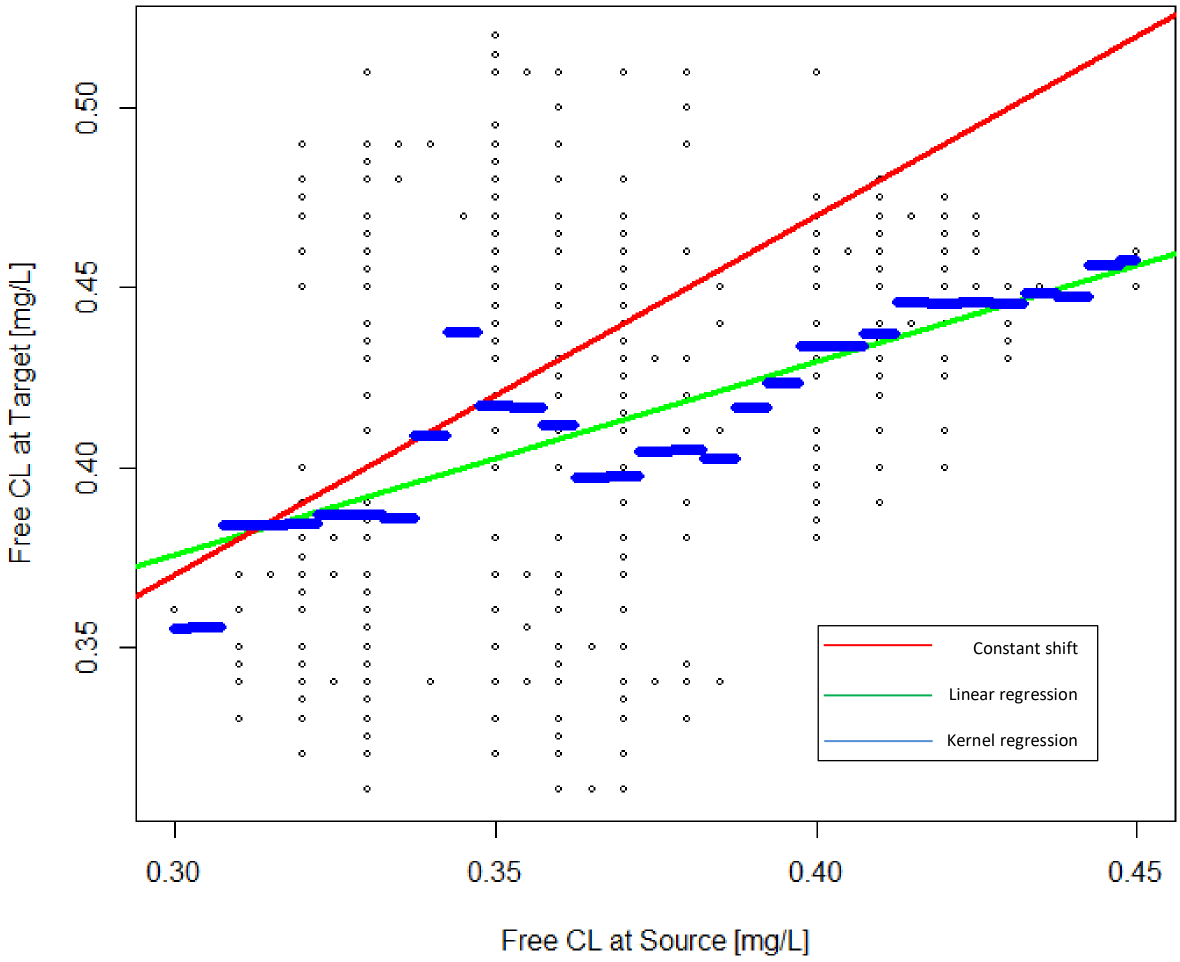


Figure 43: Free CL measurements and models for the two stations described in Section 2.5.4, over a period of 60 days.

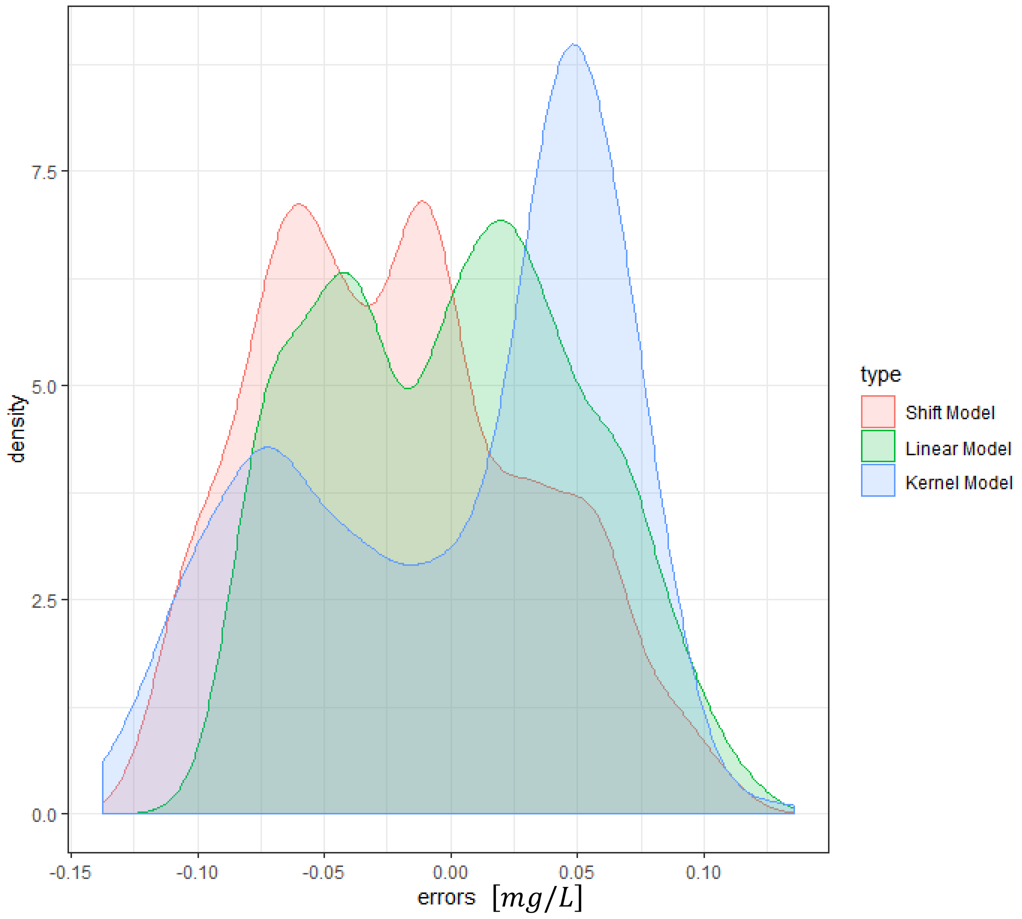


Figure 44: Error distributions for the three models of Figure 43.

2.6 Quiz for Unit A

1. What is the ACF?

2. Why should measurements be binned when analyzing time series? When is binning mandatory?

3. Consider a water measurement time series with no binning of measurements. If the sampling resolution were to increase, how would the following values change:
   a. ACF(shift = 0)
   b. ACF(shift = 1 single measurement)
   c. ACF(shift = a number of measurement equivalent to 1 hour)

4. What is the CCF?

5. Why is the shift argument value for which the CCF receives it's maximum interesting? How can this value be interpreted?

6. What possible types of noise are observed in water quality data, and how are they handled? Name three types of noise in the data, and way to counter them.

7. Except Pearson's correlation coefficient, we learned two other correlation measures. How are these correlation measures called? Why are these measures preferred for some data types?

8. Compute Kendall's Tau for the following sample of data points:
$$\{(1,1),(2,2),(3,4),(4,5),(5,3)\}$$

9. Can the bootstrap sample $1,2,4,0,5$ be obtained from the original data: $1,2,3,4,5$?

10. Our water utility engineer, named Ahab, computed an ACF for a timeseries of conductivity measurements. The time series was measured for 1 Year, in a 10 minute resolution, at some location. Aaron would like to estimate the sampling error for the values of the ACF. In order to do so, he plans to sample 52560 measurements (the equivalent of 1 year) from his original data series. Aaron has the following claim: "I can estimate the sampling error of my ACF, across all possible shifts, using the bootstrap samples produced in this method!". Jezebel, also working at the water utility: "You cannot use the bootstrap technique to estimate the value of the ACF at any possible shift!".Is Ahab correct? Is Jezebel Correct?

11. What types of prediction models were discussed? What are their Pros and Cons?

12. What is the prediction model's error distribution used for?

3. Unit B – Algorithmic and Statistical Methodology in the Spatial Model.

Unit B discuss the statistical methodology for the Spatial Model. Section 3.1 provides an introduction to the SM methodology, and motivates the construction of Statistical models involving two or more water network locations. Sections 3.2-3.5 discuss pair models in depth, with the different subsections focusing on flow time estimation, selection of flow regimes, error estimation for flow time estimations and predictions for of water quality measurements, respectively. Sections 3.6-3.7 discuss models for triplets of stations. Section 3.8 recaps up on the complete methodology.

3.1 Introduction to the Spatial Model.

This subsection motivates the use of flow time estimators, construction of statistical models for the relation in water quality measurements between pairs of stations, and introduces the idea of models for triplets of stations.

Water quality sensors in a water distribution network are generally found in key location in the network, with flowing between network locations, and several types of measurements gathered at each network locations. An illustration of such a water distribution network is shown in Figure 45.

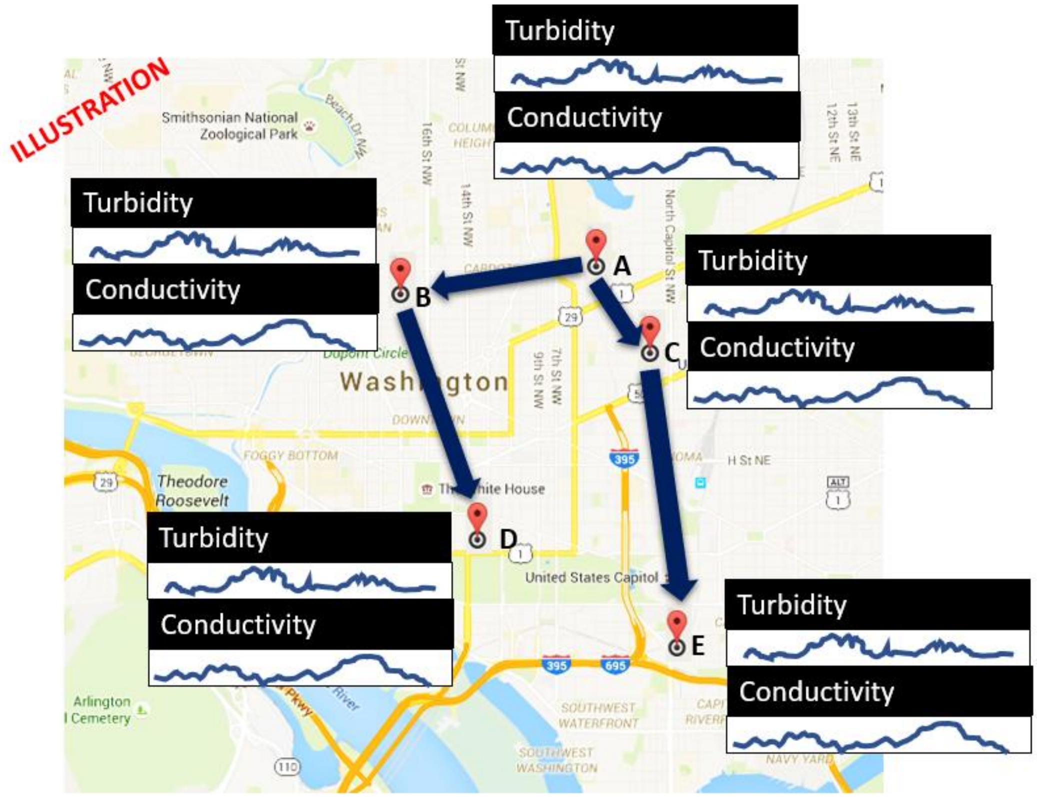


Figure 45: An illustration of water quality analyzers located across a city. The city has 5 measurement locations named A-E, with time series for conductivity and turbidity measurements depicted above each measurement station. Blue arrows depict the general direction of water flow between the different stations. The purpose of the Spatial Model is to model the relation between water quality measurements from the same type, across the different stations.

The construction of spatial models being with modeling the relationship between the measured time series in two connected network nodes, as shown in Figure 46. By modeling the different pairs of sensors in the network, we aim to provide good coverage for water quality events.

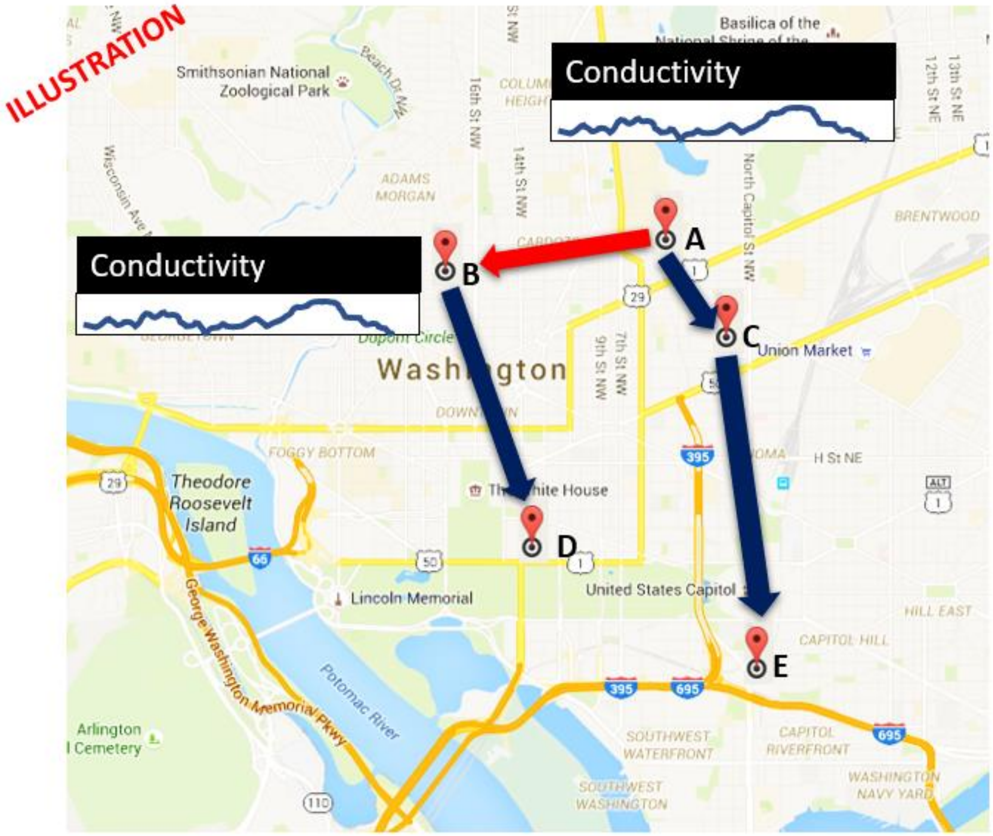


Figure 46:Spatial Model pair model focus on modeling one pair of measurement locations, with the same measurement type at a time. In this figure, we illustrate this concept by focusing on a specific pair of connected stations (or link), out of the different links presented in Figure 48.

Our working example for this section, depicted in Figure 47, shows conductivity measurements for 16723 hours, measured in two network locations. The two locations are connected, and water flows from the source station (time series shown in black) to the target station (time series shown in red). The two stations show substantial variability in terms of conductivity measurements, with measurements varying between 350 to 800 $\mu S/cm$. While each station shows substantial variability, the two stations closely resemble one another, with the 'red' curve following in motion after the 'black curve'. Consider a water engineer monitoring the Conductivity measurements for the target station. Looking at the 'red' curve alone, the engineer might be 'surprised' if a the 'red' curve is susceptible to abrupt changes. However, if the black curve is known, changes in the 'red' curve can be foreseen, and the water engineer knows changes in the red curve are not due to a water quality event happening in the network between the two stations

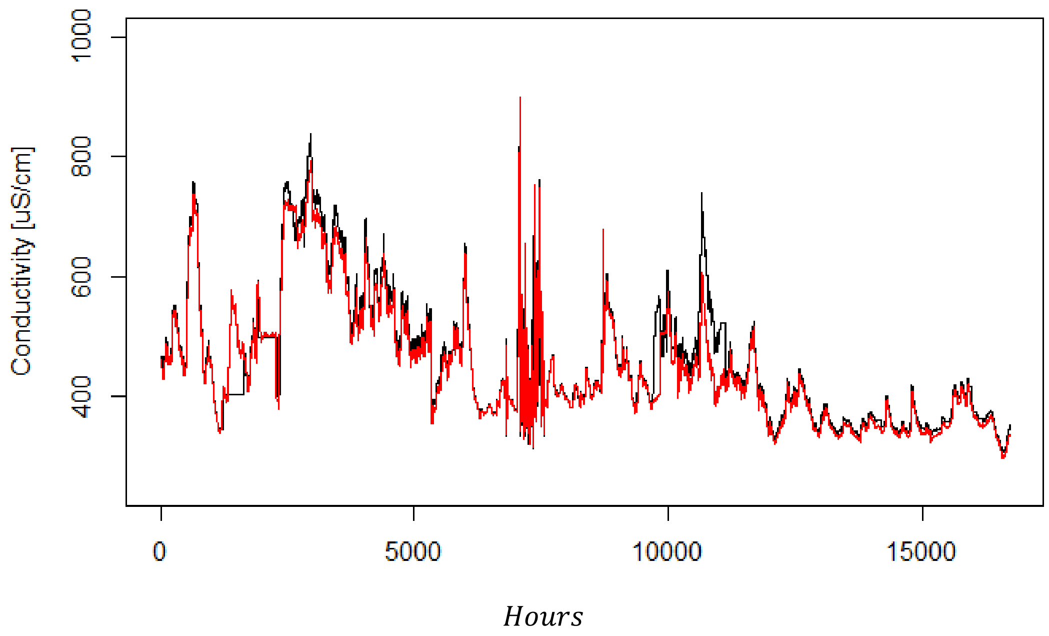


*Figure 47: Conductivity measurements across 16723 hours *697 days), in two measurement locations – a source station (black) and a target station (red). Water flows from the source station to the target station.*

Figure 48 shows the CCF for the two series depicted in Figure 47. The CCF shows a distinct peak at a lag of three hours, hinting the flow time between stations. For the purpose of this example, we consider the flow time between the two stations to be three hours and the flow time to be constant across all times. Next, we show how the flow time estimates can be used to improve predictions of water quality measurements at the target node.

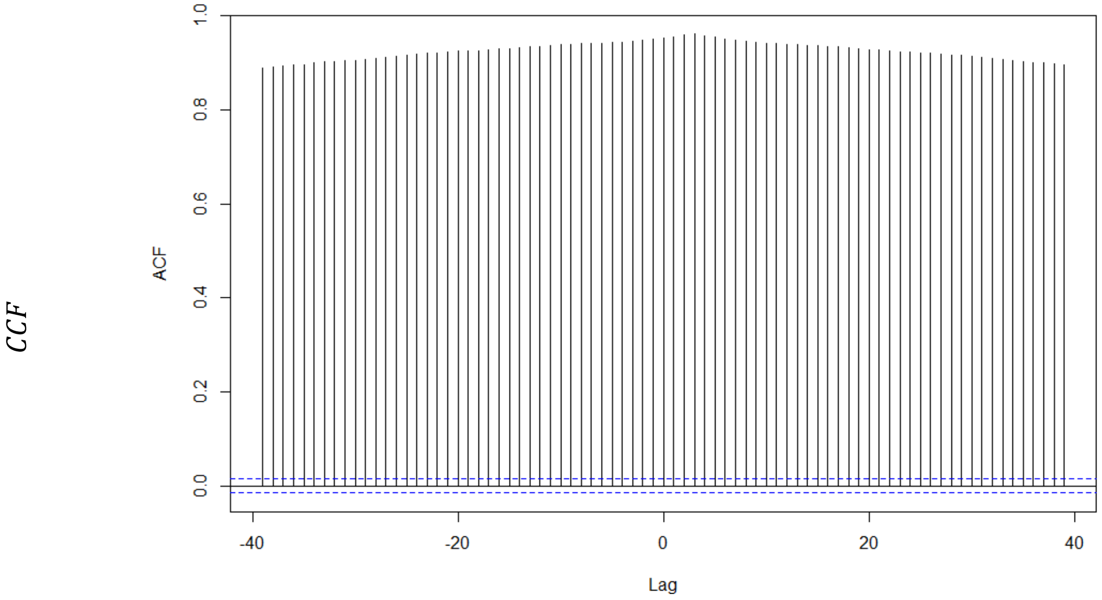


Figure 48:The CCF for the two conductivity time series presented in Figure 47. The CCF shows a distinct maximum at LAG = 3 Hours.

$[Hours]$

Figure 49 shows two scatter plots. The first scatter plots, on the left, show the bi-variate distribution of conductivity measurements taken at the two stations at the same time. The second plot, on the right, shows the bi-variate distribution of conductivity measurements taken with a time gap of three hours between source and target stations, i.e., a pair of points (X,Y) is taken with X being the measurement station at time T, and Y being the measurement at the target station at time T + 3 hours. The graph on the right, accounting for the flow-time between stations, shows a substantially better correlation between the two stations. Table 3, below Figure 49, shows the sample correlations in the two graphs, using our three measures of correlation, discussed in Section 2.3. We can see that correlations that account for the flow time between network locations (based on the subfigure on the right of Figure 49) are higher than "raw" correlations (based on the subfigure on the left of Figure 49).

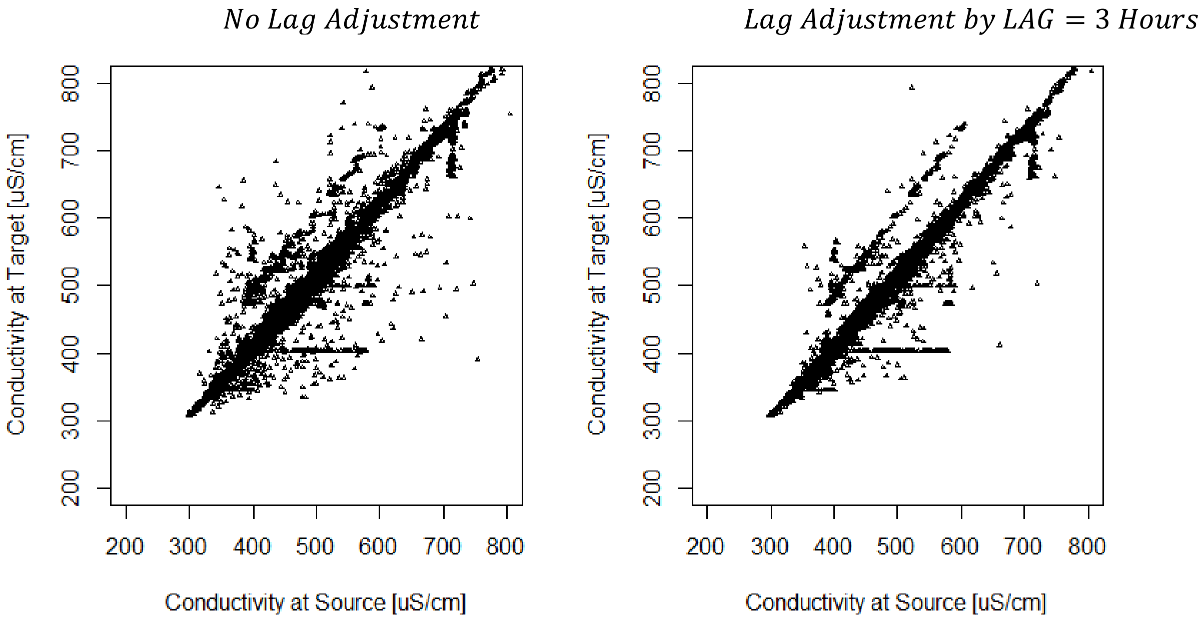


Figure 49: Plotting the hourly median conductivity measurements for the two time-series presented in Figure 45. Points on the left plot show the hourly median measurements, for the same hour at both stations, while the right plot shows hourly median measurements with a lag time of three hours: for example, a point whose X coordinate is the median conductivity value measured between 07:00-08:00 AM at the source station, (at some specific day), had a Y coordinate corresponding to the hourly median conductivity at target station, for 10:00-11:00 AM at the same day.

Table 3: Correlation statistics, for the two subfigures presented in Figure 47

	No lag adjustment	With lag adjustment, lag = 3 Hours
Pearson	0.953	0.961
Spearman	0.948	0.957
Kendall's τ	0.870	0.886

This example motivates the following methodology:

- Define the list of pairs in the network which exhibit meaningful correlations.
- Find flow time estimates for each pair defined. Flow time estimates should be allowed to vary across time of the day and days of the week (unlike in the above example).
- For each pair in the network, model the relationship in measurements between source and target stations
- While monitoring the network, continuously predict the measurement for the target node measurements in each pair based on the source station measurements. If measurements deviate from their predictions for a substantial amount of time, provide a notification.

The next chapters will flesh out this high-level methodology, and provide examples and case studies for its application.

The Spatial Model also features triplets: higher-level models that aggregate over the deviations observed in two sensor pairs. Figure 50 depicts the different kinds of triplets that can be constructed using pairs, with arrows depicting the direction of water flow. Model for triplets of sensors are discussed in subsections 3.6-3.7.

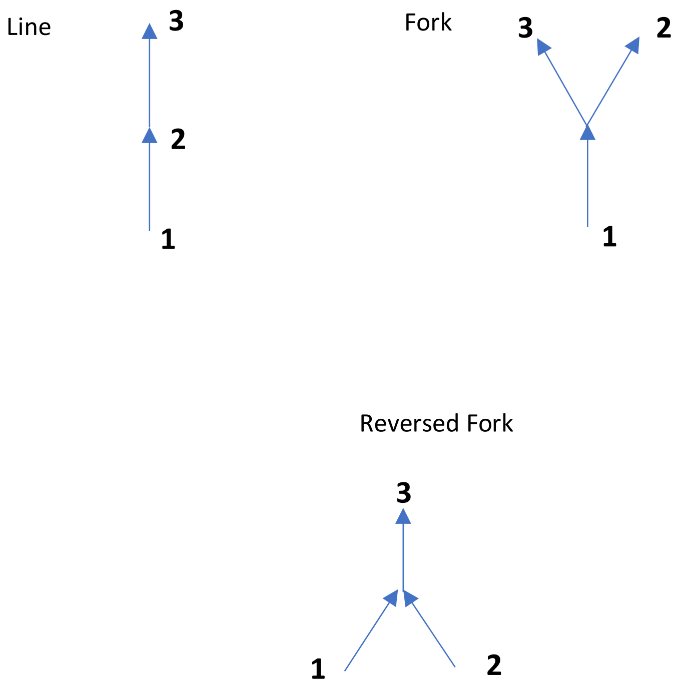


Figure 50: Three possible types of sensor triplets that can be found in a water distribution network. Arrow directions depict the flow direction of the water.

3.2 Pair model – definitions and flow time estimation

This section describes the methodology for estimation of flow times between two network locations, the notion of flow regimes, and how flow time estimates are obtained for several flow regimes.

We start by providing a simple estimator for flow time, when flow time is fixed across all dates and times. We assume the underlying signal for water quality measurements is a continuous function in time. Values for the source station are given by $y^{(1)}(t)$, and values for the target station are given by $y^{(2)}(t)$, with $0 \leq t \leq T$. We assume that the measurements in source and target station are associated via the following relationship:

$y^{(1)}(t) = y^{(2)}(f(t))$,
meaning measurements appearing at time t at the source station appear at a later time, $f(t)$, at the target station. The actual values of $y^{(1)}(t)$ and $y^{(2)}(t)$ are not observed directly. Instead, the observed quantities are averages of the above time series across time. Our discretization for the time domain is such that quantities are observed at equally space points in time, $t = \Delta, 2\Delta, 3\Delta \ldots, T - \Delta, T$. The observed measurements are indexed via i, are will be referred to as $y_i^{(1)}, y_i^{(2)}$, with i receiving values in the range $1,2, \ldots, \frac{T}{\Delta}$.
The relationship between $y^{(1)}, y^{(2)}$ to $y_i^{(1)}, y_i^{(2)}$ is given by:

$$y_i^{(1)} = \frac{1}{\Delta}\int_{\Delta\cdot(i-1)}^{\Delta\cdot i} y^{(1)}(u)du + \varepsilon_i^{(1)},$$
$$y_i^{(2)} = \frac{1}{\Delta}\int_{\Delta\cdot(i-1)}^{\Delta\cdot i} y^{(2)}(u)du + \varepsilon_i^{(2)},$$

with $\varepsilon_i^{(1)}, \varepsilon_i^{(2)}$ describing random measurement noise.
The parameter of interest is $\delta = f(t) - t$, describing the flow time between the two stations.

3.2.1 Estimating δ

Let $y_{(a):(b)}^{(1)}, y_{(a):(b)}^{(2)}$ denote the subvectors from time index a to time index b for the two series.
We will construct a score function estimating how well does a candidate lag time, e.g. u time units, do in explaining the association between $y_i^{(1)}$ and $y_i^{(2)}$, i.e., how reasonable it is for the actual lag time to be a value of u. Our idea for the score function is the following. We will take two windows of measurements, each of which is w measurements in size, and their centers u measurements apart. We will slide the pair of windows across the data, and for each pair of locations, will compute a correlation score, e.g. the Pearson correlation coefficient. The score function for the candidate lag time u will be the sum of Pearson coefficients in all window pairs. Figure 51 depicts the pair of windows, w measurements in size, and u measurements apart, positioned over the two time-series.

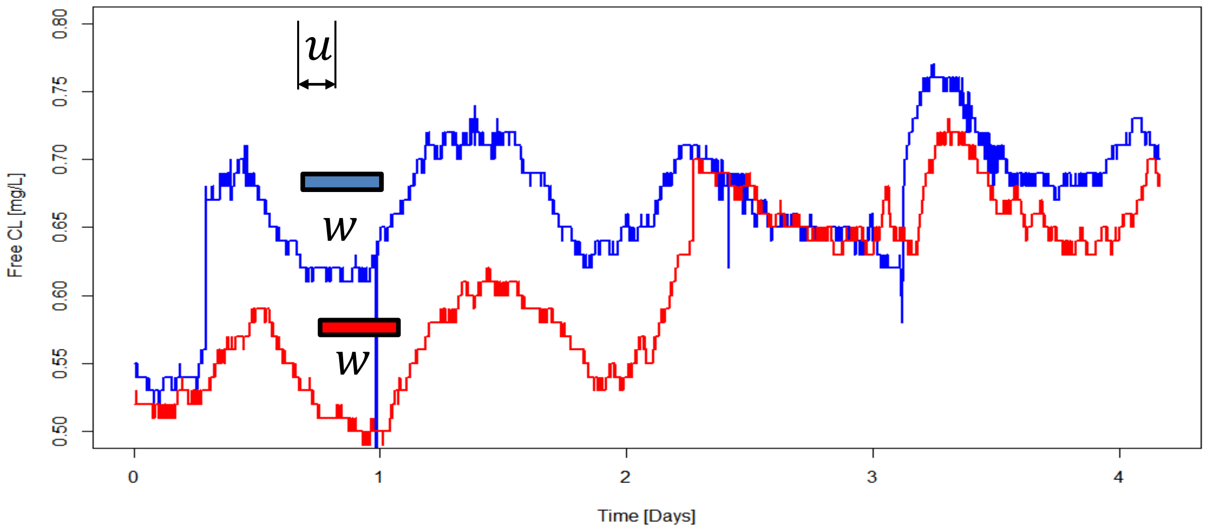


Figure 51: Two times series of measurements, the blue series is the source stations, and the red series is the target stations. In order to compute the rhoCCF function for a shift u: A) Two windows of measurements, w measurements in size and u measurements apart are "slid" across the data, B) for each location point, the correlation between two subseries is computed C) correlation subseries are aggregated.

Our score function is therefore defined to be:

$$\rho CCF(u) = \frac{1}{N} \sum_{\substack{i=\frac{w}{2}+1,\\ \text{and pair of windows exists}}}^{N-w/2} \hat{\rho}\left(\boldsymbol{y}^{(1)}_{(\boldsymbol{i}-\boldsymbol{w}/2):(\boldsymbol{i}+\boldsymbol{w}/2)}, \boldsymbol{y}^{(2)}_{(\boldsymbol{i}-\boldsymbol{w}/2+\boldsymbol{u}):(\boldsymbol{i}+\boldsymbol{w}/2+\boldsymbol{u})}\right),$$

Where $\boldsymbol{y}^{(1)}_{(\boldsymbol{i}-\boldsymbol{w}/2):(\boldsymbol{i}+\boldsymbol{w}/2)}$ represents the w-long subseries of measurements in the first time series, from index $i-\frac{w}{2}$ to $i+\frac{w}{2}$; and $\boldsymbol{y}^{(2)}_{(\boldsymbol{i}-\boldsymbol{w}/2+\boldsymbol{u}):(\boldsymbol{i}+\boldsymbol{w}/2+\boldsymbol{u})}$ represents the w-long subseries of measurements in the second time series, from index $i-\frac{w}{2}+u$ to $i+\frac{w}{2}+u$.

If we draw the function ρCCF as a function of u (Figure 52), we hope to find a maximum in the function. The X coordinate for the maximum will give the flow-time that best explains the shift in the two time-series:

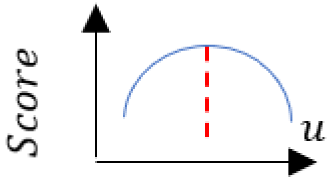


Figure 52: Visualization of the score function

We recall that the units of u are given in time measurements, therefore our approximated lag, $\hat{\delta}$, is found by multiplying the X-coordinate for the maximum by the resolution constant:

$$\hat{\delta} = \Delta \cdot \arg\max_{u} \rho CCF\ (u)\ .$$

We note that since correlations are valued between -1 and 1, an outlier measurement can damage only w windows it participates in. Therefore, its damage is bounded.

The above method works well when flow-time is approximately constant across all time of day. Generally, there is reason to believe this is not the case: flow times change across days and times of the day. We therefore suggest a semi-automatic method: the user selects a partition for days of the week and times of the day, so that the flow time in each part of the partition is approximately constant. Then, flow times are approximated in each flow regime. For example, Figure 53 shows such a partition for Figure 51, with orange and blue backgrounds, dividing days to a day flow-regime and a night flow-regime.

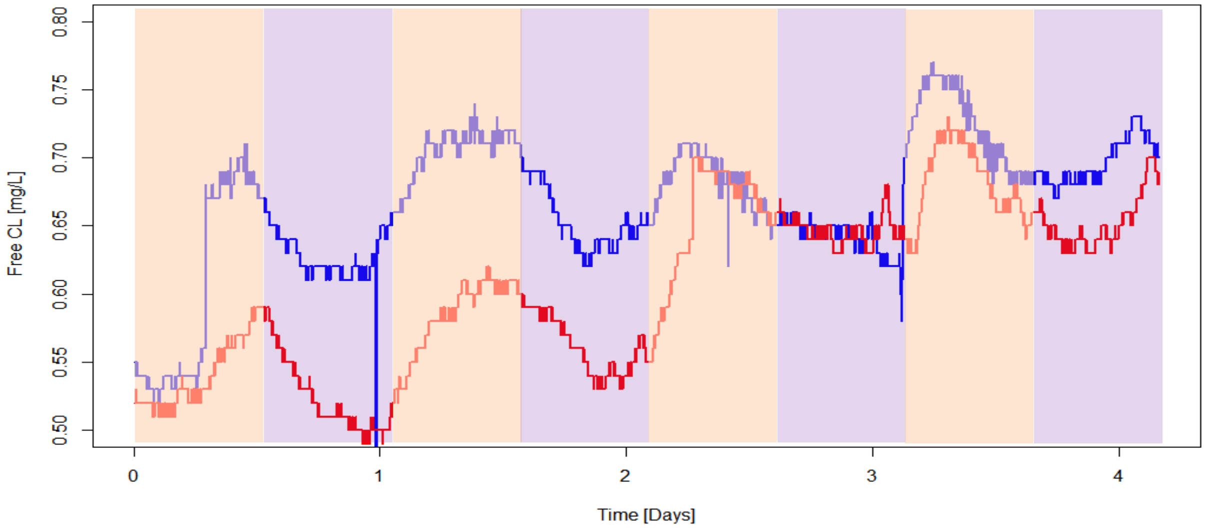


Figure 53: A partitioning of Figure HHH to different flow regimes. Flow regimes are depicted by background color. In this Figure the partitioning is to day and night regimes.

Once such a partitioning is defined, flow times are approximated in each flow regime by summing only over the window pairs that below to the flow regime. Regime identity is defined by the center of the window of measurements for the target station. Figure 54 shows an illustration of three score functions, for three regimes, with each score function receiving its maximum at a different X point.

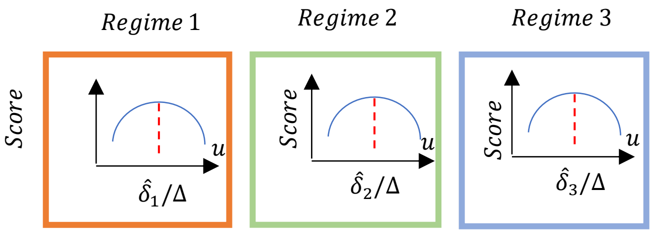


Figure 54: Visual representation for the Spatial Model score function.

Mathematically, the estimated delay time for regime h, denoted by $\hat{\delta}_h$, is given by:

$$\hat{\delta}_h = \Delta \cdot \arg\max_u \rho CCF_h\,(u),$$

Where ρCCF_h is a version of ρCCF that sums correlation scores only over window pairs belonging to regime h.

Constructing a partition of the days of the week and times of the day so that flow times can be estimated is a network-dependent and site-dependent task, and is the subject discussed in Section 3.3.

In terms of the correlation function used, we've found that the cross-covariance function, over ranks of observations, works best in most cases. The value of w is usually taken to be the number of measurements corresponding to 60-90 minutes (with 90 minutes taken for more noisy sensors). For low energy sensors, the measure only once every 15 minutes, we take the w to be the number of measurements that corresponds to 120-150 minutes.

In terms of measurements that should be used: We've found the conductivity is the most reliable in terms of flow time estimation, followed by redox, and then pH. Turbidity and Free CL measurements are generally not reliable for flow time estimation.

The remainder of this section discuss real life examples of score function charts, and how they can be viewed and analyzed.

3.2.2 Real data examples for score functions

Figures 55 and 56 show two examples for regime score functions by regimes, for two spatial model pairs based on Conductivity measurements. We observe see that the maxima in each figure, across the different flow regimes, is obtained at different X-locations, corresponding to different estimated flow times. Figure 55 and 56 show very distinct peaks for many of the different flow regimes (except Friday evening and Saturday morning, which have relatively limited water flows), and are given here as a yardstick for score functions of a model fitted correctly to the data.

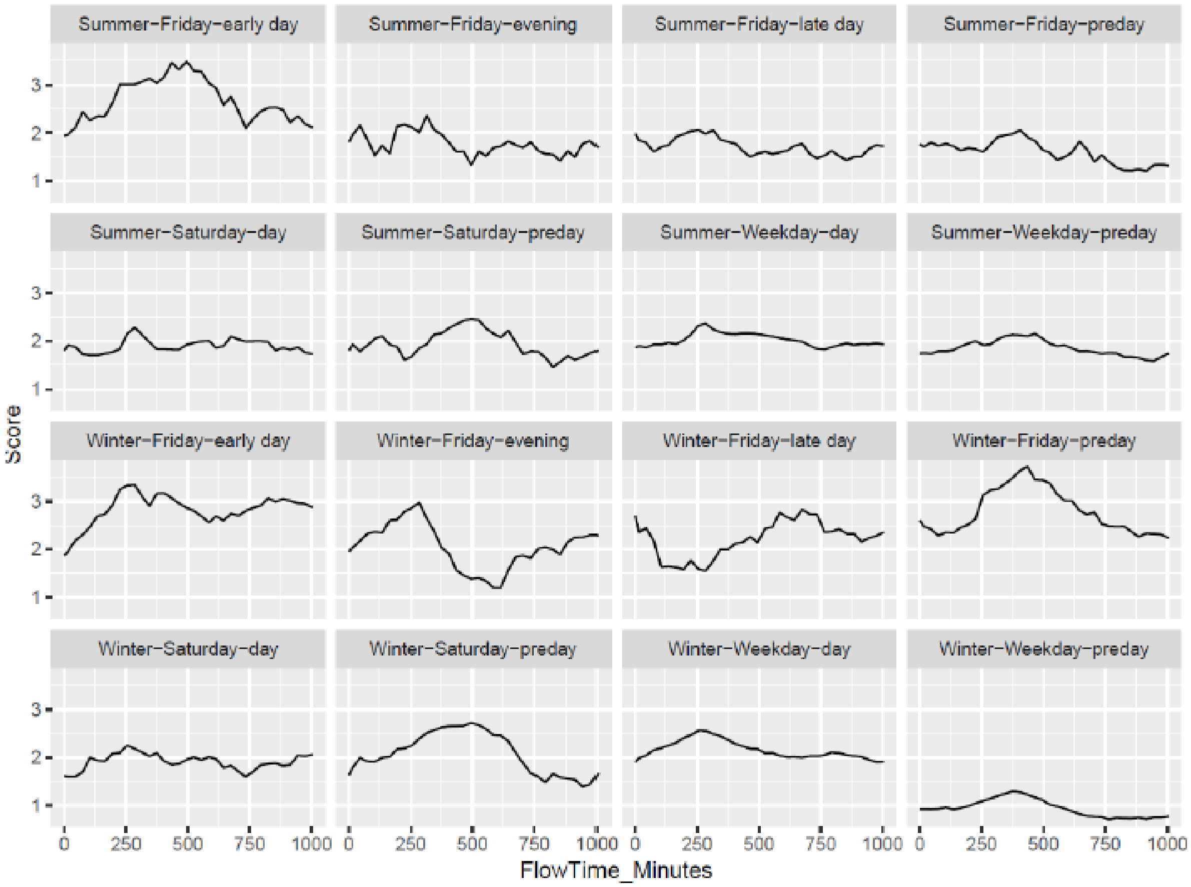


Figure 55: An example for flow regime score functions, for a Spatial Model pair connecting two stations in the same pressure zone. Flow time estimation is based on Conductivity measurements.

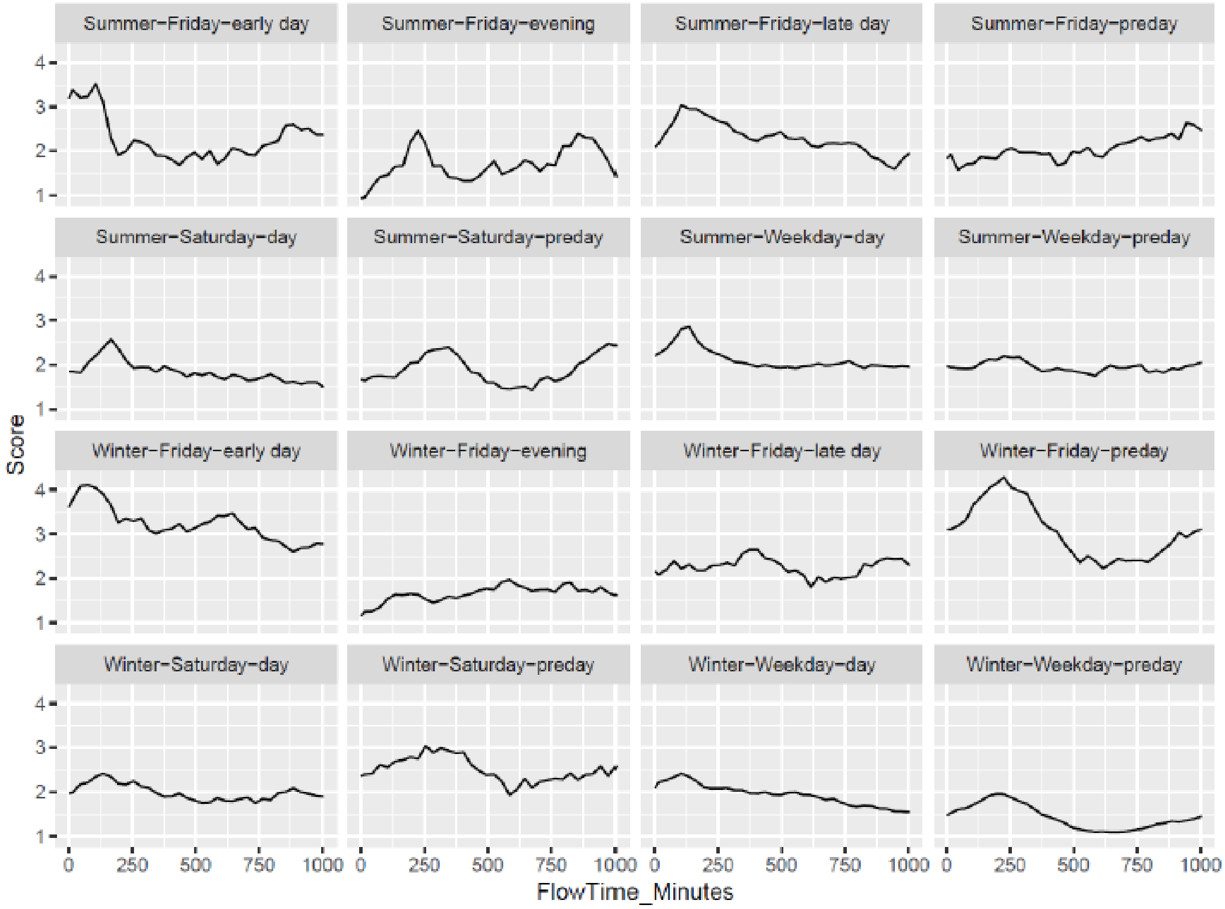


Figure 56: A second example for flow regime score functions, for a Spatial Model pair connecting two stations in the same pressure zone. Flow time estimation is based on Conductivity measurements.

Figures 57 and 58 show the score functions for the links depicted in Figure 55 and 56, respectively, with score function for several measurements depicted together. Score functions were normalized so that the maximum score value in each function is normalized to 1. In Figure 57, we observe Conductivity and Redox measurements to give similar flow time estimates; the maxima across different subplots is found in the same X location for Conductivity and Redox measurements. Generally, Conductivity gives the most distinct peaks in the data. Score functions computed based on Free Chlorine are the most noisy, and are found in the least agreement with other functions. In Figure 58, Conductivity and pH also relatively agree in estimated flow times.

Additional examples for score functions, and comparisons between score functions obtained using different measurements are found in Appendix A.1.

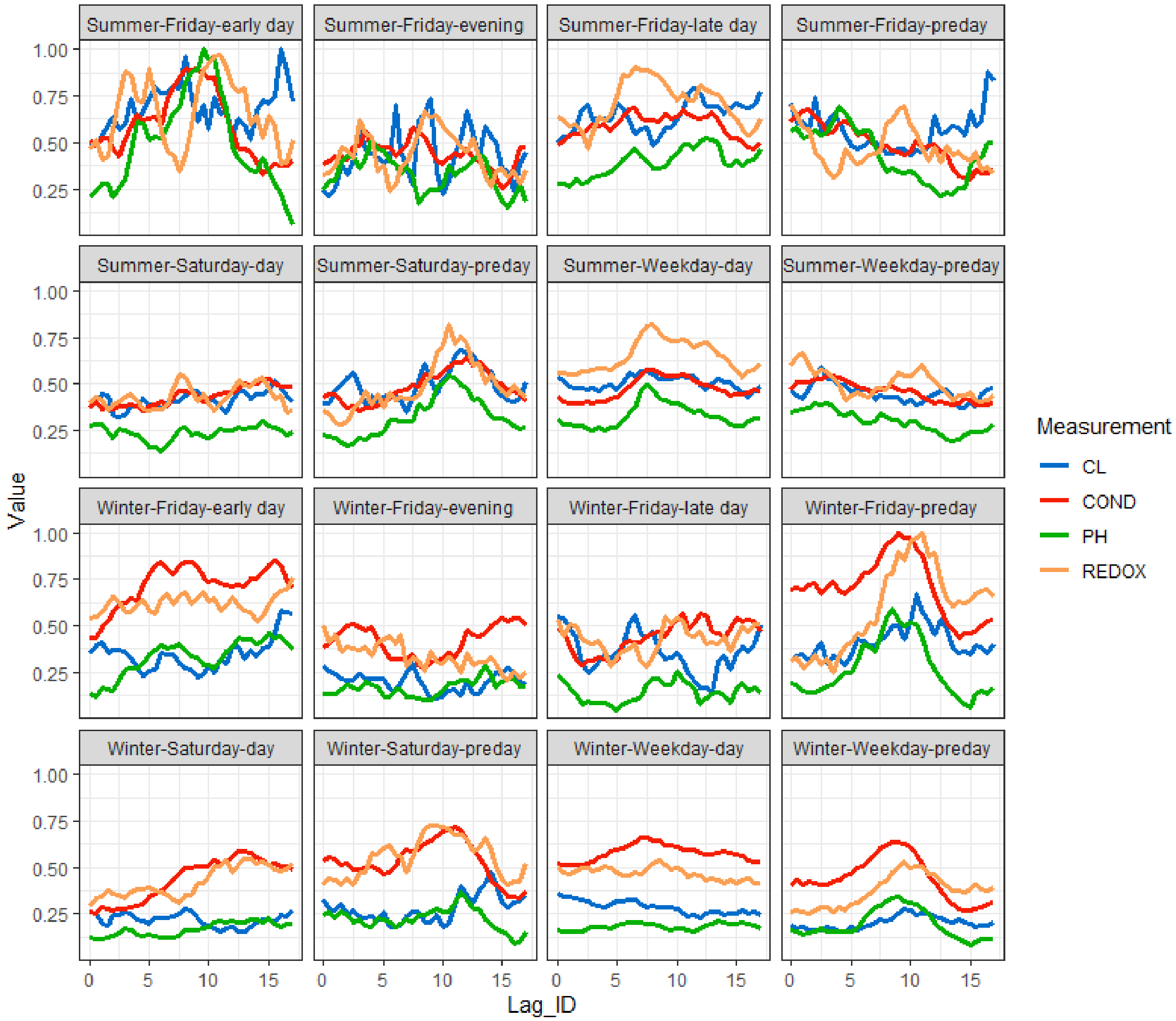


Figure 57: Score functions for Conductivity, Redox, pH and Free Chlorine measurements, for the pair model presented in Figure 55. Score functions were normalized to have a joint maximum value of 1.

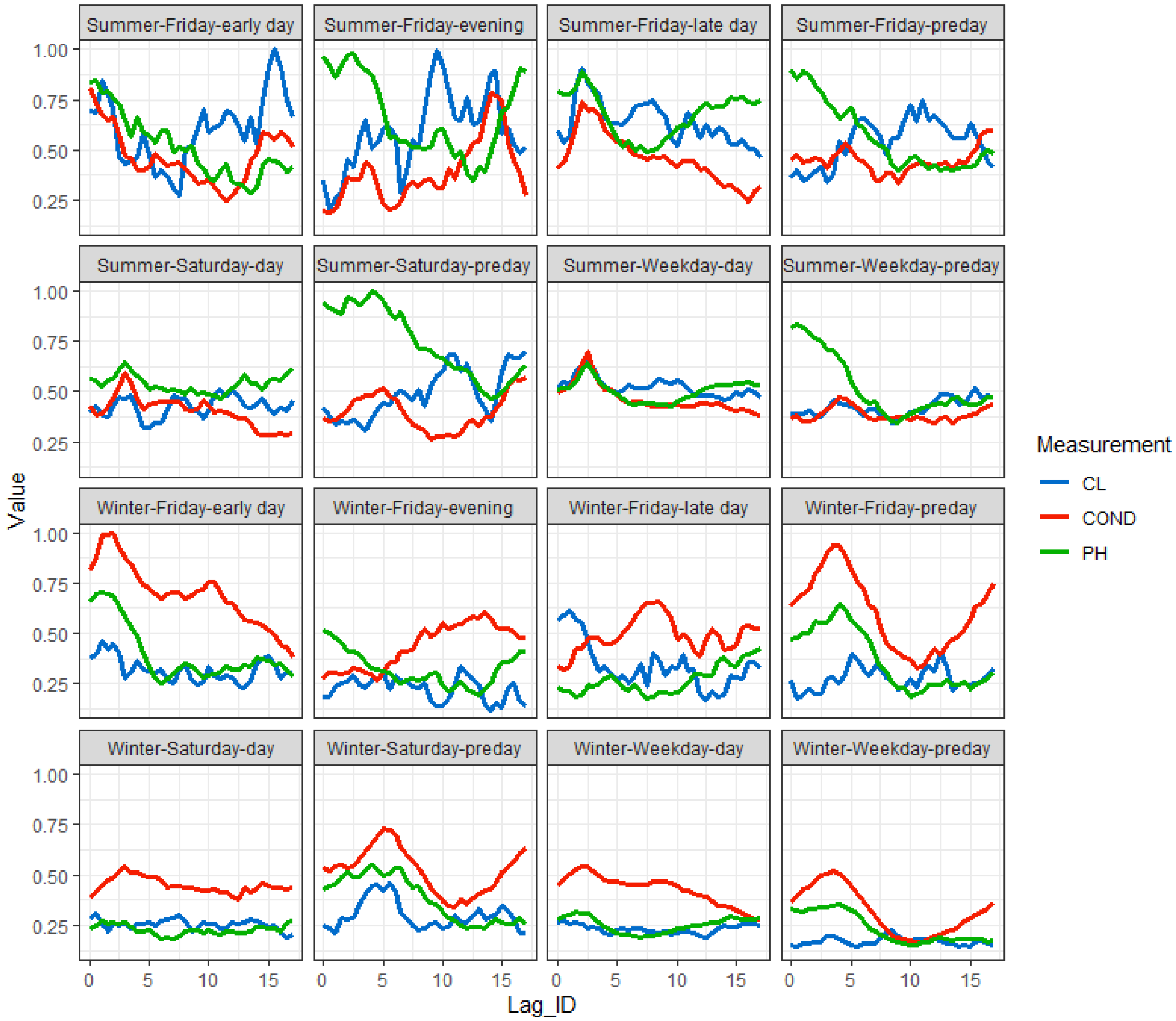


Figure 58: Score functions for Conductivity, pH and Free Chlorine measurements, for the pair model presented in Figure 56. Score functions were normalized to have a joint maximum value of 1.

3.3 Selecting flow regimes

In Section 3.2, we reviewed the SM methodology for flow time estimation, given a fixed partition of measurement times into different flow regimes. We presented a methodology that allows the user to estimate a representative flow time for each flow regime. This section discusses how flow regimes are selected.

In order to find a suitable partition of the dates and times to flow regimes, we use an additional flow meter measurement found inside the network, ideally in a central location, or in the entrance to the network. The major assumption is the times of low and high flow identified using the flow meter are similar to the times of low and high flow in the modeled link.

As a motivating example, we view Figure 59, showing the median, 10% percentile and 90% percentile of hourly flow, in a major water distribution network in Israel. The figure shows the distribution of flow time, conditioned on the time of the day, day of the week, and season of the year. We notice the values of flow may substantially vary, from values of 1000 m^3/h and up to 5500 m^3/h. However, when looking at the distribution of flow for a specific type of day and time of the day, the variance in hourly flow values is substantially lower. This motivates us to "breakdown" the time of the day, and hour of the week, to different regimes based on different partitions, so that the distribution of flow with-in each flow-regime is with a smaller span of values.

We continue after Figure 59 with our working example for this section.

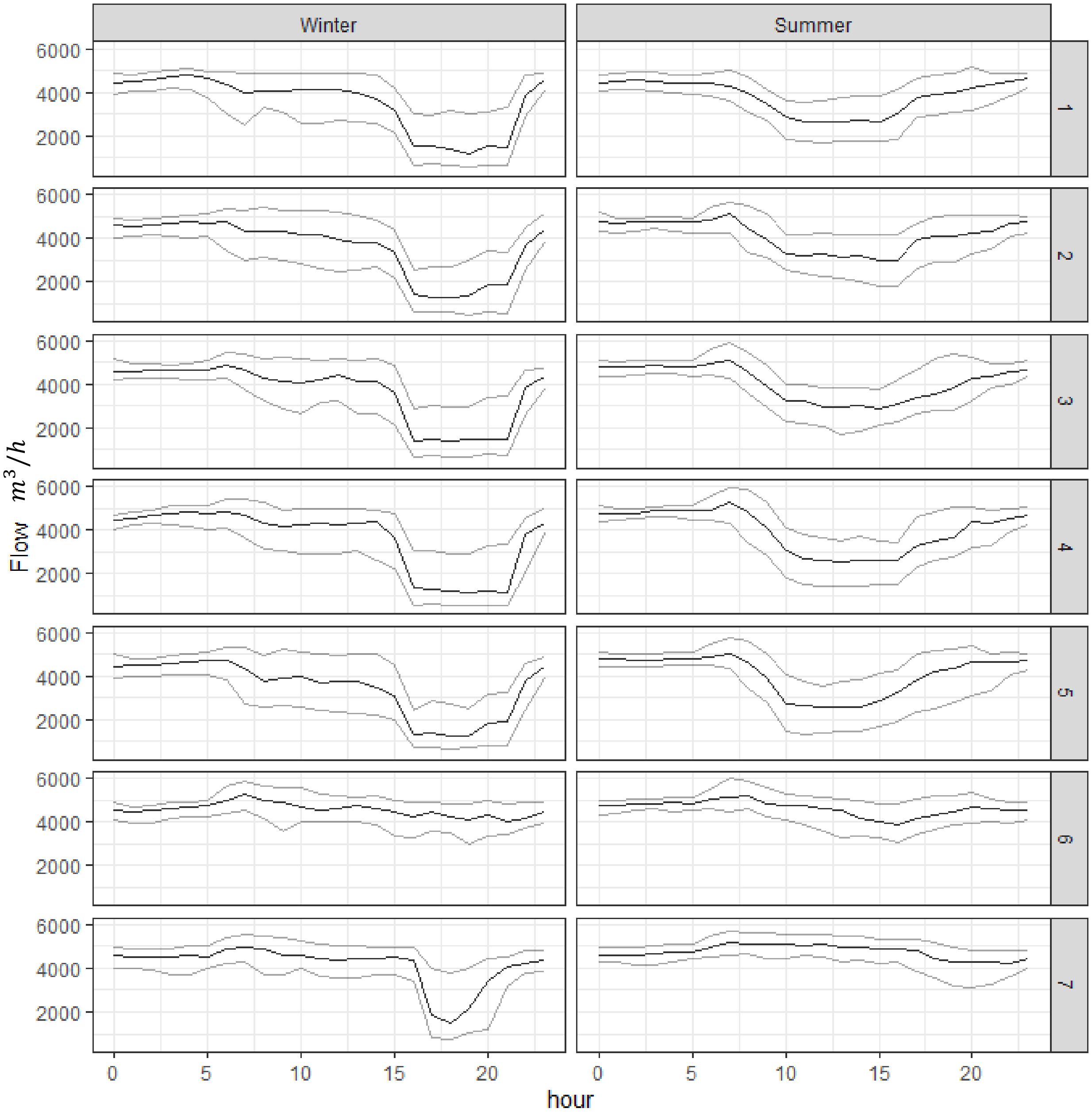


Figure 59: Motivating example - Median, 10% and 90% percentiles for flow, in a flow meter located in a water distribution network, for different hour of the day, days of the week and Seasons of the Year. The flow meter is located in Israel, and Seasons are defined by the daylight saving time. Day 1 is Sunday.

The data for our working example for this section is shown on Figure 60. Figure 60 shows the distribution of hourly flow time, by the different hours of the day. The summary statistics plotted are the median hourly flow, the 10% percentile and the 90% percentile. Different times of the year and days of the week are disregarded in this graph.

Next, we view the distribution of flow time by the hour of the day **and the day of the week**, on Figure 61. The plot shows the variance of flow times for each hour of the day is substantially reduced, compared to the variance of hourly flows (within each hour) on Figure 60. This can be seen by the range between the two gray lines in each hour.

We continue partitioning our data by the seasons of the year. Figure 62 shows the distribution of hourly flow times by the hour of the day, day of the week and season of the year. The season is the year is defined by the use of daylight-saving time in the location of the flow meter. We notice there are significant differences in the distribution of flow times between the two seasons. We also notice the gap between gray lines, marking the 10% and 90% percentiles of hourly flow values, is substantially smaller than the gap we originally started with on Figure 60. Our discussion of the working example continues after Figure 62.

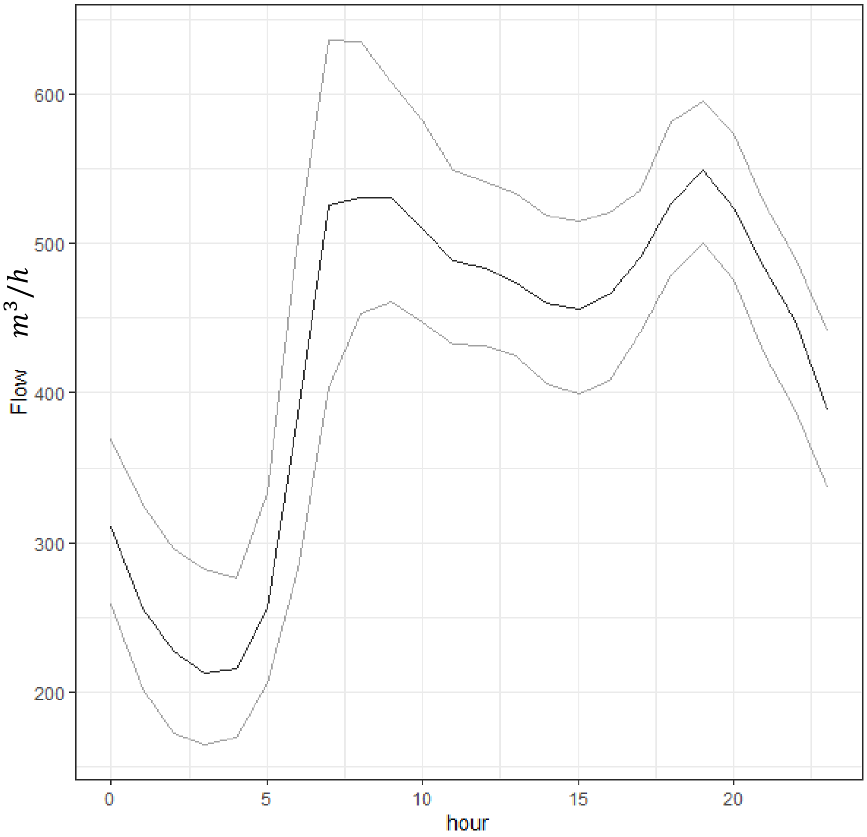


Figure 60: Working Example -Median, 10% and 90% percentiles for flow, in a flow meter located in a water distribution network, for different hour of the day.

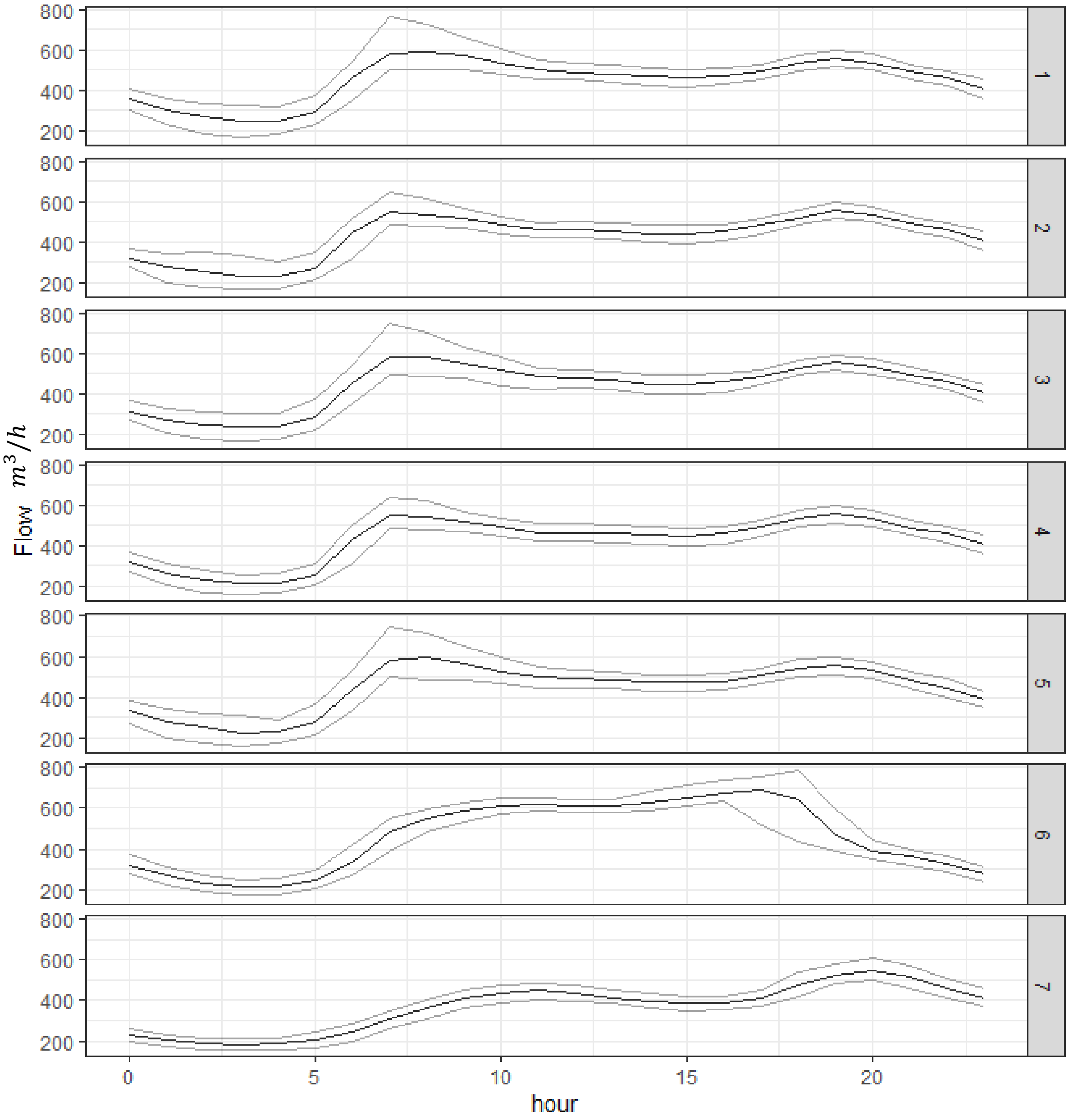


Figure 61: Working Example (continued) -Median, 10% and 90% percentiles for flow, in a flow meter located in a water distribution network, for different hour of the day and days of the week. Day 1 is Sunday.

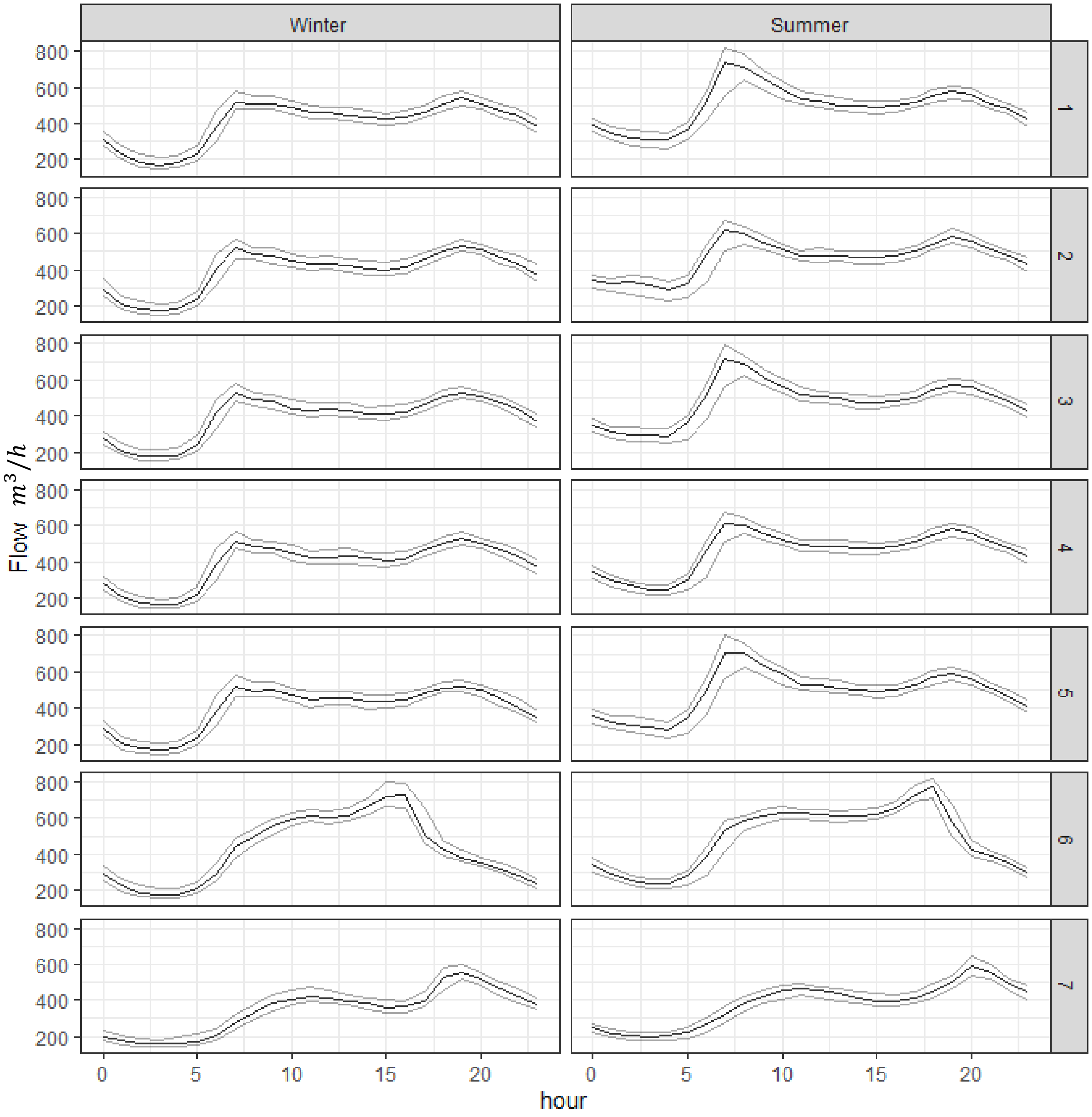


Figure 62: Working Example (continued) -Median, 10% and 90% percentiles for flow, in a flow meter located in a water distribution network, for different hour of the day and days of the week and seasons of the year. The flow meter is located in Israel, and Seasons are defined by the daylight saving time. Day 1 is Sunday.

Next, we can define flow regimes by observing adjacent times and dates which have similar flow values. For example, we can define "Summer-midweek-day" to be the flow regime with summer days , which are also Sunday-Thursday, and hours 8AM to 23PM. Similarly, "Winter-midweek-day" may be defined for Winter days. "Summer-midweek-night" and "Winter-midweek-night" may be defined for the remaining hours of the day. For Friday and Saturday, different regimes could be defined based on the hourly flow values.

For this example, partitioning measurement times by three variables (time of the day, day of the week and season of the year) was sufficient in order to achieve well defined flow regimes (small variation of flow with-in each regime). Figure 62 shows substantially lower values of conditional variance compared to Figure 60. However, this may not always be the case: it could be that partitioning measurement times by the above three variables may not lead to lower conditional variances of hourly flows. For these cases we suggest the following guidelines:

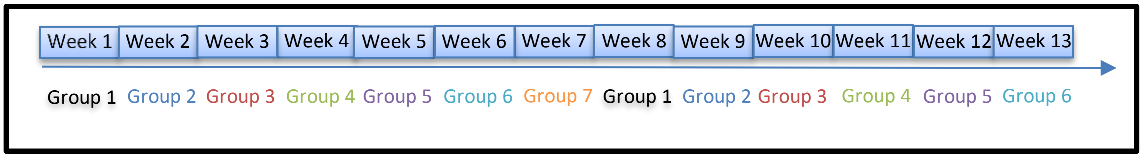
- Draws figures such as 59-62 for your flow data. See if the variation in hourly flow values decreases as more variables are added into the analysis.
- Variables which are not helpful, i.e., do not decrease the variation in flow time, should generally be disregarded when selecting the flow regimes.
- Find additional variables that may be indicative of the flow in the network. This includes operational variables such as the states of pumps, levers and the source of water used (if more than one source is available). As an example, see the definition of network regimes at the start of unit C, and Sections 4.2.2-4.2.3.
- For the above water distribution network, positioned in Israel, Friday and Saturday were known a-priori to be substantially different from the other days of the week. If you know of a different time of the day, or day of the week, that behaves substantially different from other times, use an indicator variable to identify these times.

An important point to note, is that as more flow regimes are defined for the data, less data points are available in each regime, thus making flow time estimation harder. We generally suggest at least one month of data in each flow regime, in order to have good estimation of flow times, along with the ability to use the estimation of the sampling error as detailed in the next section.

3.4 Estimating the sampling error in flow-time estimation

When analyzing flow-time estimates using the Spatial Model, we are also interested in estimates for the sampling error of estimators, i.e., if data was sampled from a slighlty different time period, how would the flow-time estimates change. We utilize a resampling procedure in order to generate our error estimate.

Step I- generating several "folds" of the data: We partition the different weeks of the time line into several groups, e.g. 7, in a cyclical manner. Let $N_{partitions}$ denote the number of week groups. Figure 63 shows such a partition of the time line. We than compute our flow time estimators $N_{partitions}$ times, by each time removing one of the week groups from the data. At the end of this step, each flow regime has $N_{partitions}$ estimated flow times, each computed with $\frac{1}{N_{partitions}}$ of the data removed.



Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13
Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6

Figure 63: A visualization of a timeline with a parititon of weeks into 7 different subgroups.

Step II- compute error estimates: for each of the flow regimes, we compute the InterQuantile Range (IQR) for the $N_{partitions}$ flow estimates. The IQR is defined to be the distance between the 3rd quartile and the 1st quartile. **Our error estimate is half the IQR**: it is the distance from the center of the distribution of flow time estimates in which you find the central 50% of estimates.

The above resampling scheme uses a time period of a week as its basic block. This scheme was chosen so that autocorrelation of water quality time series is taken into consideration as discussed in Section 2.4.

We note that by using too few folds, flow time estimates will become too noisy compared to the original estimates. By using too many folds, the flow time estimates in each fold will be effectively computed using the same data using for the original estimate, and therefore will be non-informative. We suggest to set $N_{partitions}$ by seeking to estimate **"how much will flow-time estimates vary if another $\mathbf{1/N_{partition}}$ of the data were sampled?"**.

Figure 64 show the score functions for 7 different data-folds, for the data and score functions depicted in Figure 56. Each of the 7 different curves, depicts a score function computed after removing 1/7 of the data. Vertical lines represent the estimated flow times, for each fold of the data, for each flow regime. Different flow-time estimates obtained for the different folds in each flow regime are found in relatively good agreement.

Additional graphs showing score function for cross-validation estimates of flow times are found in appendix A.2.

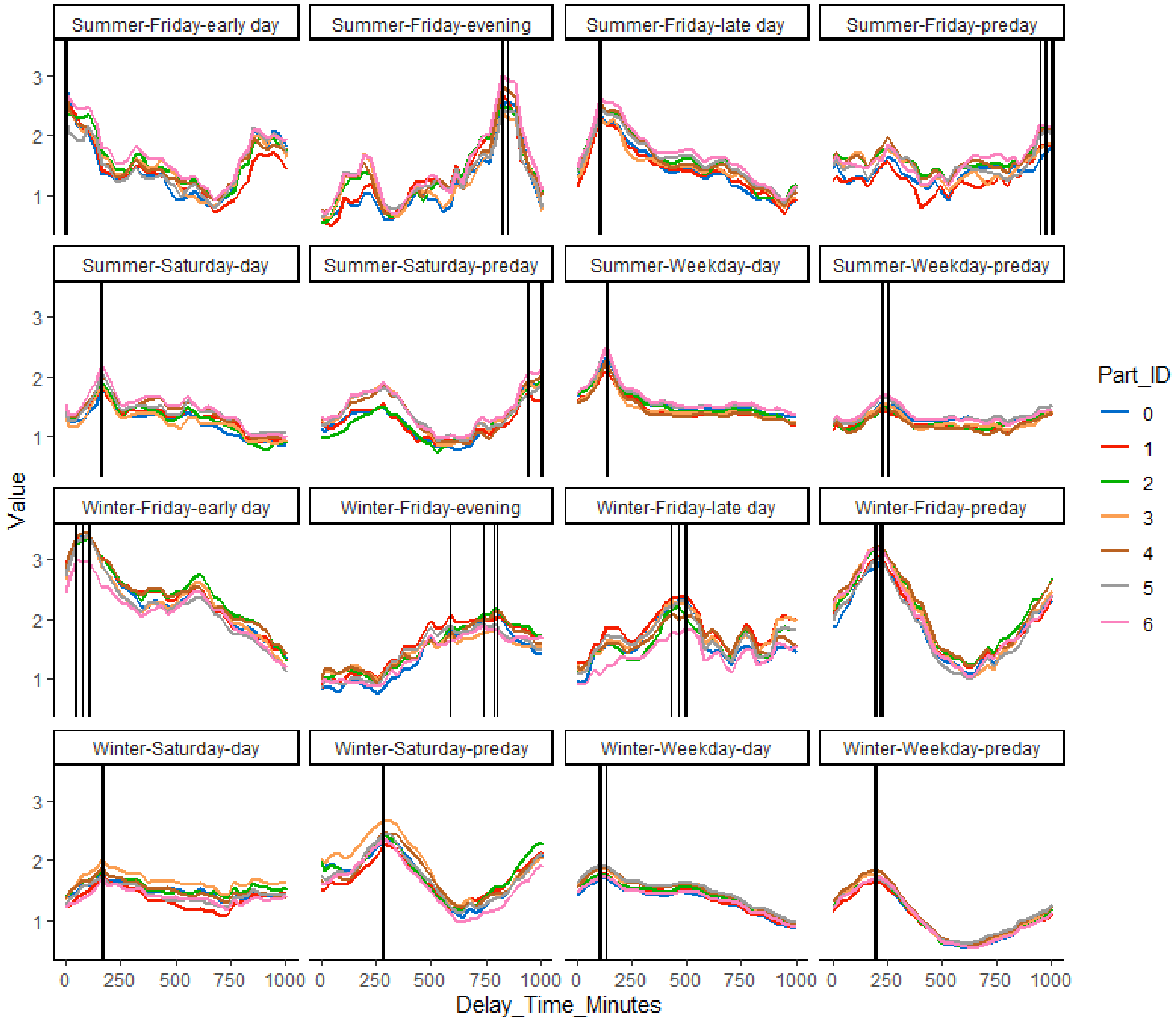


Figure 64: Score functions for 7 data folds, for the data and score functions depicted in Figure 56, for an SM model estimating flow time between two network locations using conductivity measurements. Different colors represent score functions with different week groups removed. Vertical lines mark the flow-times with maximal score values in each fold of the data.

3.5 Pair Model – Prediction and Monitoring

This section discusses how prediction models are built for SM pair objects, and how prediction models are used for monitoring. Construcion of prediciton models is discussed in Section 3.5.1, and monitoring is discussed in 3.5.2.

3.5.1 Construction of prediction models

The first step in construction of prediction models is the construction of a training set according to the following steps

Step I: for a preset period of time, e.g. most recent 30 days prior to model construction, collect all the times for which the target node has measurements. We denote times in which target station measurements were available by $t_1, t_2, \dots, t_N$.

Let $\hat{\delta}(t)$ be the estimated flow time, for water arriving to the target station at time t. For example, water arriving at time t_1 at the target station passed the source station at time $t_1 - \hat{\delta}(t_1)$. Let $y^{(1)}(t)$ and $y^{(2)}(t)$ represent water quality measurements at the source and target station, respectively.

Step II: create a dataset with the entries:

$$\left(y^{(1)}\left(t_1 - \hat{\delta}(t_1)\right), y^{(2)}(t_1)\right)$$

$$\left(y^{(1)}\left(t_2 - \hat{\delta}(t_2)\right), y^{(2)}(t_2)\right)$$

$$\dots$$

$$\left(y^{(1)}\left(t_N - \hat{\delta}(t_N)\right), y^{(2)}(t_N)\right)$$

The values of $y^{(1)}(t)$ and $y^{(2)}(t)$ are given by median time bin values, as explained in Sections 2.1-2.2. If a source value $y^{(1)}\left(t_i - \hat{\delta}(t_i)\right)$ is not available, (either due to missing data, fixed data, or data out of value limits) then the ith point is removed from the data.

The values of the form $y^{(1)}\left(t_i - \hat{\delta}(t_i)\right)$, for $i = 1, \dots, N$, will be known as the **flow-time adjusted source values**.

We proceed with model construction.

Step III: construct kernel regression models for the training set, using a kernel regression (Section 2.5), with a predefined number of bins and kernel size.

Step IV: for each data point, computed its predicted value, $\hat{y}^{(2)}(t_i), for\ i = 1, \dots, N$. Compute also the residuals:

$$e_i = y^{(2)}(t_i) - \hat{y}^{(2)}(t_i), i = 1, \dots, N$$

The residuals will be used for monitoring and event detection in Section 3.5.2.

The above model is a prediction model for all flow regimes, with the variation in flow-time only taken into consideration when selecting the X-coordinate values for the training set. Empirically, we have found constructing different prediction models for each flow-regime leads to lower prediction accuracy, especially in flow-regimes which span relatively short time span, e.g. weekend regimes.

Figure 65 shows three graphs that are used in the diagnostic analysis of SM pair object prediction models. The three graphs, from left to right, are: **A source by target graph**, showing the values of $y^{(1)}\left(t_i - \hat{\delta}(t_i)\right)$ in the source station against the values of $y^{(2)}(t_i)$ in the target stations, for different times (by values of t_i; A histogram of prediction errors; A scatter plot of prediction values vs. actual values measured. The three Graphs in Figure 65 are shown for a prediction model for conductivity measurements, and show a model with good fit to the data. Additional examples that can be used as yardsticks, along with examples for a bad model fit, are shown in Section 4.8.

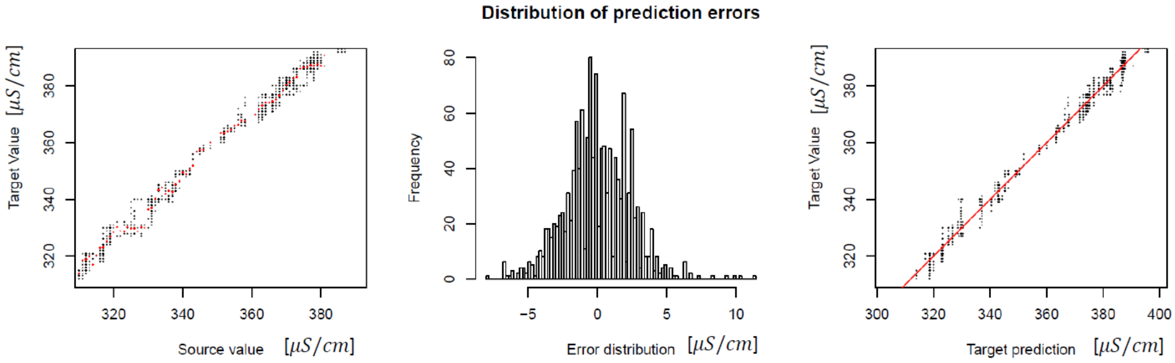


Figure 65: Diagnostic graphs for a pair prediction model. Left: Comparison of Target measurements, flow time adjusted flow measurements. Center: Histogram of prediction errors. Right: Scatterplot comparing Target station predicted measurements, against their actual measured values.

Figure 66 compares shows an Actual vs. Predicted graph for two models – the left model taken flow-time into consideration. The model on the right uses $\hat{\delta}(t) \equiv 0$ in model estimation. A better fit is achieved when considering flow times in model construction.

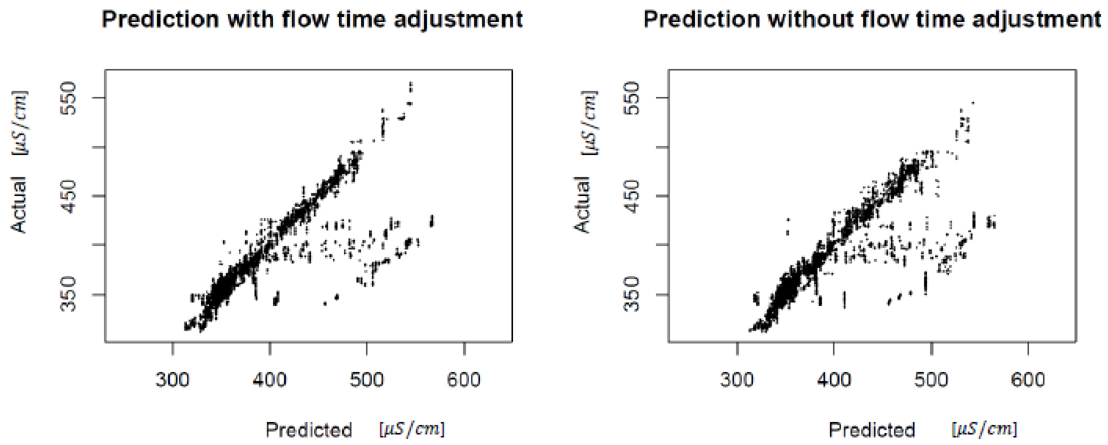


Figure 66: Two "Actual vs. Predicted" scatter plots for a model predicting target station conductivity values based on source conductivity values. The left plot show uses flow-time estimates in model construction. The right plot is for a model assuming the flow-time is identically zero at all times.

3.5.2 Monitoring target station measurements using prediciton models.

We explain how SM pair objects prediction models can be used for detecting water quality events in the target node stations. Our main idea is the following:

We continuously predict target node water quality measurements based on the source values. If prediction errors are irregularly large for a prolonged period of time, we need to notify the operator.

Specifically, event detection happens using the following procedure:

Step I: From the empirical distribution of e_i's (prediction errors), remove the most recent 5% of errors. Sort the distribution of errors from smallest to largest. Denote the sorted residuals by $e_{(1)}, e_{(2)}, \dots, e_{(0.95\cdot N)}$, assuming for simplicity $0.95 \cdot N$ is an integer.

Step II: Find an interval $[L, U]$ holding the $1 - \alpha_{Lower} - \alpha_{Upper}$ central values in the distribution of errors by computing: $L = e_{(\alpha_{Lower}\cdot 0.95\cdot N)}$; $U = e_{\left((1-\alpha_{Upper})\cdot 0.95\cdot N\right)}$.

Step III: For a size M defined by the user as the minimum size for ***half the prediction interval***, set:

$$L \leftarrow \min(L, -M)$$

$$U \leftarrow \max\,(U, M)$$

By default, M is set to be 5% of the user limits for the variable.

Step IV: Compute in addition a set of **"soft limits"**, by computing:

$$L' \leftarrow L - M'$$

$$U' \leftarrow U + \mathrm{M}'$$

Where M' is set to be 5% of the range between statistical limits for the variable.

Step V: Form a prediction interval using $[\hat{y}^{(2)}(t) - L', \hat{y}^{(2)}(t) + U']$

Step VI: Check if $y^{(2)}(t)$ is found outside the range $[\hat{y}^{(2)}(t) - L', \hat{y}^{(2)}(t) + U']$.

The system performs steps III-VI continuously. If $y^{(2)}(t)$ is found irregular using step VI for period larger than a set delay time, e.g., 30 minutes, than a Spatial Model limits violation alarm is raised.

The filtration of the 10% most recent prediction errors is performed so that a water quality event will not "contaminate" the distribution of prediction errors used to declare an event. See Section 4.8.9 for an example demonstrating the importance of this filtration.

The next example illustrates why the requirement that the length of prediction intervals is no more than a minimal size is needed. Figure 67 shows the distribution of prediction intervals sizes $(U - L)$, for a model predicting Free Chlorine in a target node based on a Free chlorine measurements in the source node, for 1 month of data (one observation taken every 10 minutes). The median prediction interval length is around 0.1 mg/L. This indicates that for a vast proportion of times, a decrease in F-CL measurements that is less than 0.05 in size compared to the predicted value, will result in the system declaring the drop in F-CL as abnormal. However, we know that such a drop in F-CL is commonly experienced in water distribution systems with a base line value of 0.4 mg/L F-CL (as is the case here). Hence, we require that the difference in Spatial Model limits in no less than 10% of the user limits, and add an additional "grace period" using the above soft limits (step IV in the above algorithm).

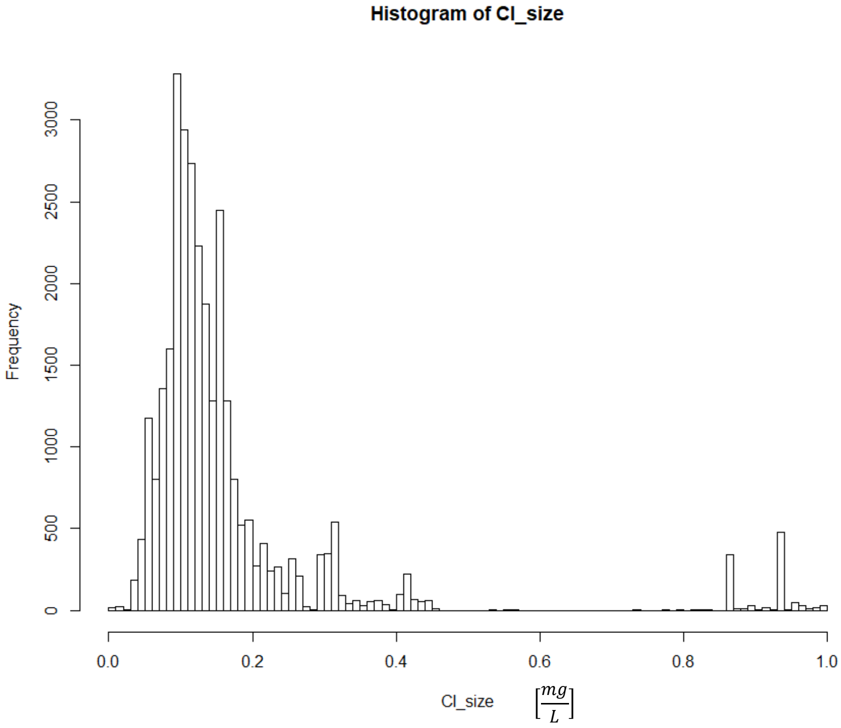


Figure 67: Distribution of prediction interval lengths, for a model predicting Free Chlorine measurements in a target station based on source values, when prediction interval length is not required to be of a minimum size. Measurements taken at 10 minutes intervals, across 1 month.

3.6 Triplets – definitions and types of triplets

In this section, we motivate the use of statistical models combining water quality measurements from three stations in order to detect water quality events. Figure 68 shows the different types of triplet models that can be constructed in the Spatial Model:

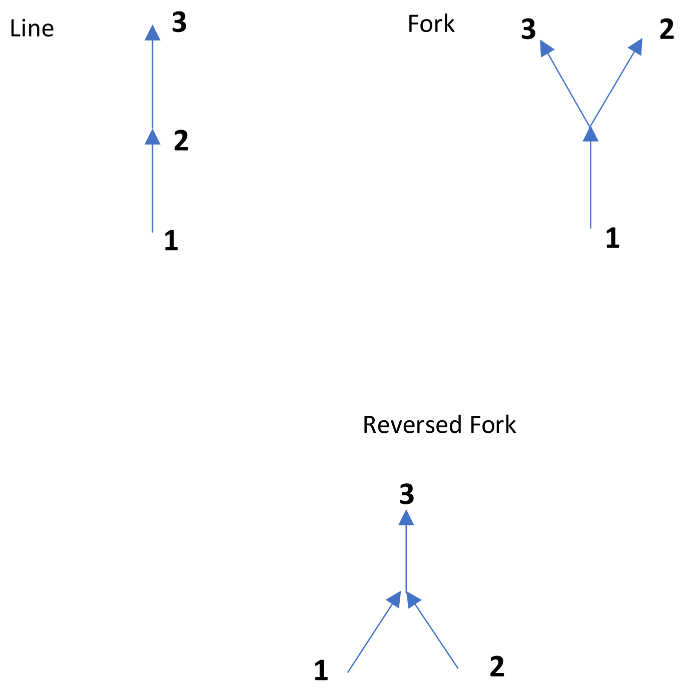


Figure 68: (Copy of Figure 50) Three possible types of sensor triplets that can be found in a water distribution network. Arrow directions depict the flow direction of the water.

We provide several motivating examples, demonstrating cases in which triplet models may be more sensitive to detect abnormalities in water quality events:

- **A water quality event starting after the first node in a Fork triplet –** As an example, we consider a setting where a water quality event with a relatively small footprint on water quality measurements takes place at a fork triplet, with the origin of the event located immediately after the source station (station number 1 in the fork triplet on Figure 68). The water quality event does not affect station 1, but it is measured in stations 2 and 3. If the water quality event has little effect on the gathered water quality measurements, it is possible the SM pair models alone will not detect the event. Since Prediction errors in both stations 2 and 3 will be irregular and in the same direction, we suggest constructing a model that monitors for water quality events in stations 2 and 3 simultaneously.
- **A case detecting calibration is needed in a line triplet**– Consider a line triplet of Free- Chlorine sensors showing the values 0.4 mg/L, 0.2 mg/L, and 0.4 mg/L, for stations 1,2, and 3, respectively. These values hint that at least one of the stations 2 or 3 are not calibrated properly since Free-Chlorine can only dissipate as water travels into the network. Constructing a triplet model can detect irregularities in these values.

We wish to motivate our detection method for sensor triplets. As an example of a fork triplet, we show the hourly median pH measurements for three stations in Figure 69, across 2500 hours (over 100 days). Station 1 (the source of the triplet) is shown in **black**. Station 2, the middle station, is shown in **red.** Station 3, the final station in the line triplet, is shown in **blue**. The three-time series are highly correlated. We continue the example on the next page.

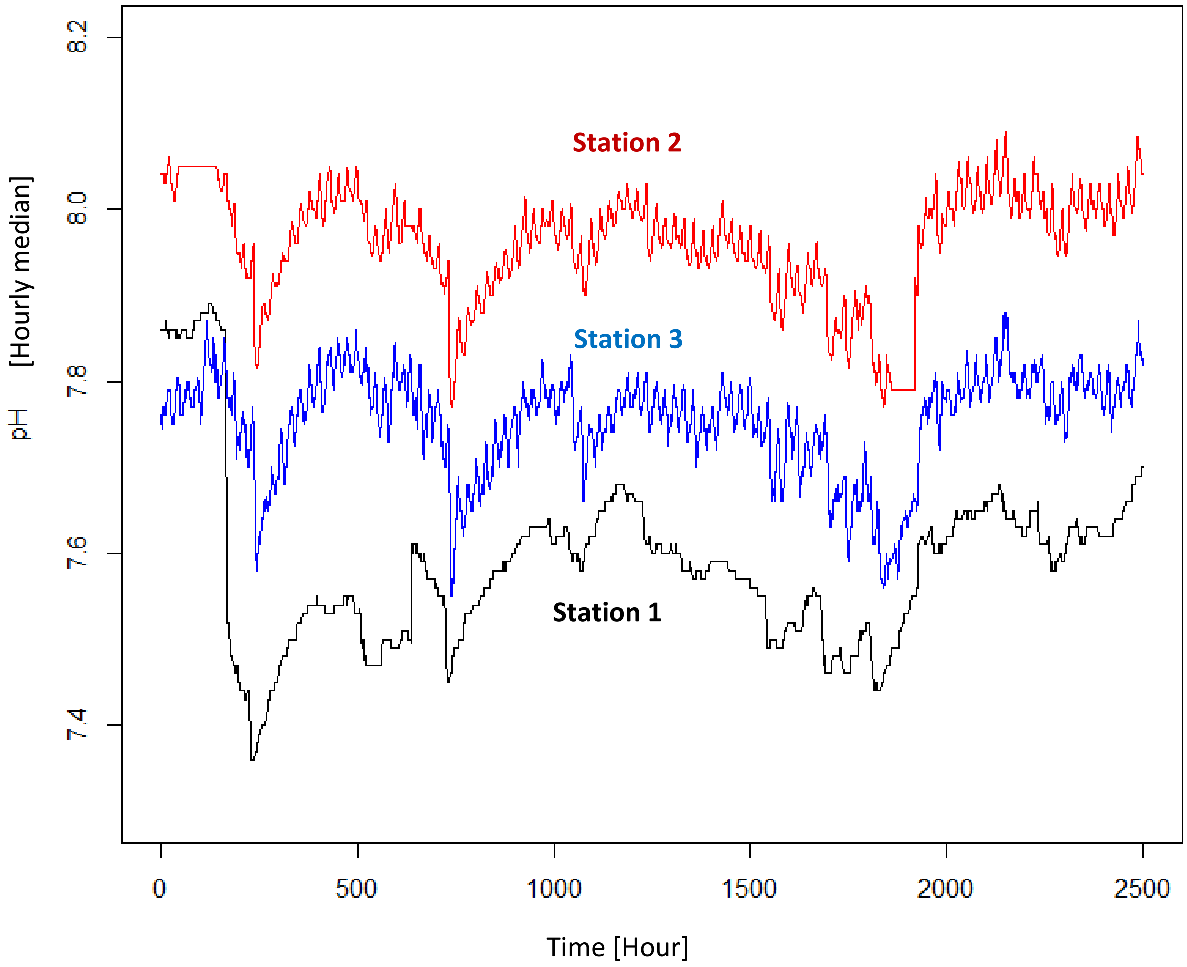
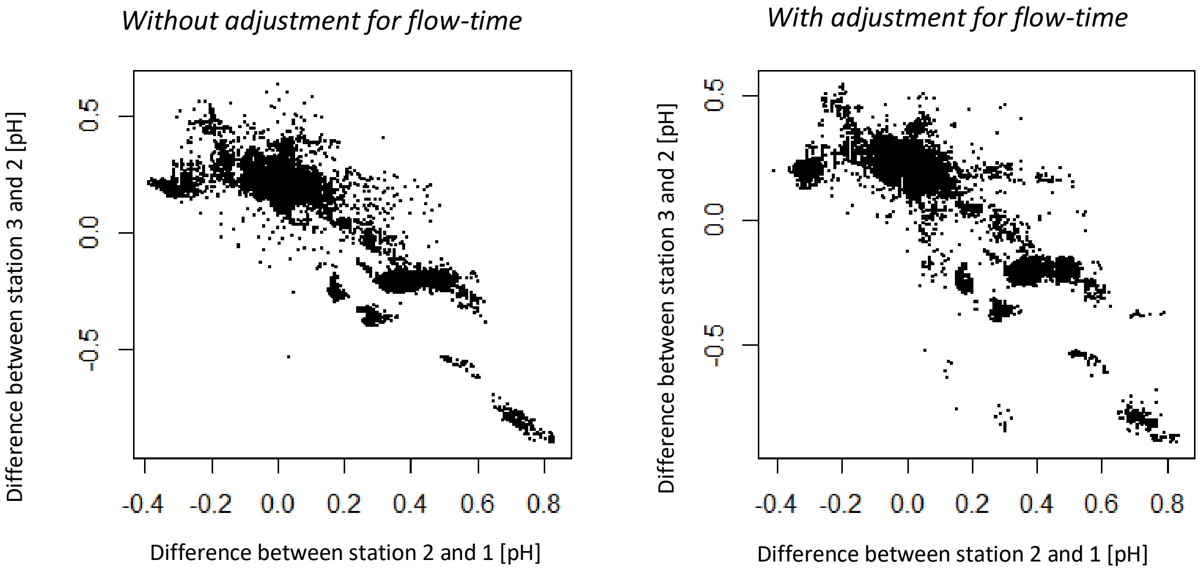


Figure 69: Three pH time series, across 2500 hours, for three stations in a line triplet. pH measurements are hourly medians. Stations 1,2 and 3 are shown in black, red, and blue, respectively.

Figure 70 shows the bi-variate distribution of inter-station differences, with and without an adjustment for flow time. The X axis in each scatterplot shows the difference in pH measurements between stations 1 and 2. The Y axis in each scatterplot shows the difference in pH measurements between stations 2 and 3. Each point represents a joint hourly measurements for the three stations. We observe the within-pair deviations in values to be highly correlated across different times. If the joint value of both within-pair deviations is abnormal, a human operator should be notified of the irregularity.


Without adjustment for flow-time

With adjustment for flow-time

Figure 70: Scatterplots for the differences in pH measurements, for each pair of sensors, for the line triplet shown in Figure 69.The left figure is shown without adjustment for flowtime, while the right figure adjusts for flow time using the method presented in sections 3.2-3.3.

Flow time has to be accounted for when monitoring triplets. However, due to the differences in geometry between the different triplet types, adjustments for flow-times vary. The remainder of this section discusses which how flow times are encountered for. **This will result in us monitoring a set of sensor values measured at the three triplet stations at different times.** Section 3.7 discuss how the sensor measurements from the three network locations at different times are combined in order to detect abnormal water quality events.

In section 3.2, we defined $\delta(t)$ as the time it took water from station 1 to arrive to station 2 at time t:

$$y^{(2)}(t) = y^{(1)}\big(t - \delta(t)\big).$$

Similarly, we can define $\delta'(t)$ as the time it took water from station 1 at time t to reach station 2:

$$y^{(2)}(t + \delta'(t)) = y^{(1)}(t).$$

The functions $\delta(t)$ and $\delta'(t)$ are obviously related, however discussing the relation is beyond the scope of this section. However, our suggested estimator in Section 3.2.1 is built for estimating $\delta(t)$. In order to estimate $\delta'(t)$, we run the algorithm described in 3.2.1, but use the source station for the regime identity. Since triplets involve two pairs of stations, we will use a subscript to denote the pair identity for delay time functions: for example, $\delta_{2\to3}(t)$ and $\delta'_{2\to3}(t)$ will be used to denote the flow time for water arriving from station 2 to station 3 at time t, and the flow time for water leaving station 2 at time t (and heading to station 3), respectively. Similarly, the notations $\hat{\delta}_{2\to3}(t)$ and $\hat{\delta}'_{2\to3}(t)$ will be used to denote the estimators for the above delay times, respectively.

Table 4 shows the different time adjustments used when monitoring different triplet types. We now survey this table in detail. The Fork-out triplet has two operation modes, meaning over-all 4 triplet flow time adjustment methods are discussed. In the first three triplet types, a station monitored at time t denotes a station monitored at the current time, i.e. the wall time, unless stated otherwise. For the "source is pivot" mode of the fork-out triplet, time t denotes a past time.

Table 4: Monitored time points for different triplet types

Triplet type	Stations setting time	time at station 1	time at station 2	time at station 3
Line	3	$t - \hat{\delta}_{2\to3}(t) - \hat{\delta}_{1\to2}\left(t - \hat{\delta}_{2\to3}(t)\right)$	$t - \hat{\delta}_{2\to3}(t)$	t Wall time
Fork-in	3	$t - \hat{\delta}_{1\to3}(t)$	$t - \hat{\delta}_{2\to3}(t)$	t Wall time
Fork-out	2,3	$t - \hat{\delta}_{1\to2}(t)$, $t - \hat{\delta}_{1\to3}(t)$	t Wall time	t Wall time
Fork-out (Is Pivot)	1	t **(not wall time)**	$t + \hat{\delta}'_{1\to2}(t)$	$t + \hat{\delta}'_{1\to3}(t)$

For monitoring, we wish to follow the same "drop of water" across the network, as it passes different locations at different times. For monitoring a line triplet **(line 1 table 4)**, the last station appearing in the water flow is station 3, which as monitored at the wall time, time t. Station 2 is the previous station in the water flow. This station, is monitored at time $t - \hat{\delta}_{2\to3}(t)$, which is the estimated wall time, in which water that left station 2 later arrived at station 3 at t. Similarly, Station 1 is monitored at time $t - \hat{\delta}_{2\to3}(t) - \hat{\delta}_{1\to2}\left(t - \hat{\delta}_{2\to3}(t)\right)$, which is the estimated wall time in which water the left station 1 later arrived at station 2 at time $t - \hat{\delta}_{2\to3}(t)$.

For a fork-in triplet **(line 2 table 4)**, we monitor the target station t. The Estimated times in which water leaving stations 2 and 3 arrived at station 3 are $t - \hat{\delta}_{1\to3}(t)$ and $t - \hat{\delta}_{2\to3}(t)$, respectively.

For a fork-out triplet, we wish to look at the same "drop of water " as it passes through the different stations. However, we cannot monitor both stations 2 and 3 at the wall time, since the flow times in the links 1 → 2 and 1 → 3 may differ. There are two possible solutions for this problem. The first possible solution is to monitor the same "drop of water" both station 2 and 3, but with one of them not taken to be the current wall time. We call this mode of operation "source is pivot" **(line 4 table 4)**, since the source station sets the "monitoring time" for the triplet. In this mode, a water drop passing station 1 at an earlier time t is monitored by also observing times $t + \hat{\delta}'_{1\to2}(t)$, $t + \hat{\delta}'_{1\to3}(t)$ at stations 2 and 3, respectively. The monitor time t is advanced by the system to be the most recent time such that data is available for both time $t + \hat{\delta}'_{1\to2}(t)$ in station 2, and time $t + \hat{\delta}'_{1\to3}(t)$ in station 3.

A drawback of the "source is pivot" mode is that stations 2 and 3 are not both monitored at the current wall time: one of the stations is monitored at a time which is earlier than the current wall time. The danger in "source is pivot" mode, is that if a water quality event affects an end station which is not monitored at the current wall time, the triplet model will only be able to detect the event at a later time. In order to address this issue, fork-out triplets can also be monitored by observing both stations 2 and 3 at the current wall time **(line 3 table 4)**. In this operation mode, the source station is monitored at two different times, $t - \hat{\delta}_{1\to2}(t)$ and $t - \hat{\delta}_{1\to3}(t)$, when computing the predictions for measurements at stations 2 and 3, respectively.

Our test statistic for detecting abnormal water quality events is discussed in the next section.

3.7 Triplets – Data statistic used for monitoring

This section presents the test statistic used for monitoring events at the triplets level. In this section, the setup is such that a single triplet is monitored for events, and the triplet type and operation mode is one of the four modes/types listed on Table 4 in Section 3.6. The approach detailed below is general, and is relevant to all four operation modes. Hence, it avoids mentioning the measurement times and lag times (between triplet locations) directly. We will use the following notations:

- y_1 and y_2 will denote the measured values for the target nodes in the two Pair models monitored by the triplet.
- $\hat{y}_1$ and $\hat{y}_2$ will denote the predicted values for the target nodes in the two Pair models monitored by the triplet.
- $\hat{y}_{1,L}$ and $\hat{y}_{1,U}$ will denote the lower and upper thresholds for prediction, respectively, for the target node in the first pair of the triplet. Thresholds can be computed using any coverage probability, but for this example let us assume a 95% coverage probability is used.
- Similarly, let $\hat{y}_{1,L}$ and $\hat{y}_{1,U}$ will denote the lower and upper thresholds for prediction, respectively, for the target node in the second pair of the triplet.
- Let e_1 and e_2 denote the following normalized error terms:
$$e_1 = \frac{y_1 - \hat{y}_1}{\max(\hat{y}_{1,U} - \hat{y}_{1,L}, 2M)}, \quad e_2 = \frac{y_2 - \hat{y}_2}{\max(\hat{y}_{2,U} - \hat{y}_{2,L}, 2M)},$$
where $2M$ is the minimal size for a prediction interval, as defined in Subsection 3.5 (using the user limits).

Our test statistic for monitoring events is:

$$T = \frac{1}{4} \cdot \sqrt{(e_1)^2 + \gamma(e_1)(e_2) + (e_2)^2},$$

where γ is a constant set by the user.

Some additional notes and comments:

- The parameter γ is added to the formula for T in order to "gain" from cases in which e_1 and e_2 are correlated, see for example Figure 70.
- The normalization done in the formulas for e_1 and e_2 is aimed at making T a unitless size, which is $O(1)$ (see additional details on the distribution of T below).
- A triplet model will declare a triplet alarm if the value of T exceeds a set threshold for more than a defined time period.

We have found through numerous simulations that for fork-out triplets, a value of γ in the range $(2,4)$ provides good adjustment for the correlation between e_1 and e_2.

The threshold for declaring abnormal events is set by the 95% percentile of T statistics across the most recent 30 day time period, excluding document water quality events.

In terms of typical values for the distribution of T, we have found that when taking a sample of T values with a buffer of 30 minutes or more between samples, the T statistic is typically gamma distributed. See Figure 71 for a sample of T values, and Figure 72 for a comparison of the same data sample to a Gamma distribution, with distribution parameters estimated using maximum likelihood estimators, using the R package 'fitdistplus'. Appendix B shows additional examples for times series of T statistics, and their fits to a Gamma distribution.

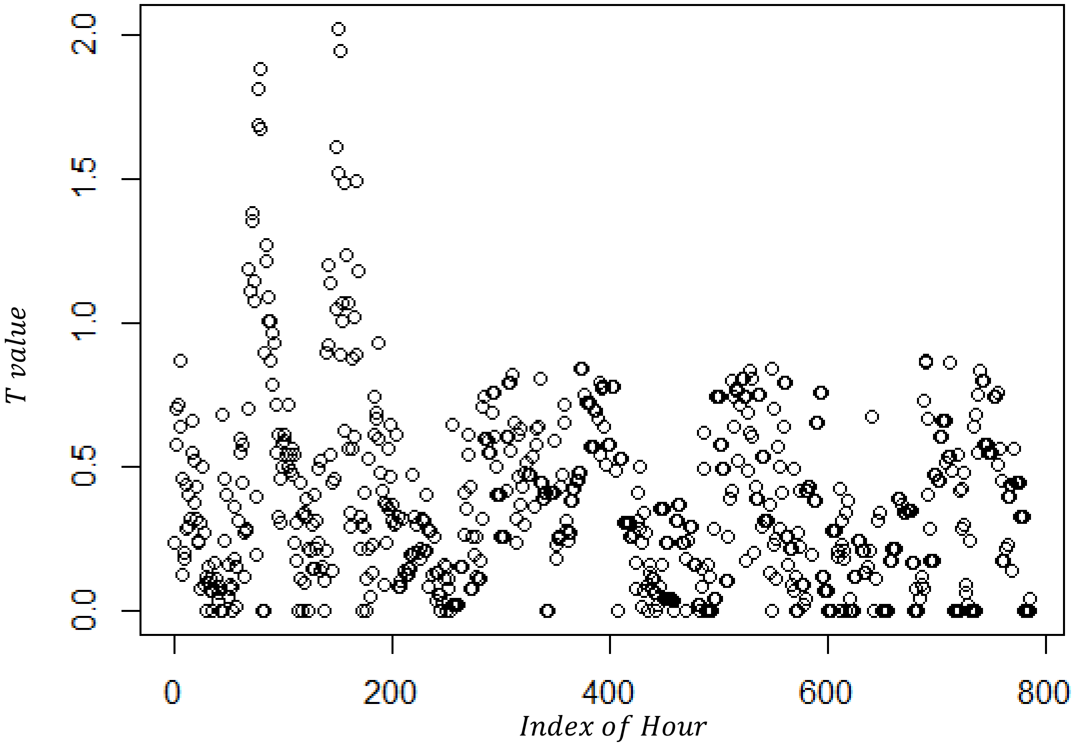


Figure 71: Time series showing the distribution of T values for a forkout triplet, across 800 hours. Taken from triplet 441.

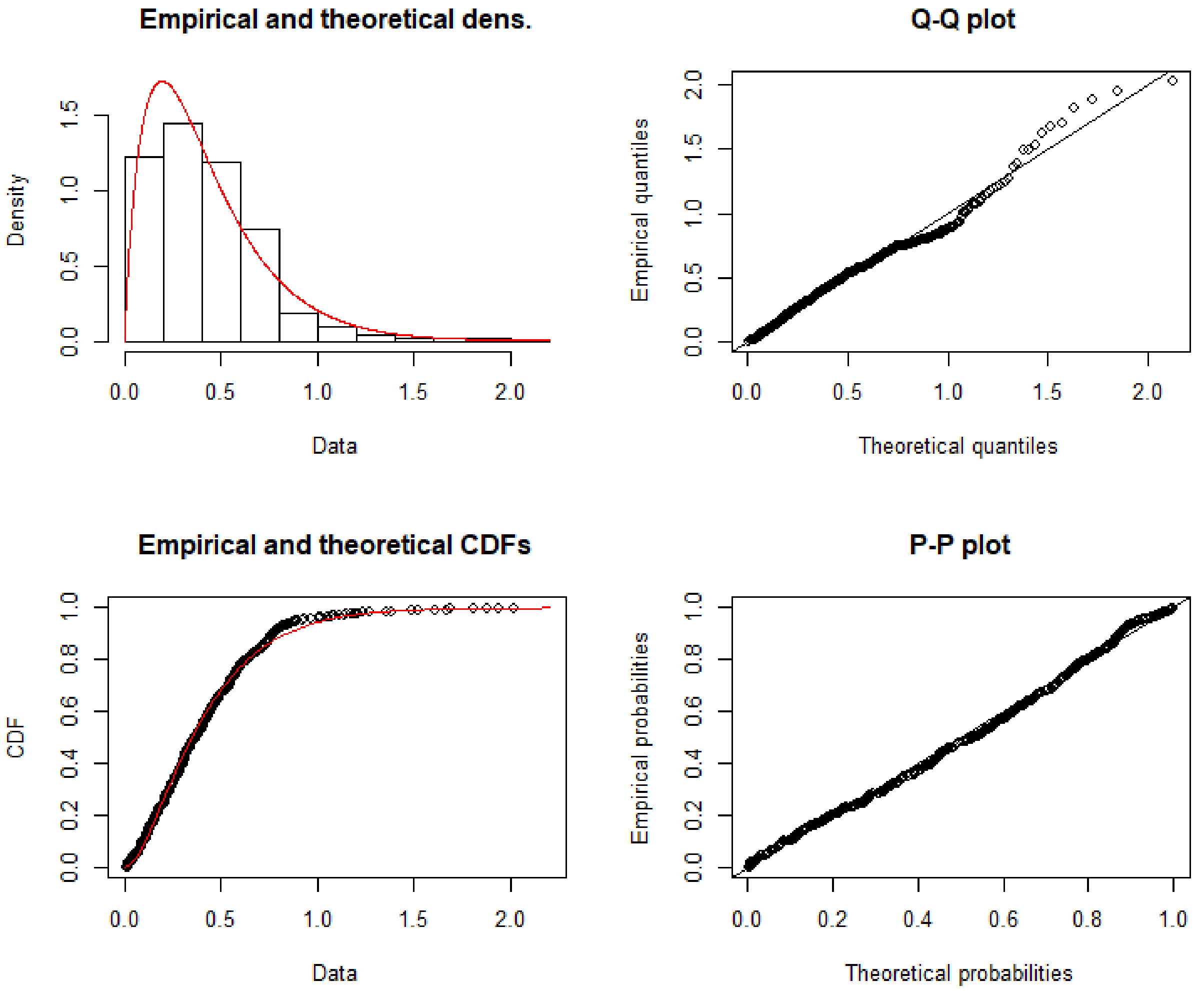
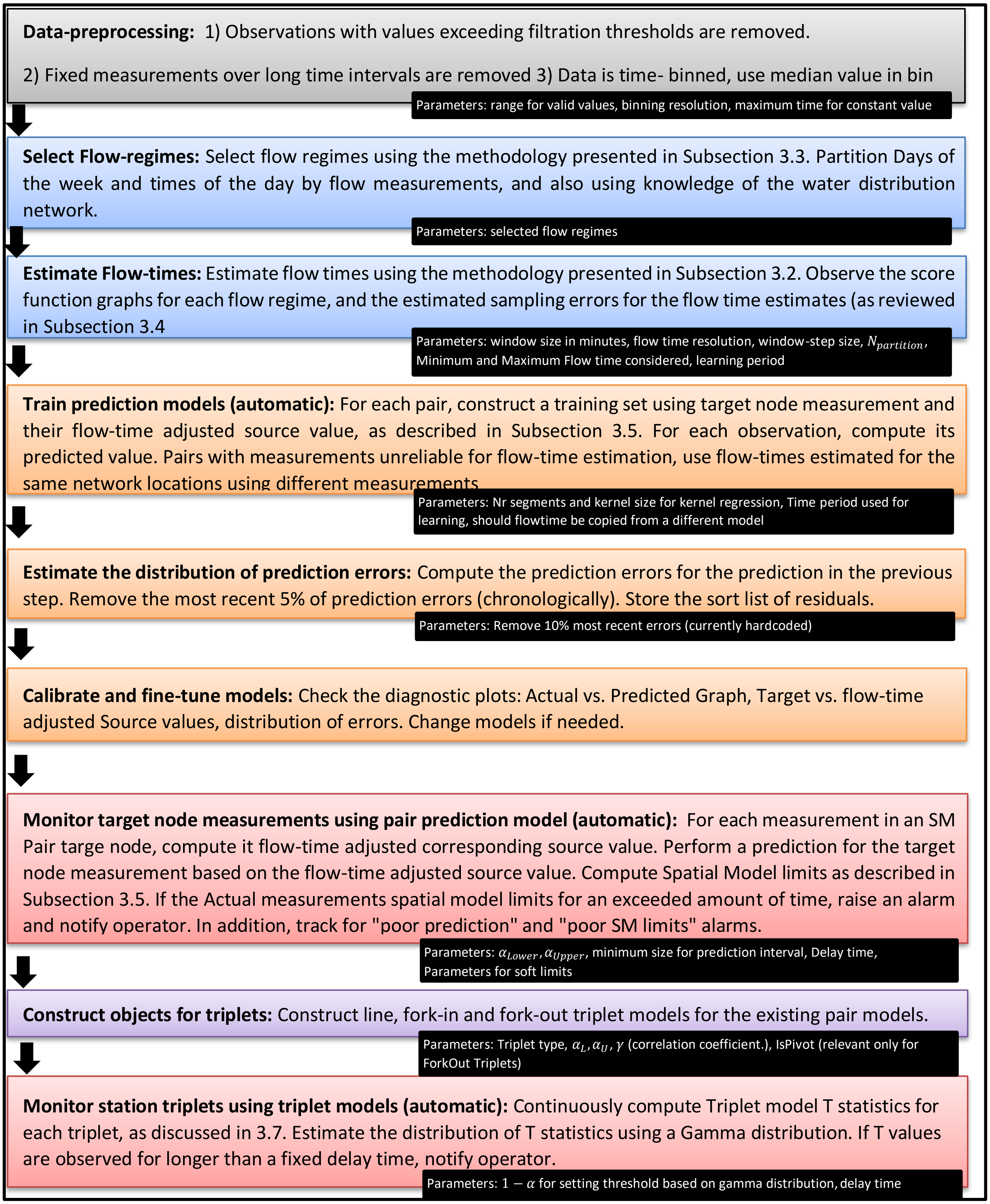


Figure 72: Comparison between the data shown in Figure 71, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.

3.8 Spatial Model – a recap on the full methodology.

Figure 73 shows a flow chart with the full SM methodology. See 4.8.1 for suggested parameter values.



Data-preprocessing: 1) Observations with values exceeding filtration thresholds are removed.

2) Fixed measurements over long time intervals are removed 3) Data is time- binned, use median value in bin

Parameters: range for valid values, binning resolution, maximum time for constant value

Select Flow-regimes: Select flow regimes using the methodology presented in Subsection 3.3. Partition Days of the week and times of the day by flow measurements, and also using knowledge of the water distribution network.

Parameters: selected flow regimes

Estimate Flow-times: Estimate flow times using the methodology presented in Subsection 3.2. Observe the score function graphs for each flow regime, and the estimated sampling errors for the flow time estimates (as reviewed in Subsection 3.4

Parameters: window size in minutes, flow time resolution, window-step size, $N_{partition}$, Minimum and Maximum Flow time considered, learning period

Train prediction models (automatic): For each pair, construct a training set using target node measurement and their flow-time adjusted source value, as described in Subsection 3.5. For each observation, compute its predicted value. Pairs with measurements unreliable for flow-time estimation, use flow-times estimated for the same network locations using different measurements

Parameters: Nr segments and kernel size for kernel regression, Time period used for learning, should flowtime be copied from a different model

Estimate the distribution of prediction errors: Compute the prediction errors for the prediction in the previous step. Remove the most recent 5% of prediction errors (chronologically). Store the sort list of residuals.

Parameters: Remove 10% most recent errors (currently hardcoded)

Calibrate and fine-tune models: Check the diagnostic plots: Actual vs. Predicted Graph, Target vs. flow-time adjusted Source values, distribution of errors. Change models if needed.

Monitor target node measurements using pair prediction model (automatic): For each measurement in an SM Pair targe node, compute it flow-time adjusted corresponding source value. Perform a prediction for the target node measurement based on the flow-time adjusted source value. Compute Spatial Model limits as described in Subsection 3.5. If the Actual measurements spatial model limits for an exceeded amount of time, raise an alarm and notify operator. In addition, track for "poor prediction" and "poor SM limits" alarms.

Parameters: α_{Lower}, α_{Upper}, minimum size for prediction interval, Delay time, Parameters for soft limits

Construct objects for triplets: Construct line, fork-in and fork-out triplet models for the existing pair models.

Parameters: Triplet type, α_L, α_U, γ (correlation coefficient.), IsPivot (relevant only for ForkOut Triplets)

Monitor station triplets using triplet models (automatic): Continuously compute Triplet model T statistics for each triplet, as discussed in 3.7. Estimate the distribution of T statistics using a Gamma distribution. If T values are observed for longer than a fixed delay time, notify operator.

Parameters: $1-\alpha$ for setting threshold based on gamma distribution, delay time

Figure 73: A flowchart with the SM methodology. Parameters for each step are detailed in small, white on black text boxes.

3.9 Quiz for Unit B

1) What types of object types are available in the Spatial Model?

2) Explain the importance of flow regime selection and flow time estimation. What are the dangers of large errors in estimated flow times?

3) Explain the general idea for the algorithm used to estimate flow times from the data

4) How are flow regimes selected? mention several methodologies are auxiliary measurements that can help in the selection of flow regimes.

5) What measurements are best used for the estimation of flow times? what can be done in SM pair model that are not based on these measurements?

6) What is the sampling error for flow time estimators? how can it be estimated?

7) Sort the following terms, into their methodological order, as performed when constructing prediction models:

- Collect and sort the distribution of errors, remove errors from the last portion of the data, e.g. most recent 10% of the data.
- Create a prediction value for each pair of measurements.
- For each observed target measurement, find it's associated flow-time adjusted source value.
- For each observed target measurement, find it's relevant flow time.

8) Sort the following terms, into their methodological order, as performed when monitoring for events in a target station:

- If alarm duration passes a prespecified delay time, notify operator.
- Compare the observed measurement to the prediction interval. If the measurement is outside the prediction interval, start a "limits violation event".
- Adjust prediction interval to be at least the required minimum size.
- Use the historical distribution of errors to compute a prediction interval
- Use the lag-time adjusted source value for performing a prediction.
- Find the lag-time adjusted source value.
- Find lag time for current target station date time.

9) What is the danger of a model with poor prediction capabilities?

10) Explain in what scenarios may triplet sensors be useful.

4. Unit C – The Spatial Model Software

This Section describes the Spatial Model module of the EDS. Screenshots in this section were taken from a working Spatial Model system, and model names were changed to Station A, Station B, Station C and so on, so that real locations are not identifiable.

System Objects and life cycle:

Pair models are defined using existing EDS models for network stations, and Triplet models are defined using existing Pair models. A pair model can be in one of four possible states: Stopped, Running, Flow time estimation and Performing prediction. Models that are in run state are evaluated once every minute. If new records are available for the target node, predictions for target value are computed and compared to the actual values measured. If the actual value measured exceeds the computed thresholds, an alarm is raised. The SM module retrains the flow time for pair models once every 24 hours or when manually requested. Note that a pair model may use another pair model as a reference for flow time estimation, e.g. a Turbidity model may be unreliable for flow time estimation and may use the Conductivity based model between the same two stations for flow time estimates. Model that have been stopped are not evaluated.

Types of Alarms

The spatial model can raise one of the following alarms:

- **Limits violation –** A measurement is the target node of a pair has exceeded it's computed limits
- **Poor limit –** The ratio $\frac{\text{Range between Spatial limits}}{Range\ between\ Statistical\ limits}$ for a pair is larger than 50%.
- **Poor prediction –** The ratio $\frac{|Predicted\ target\ node\ value - Actual\ target\ node\ value|}{Actual\ target\ node\ value}$ for a pair is larger than 25%.
- **No data-** A pair object has no data available in either the Target or Source station.

The above four alarms are associated with a pair model.

- **Triplets alarms –** A triplet's T value has exceeded it's computed threshold.

Multiple alarm types can identify a single event, since several of the above conditions can be met simultaneously. **However, the user will only be notified only of events including a "limits violation" type alarm.**

<u>Network States</u>

In addition to the different flow-regimes defined at the pair level, the Spatial Model module supports network-wide states. Currently, the following states are available for the monitored network:

State E: Water supplied from an external source

State R: Water supplied from reservoirs

State E→R: Network is transitioning from state E to R.

State R→E: Network is transitioning from state R to E.

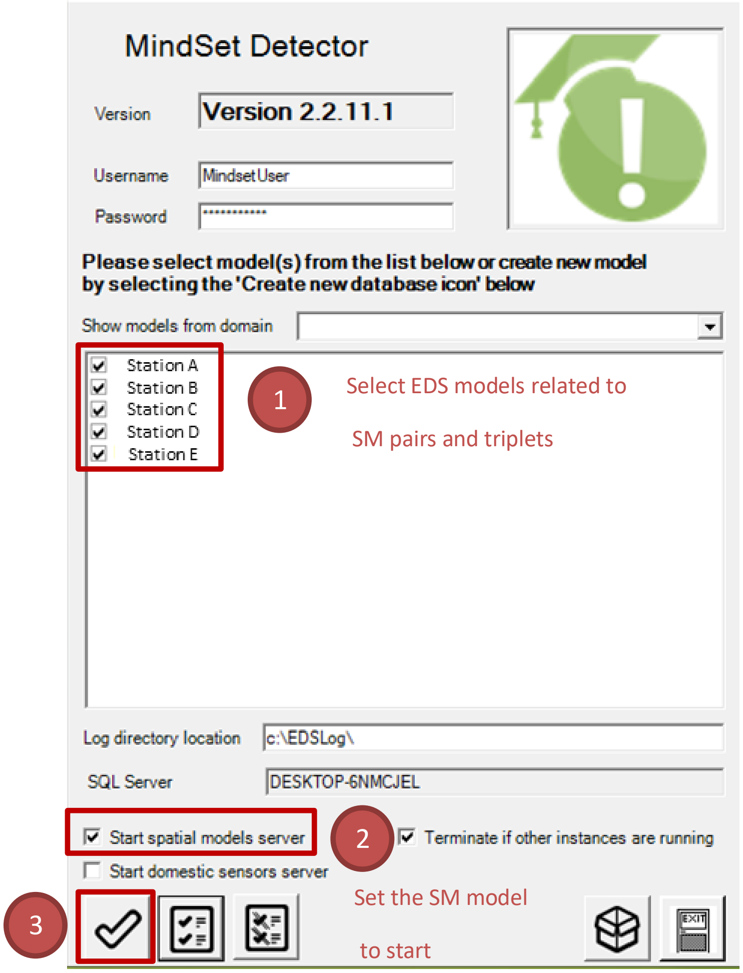
The state machine and transition rules for these four states is fully discussed in Section 4.2. Pair models may have a set time period where no alarms are raised after a Network State change, see Section 4.3. Modes may be set to monitor and raise alarms in state E, R or both. For transitions states, no alarms are raised.

4.1 Intro to the SM software and main software panels

This section describes the main screens and software panels of the SM software module. To gain access to the SM screen, begin by logging in to the MindSet Detector EDS system. Figure 74 shows the log in screen of the MindSet Detector. To log in:

- Select all station related to the SM, see the label "1" in Figure 74 (you also just select all stations)
- Select the option "Start spatial models server", see the label "2" in Figure 74.
- Click the "OK" icon, see the label "3" in Figure 74.

After log-in, you will see the main EDS screen, as in Figure 75.



MindSet Detector

Version: Version 2.2.11.1

Username: MindsetUser

Password: ************

Please select model(s) from the list below or create new model by selecting the 'Create new database icon' below

Show models from domain

☑ Station A
☑ Station B
☑ Station C
☑ Station D
☑ Station E

Log directory location: c:\EDSLog\

SQL Server: DESKTOP-6NMCJEL

☑ Start spatial models server

☑ Terminate if other instances are running

□ Start domestic sensors server

Select EDS models related to SM pairs and triplets

Set the SM model to start

Click the 'OK/Accept"' button

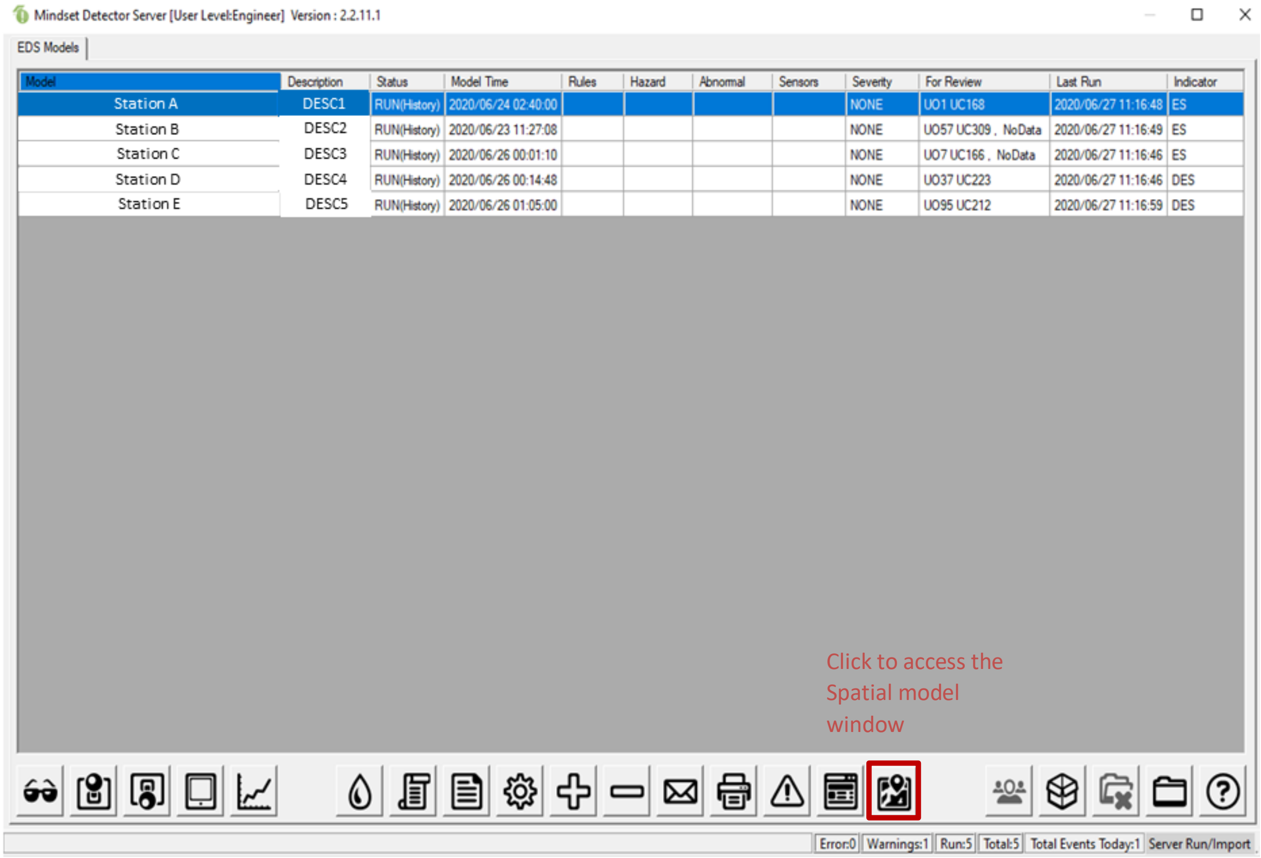
Figure 74: The EDS log-in screen. Steps to log in marked in labels

Figure 75 shows the loaded EDS models and their current state. For each model the following key fields are presented:

- Model – gives the model name, used to identify the model in all screens
- Description – a description of the model, in simple English
- Status – current status. Model may be either running or stopped. Transition between states can be done using the 'lever up'/'lever down' button at the bottom of the screen.
- Model Time – Last data timestamp processed by the EDS model.
- Last Run – Last time model was evaluated, in server time.

Details about additional fields, toolbars and EDS screens can be found in the EDS manual.

Click on the "Spatial Model window" icon, depicting a map (and showed in red in Figure 75) to access the main Spatial Model window.



Mindset Detector Server [User Level:Engineer] Version : 2.2.11.1
EDS Models

Model	Description	Status	Model Time	Rules	Hazard	Abnormal	Sensors	Severity	For Review	Last Run	Indicator
Station A	DESC1	RUN(History)	2020/06/24 02:40:00					NONE	UO1 UC168	2020/06/27 11:16:48	ES
Station B	DESC2	RUN(History)	2020/06/23 11:27:08					NONE	UO57 UC309 , NoData	2020/06/27 11:16:49	ES
Station C	DESC3	RUN(History)	2020/06/26 00:01:10					NONE	UO7 UC166 , NoData	2020/06/27 11:16:46	ES
Station D	DESC4	RUN(History)	2020/06/26 00:14:48					NONE	UO37 UC223	2020/06/27 11:16:46	DES
Station E	DESC5	RUN(History)	2020/06/26 01:05:00					NONE	UO95 UC212	2020/06/27 11:16:59	DES

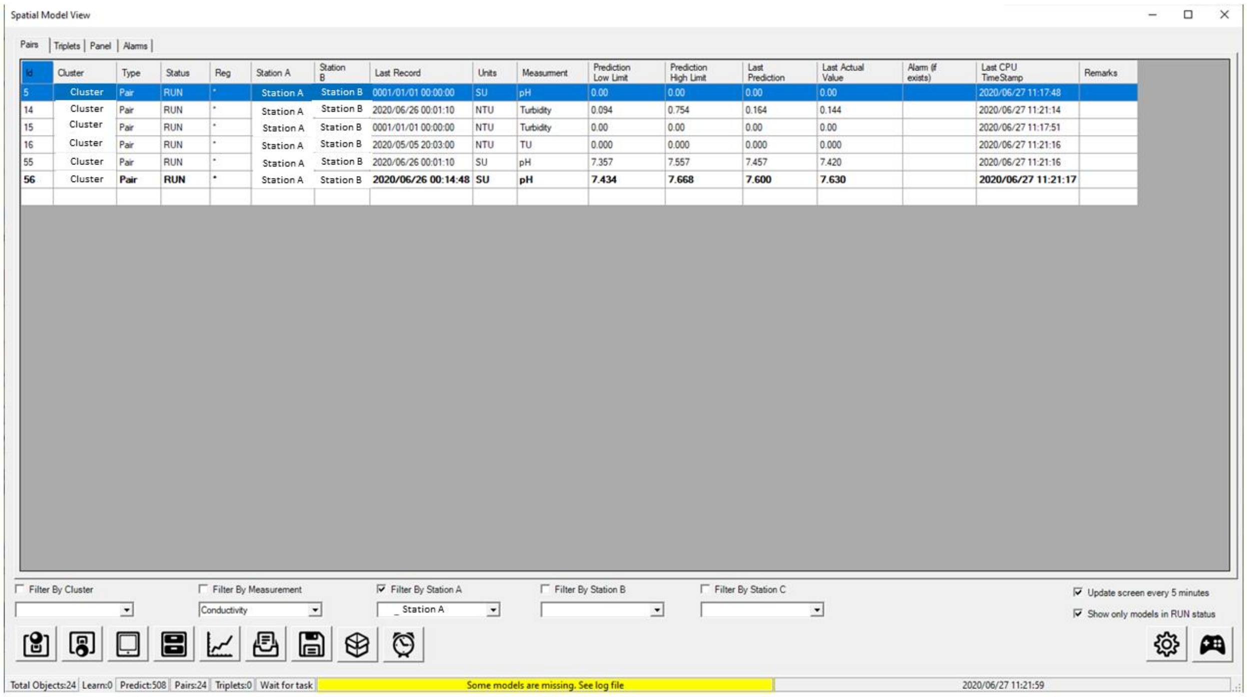
Click to access the Spatial model window

Figure 75: The EDS models view

Figure 76 shows the Spatial Model main window. The window features four views separated by tabs, filter selection boxes, a tool bar and a status panel. The four views are:

- Pairs tab- shows the list of pair models and their current state. The pairs tab is shows in Figure 76
- Triplets tab- shows the list of triplet models and their current state. Triplets and their respective tab are discussed in Section 4.9.
- Panel tab – shows a " birds eye view" of all the models. The panel tab is shown in Figure 79.
- Alarms tab- shows all alarms opened by the Spatial Model. The alarms tab is shown in Figures 81-82.

Next, we survey the pairs tab, toolbar, filter selection boxes, status bar, Panel tab and alarms tab, in this order.



Id	Cluster	Type	Status	Reg	Station A	Station B	Last Record	Units	Measurment	Prediction Low Limit	Prediction High Limit	Last Prediction	Last Actual Value	Alarm (if exists)	Last CPU TimeStamp	Remarks
5	Cluster	Pair	RUN	*	Station A	Station B	0001/01/01 00:00:00	SU	pH	0.00	0.00	0.00	0.00		2020/06/27 11:17:48	
14	Cluster	Pair	RUN	*	Station A	Station B	2020/06/26 00:01:10	NTU	Turbidity	0.094	0.754	0.164	0.144		2020/06/27 11:21:14	
15	Cluster	Pair	RUN	*	Station A	Station B	0001/01/01 00:00:00	NTU	Turbidity	0.00	0.00	0.00	0.00		2020/06/27 11:17:51	
16	Cluster	Pair	RUN	*	Station A	Station B	2020/05/05 20:03:00	NTU	TU	0.000	0.000	0.000	0.000		2020/06/27 11:21:16	
55	Cluster	Pair	RUN	*	Station A	Station B	2020/06/26 00:01:10	SU	pH	7.357	7.557	7.457	7.420		2020/06/27 11:21:16	
56	Cluster	**Pair**	**RUN**	*	Station A	Station B	**2020/06/26 00:14:48**	**SU**	**pH**	**7.434**	**7.668**	**7.600**	**7.630**		**2020/06/27 11:21:17**	

□ Filter By Cluster
☑ Filter By Measurement: Conductivity (unchecked box □)
☑ Filter By Station A: _ Station A
□ Filter By Station B
□ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:24 | Learn:0 | Predict:508 | Pairs:24 | Triplets:0 | Wait for task | Some models are missing. See log file | 2020/06/27 11:21:59

Figure 76: The Spatial Model Pairs view

Pairs tab

The pairs tab shows a table with rows corresponding to different pairs model, and the following fields:

- Id- A unique numeric identifier for the link
- Cluster – A string describing the cluster the link is in. Clusters are groups of links identified by a common name. Every link must be a member of a single cluster
- Type- The type of SM model, "Pair" for all models in this view
- Status – The current status of the model: stopped, running, estimating flow time or performing prediction.
- Reg – The network flow regimes in which the model is active, see Section 4.3.
- Station A – The name of the source station for this link.
- Station B – The name of the target station for this link.
- Last record – The timestamp for the last record processed in the target node.
- Units – The units of measurement
- Measurement – The name of the monitored measurement.

- Prediction Low/High Limit – The upper and lower thresholds computed for monitoring the target station.
- Last Prediction – The last prediction given for the target node measurement for this link.
- Last Actual value – The value measured at the target station at the timestamp indicated by "Last Record".
- Alarm (if exists) – Describes the current alarms opened by this model.
- Last CPU timestamp – Describes the last timestamp (in server time) this model was evaluated by the SM software.
- Remarks – Additional remarks.

Toolbar

The toolbar at the bottom of the Spatial Model window contains the following options:

Set selected model state to "run"

Set selected model state to

Open selected model in model editor

Refresh view

View graphs for selected model

Generate reports

Export models to files

Define a new model (in model editor)

Set flow regimes and network regimes

Server setup

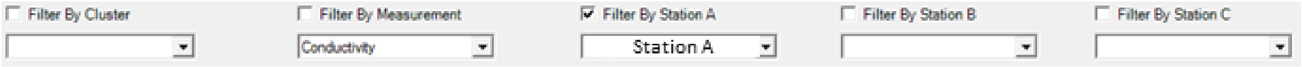
Run "what-if" simulations

The model editor is discussed in detail in Sections 4.3-4.5. The Graphs window is discussed in 4.6. Screens for Flow and network regimes are discussed in Section 4.2. The simulations screen is discussed in 4.10.

Filter selection boxes

Figure 77 shows the filter selection boxes available on the SM screens.
Models in the pairs tab can be filtered by cluster (group of models). Measurement name, source station (A), and target station (B). Triplets can be filtered by any of the three Stations in the model, see additional details in Section 4.9. To filter the rows:

1. Select filters **using the drop down menus**.
2. Selecting which filters are activated **using the checkboxes**
3. Click the **"Refresh view" button** to immediately filter the results.



☐ Filter By Cluster

☐ Filter By Measurement: Conductivity

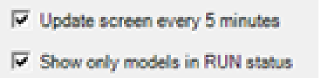
☑ Filter By Station A: Station A

☐ Filter By Station B

☐ Filter By Station C

Figure 77: Filters for the Spatial Model window

Figure 78 shows additional options for filtration and viewing. The first option in Figure HHH sets whether the view will automatically refresh once every 5 minutes. If turned off, the view will refresh only when the user clicks the "Refresh view" button. The second option in Figure 78 sets whether all models are displayed, or only models in "Run" status (including flow time estimation and prediction).



☑ Update screen every 5 minutes

☑ Show only models in RUN status

Figure 78: Additional Filters for the Spatial model window

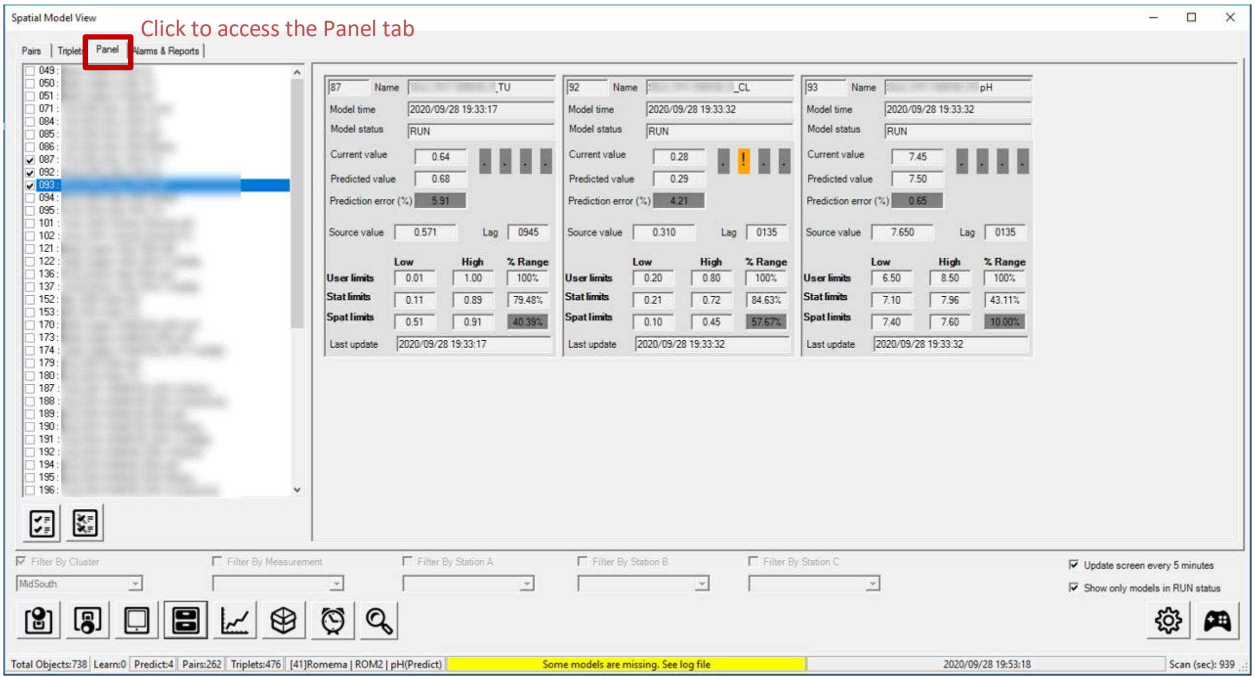
Status bar

The status bar found at the bottom of the screen features the following fields, from left to right:

- Total number of objects loaded.
- The number of objects performing flow time estimation, i.e., learning
- The total number of pairs performing prediction/ activated (HHH).
- The total number of pairs objects
- The total number of triplet objects
- A status line showing key messages, such as HHH
- The current server date and time

Panel tab

The panel tab shows an alternate view of all pair models, along with their state and common key indicators. Figure 79 shows the Panel tab, while Figure 80 shows a "zoomed-in" view of one of the pairs in the panel tab. Next, we describe the different fields available for each Pair model in this tab.



Click to access the Panel tab

□ 049 :
□ 050 :
□ 051 :
□ 071 :
□ 084 :
□ 085 :
□ 086 :
☑ 087 :
☑ 092 :
☑ 093 :
□ 094 :
□ 095 :
□ 101 :
□ 102 :
□ 121 :
□ 122 :
□ 136 :
□ 137 :
□ 152 :
□ 153 :
□ 170 :
□ 173 :
□ 174 :
□ 179 :
□ 180 :
□ 187 :
□ 188 :
□ 189 :
□ 190 :
□ 191 :
□ 192 :
□ 194 :
□ 195 :
□ 196 :

87	Name	[redacted]_TU	
Model time	2020/09/28 19:33:17		
Model status	RUN		
Current value	0.64		
Predicted value	0.68		
Prediction error (%)	5.91		
Source value	0.571	Lag	0945
	Low	High	% Range
User limits	0.01	1.00	100%
Stat limits	0.11	0.89	79.48%
Spat limits	0.51	0.91	40.39%
Last update	2020/09/28 19:33:17		

92	Name	[redacted]_CL	
Model time	2020/09/28 19:33:32		
Model status	RUN		
Current value	0.28		
Predicted value	0.29		
Prediction error (%)	4.21		
Source value	0.310	Lag	0135
	Low	High	% Range
User limits	0.20	0.80	100%
Stat limits	0.21	0.72	84.63%
Spat limits	0.10	0.45	57.67%
Last update	2020/09/28 19:33:32		

93	Name	[redacted]pH	
Model time	2020/09/28 19:33:32		
Model status	RUN		
Current value	7.45		
Predicted value	7.50		
Prediction error (%)	0.65		
Source value	7.650	Lag	0135
	Low	High	% Range
User limits	6.50	8.50	100%
Stat limits	7.10	7.96	43.11%
Spat limits	7.40	7.60	10.00%
Last update	2020/09/28 19:33:32		

☑ Filter By Cluster: MidSouth
□ Filter By Measurement
□ Filter By Station A
□ Filter By Station B
□ Filter By Station C

☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:738 | Learn:0 | Predict:4 | Pairs:262 | Triplets:476 | [41]Romema | ROM2 | pH(Predict) | Some models are missing. See log file | 2020/09/28 19:53:18 | Scan (sec): 939

Figure 79: The Spatial Model panel view.

The different fields for each Pair model in the Panel tab, as in Figure HHH, are:

- The **name** of the Pair model

- The current model **time** (last processed record)

- Current model status (Run/ Stop/ Flow time learning/ Prediction)

- The **current value** in the target station

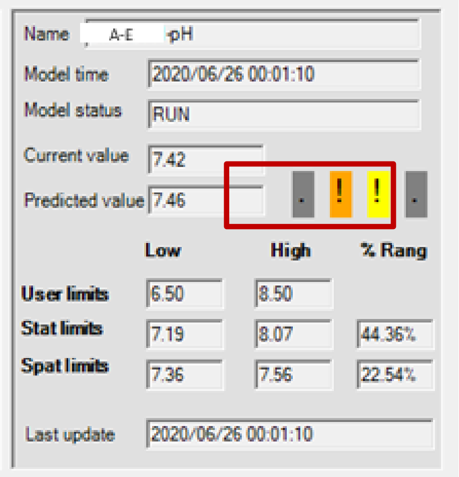
- The **predicted value** for the target station, based on the source station.

- **Alarm indicators**, from left to right: (marked by a red rectangle in Figure HHH)
    - Limits violation – marked by a red rectangle with an exclamation sign (turned off in Figure HHH).
    - Poor limits – marked by an orange rectangle with an exclamation sign (turned on in Figure HHH)
    - Poor prediction – marked by a yellow rectangle with an exclamation sign (turned on in Figure HHH)
    - No data- marked by a blue rectangle with an exclamation sign (turned off in Figure HHH).

- The **user limits**, as defined for the monitored measurement in the EDS model for the target node.

- The **statistical limits** for the target node measurement, as computed by the statistical limits detector in the EDS model for the target node. The **percentage** found to the right of limits gives the ratio between the range of the statistical limits and the user limits. Numbers under 100% represent an improvement in the statistical limits compared to the user limits.

- The **spatial limits** for the target node measurement, as computed by the Spatial Model software. The **percentage** found to the right of the spatial limits gives the ratio between the spatial model limits and the statistical limits. Numbers under 100% represent an improvement in the spatial limits compared to the statistical limits.



Name A-E pH

Model time 2020/06/26 00:01:10

Model status RUN

Current value 7.42

Predicted value 7.46

	Low	High	% Rang
User limits	6.50	8.50	
Stat limits	7.19	8.07	44.36%
Spat limits	7.36	7.56	22.54%

Last update 2020/06/26 00:01:10

-

Figure 80: A Spatial Model panel item corresponding to a single Pair model. Alarm indicator are marked by a red rectangle.

The different fields, such as percentages and alarms, feature a tool tip text label stating their computation.

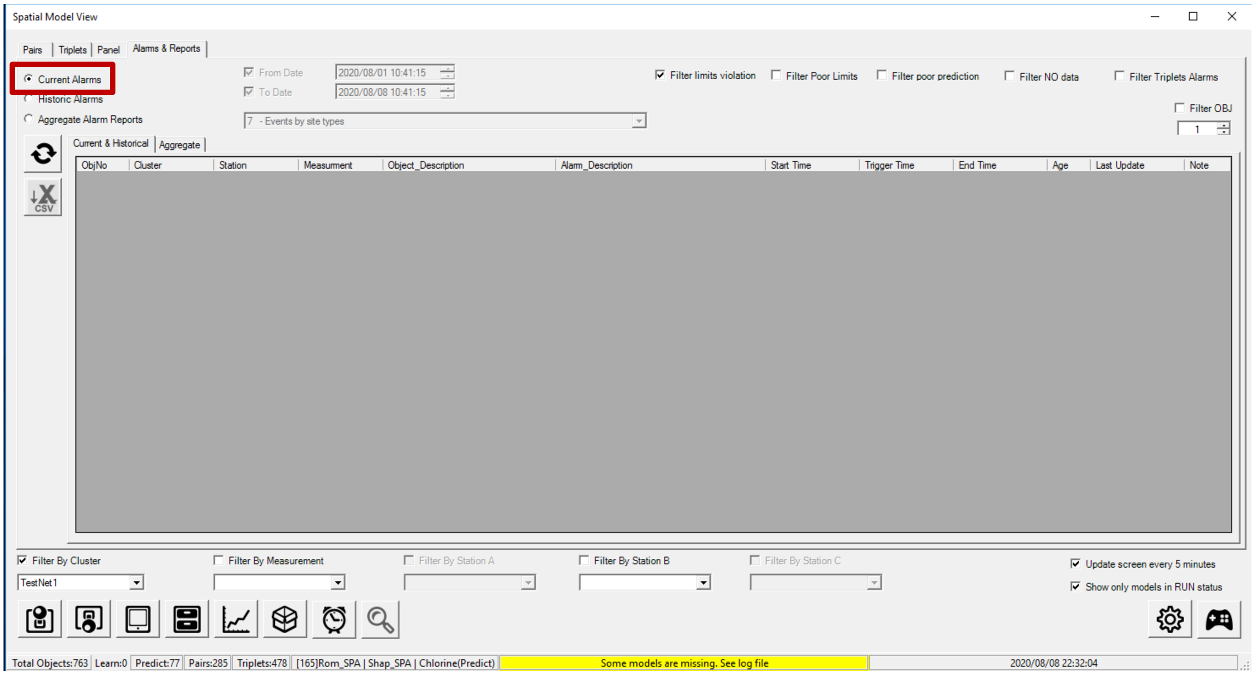
Alarms tab

The alarms tab, shown in Figure 81, shows the different the current open and historic alarms raised by the Spatial Model software. Users can view current or historic alarms, and filter alarms by type and location:

1. Switch between alarm types using the option boxes names "Current Alarms" and "Historic Alarms".
2. For historic Alarms, you may filter by a "From data" and a "To Date". Filter are activated using the checkboxes. Date selection is using the date and time field for each filter.
3. Filter alarm by type, using the check box fields on the top right. Alarms filters correspond to the 5 alarm types presented at the start of Unit C.
4. Since alarms may be raised from different objects, The user may filter alarms by clusters, measurements and stations, using the filter selection boxes at the bottom of the screen.
5. After performing 1-4 (some steps optional), click on the refresh button found on the left of the screen (⟳).

Alarms will be presented in the table at the center of the screen. Alarms will feature:

- The object Number and Description (for the object generating the alarm)
- The description for the alarm
- Alarm start, trigger and end times. The trigger time is the start time + the delay time for alarm see Sections 4.3-4.4 for discussion of delay times.
- The alarm age, in Minutes.
- The time of last update to the alarm record and additional notes.



● Current Alarms
○ Historic Alarms
○ Aggregate Alarm Reports

☑ From Date 2020/08/01 10:41:15
☑ To Date 2020/08/08 10:41:15
7 - Events by site types

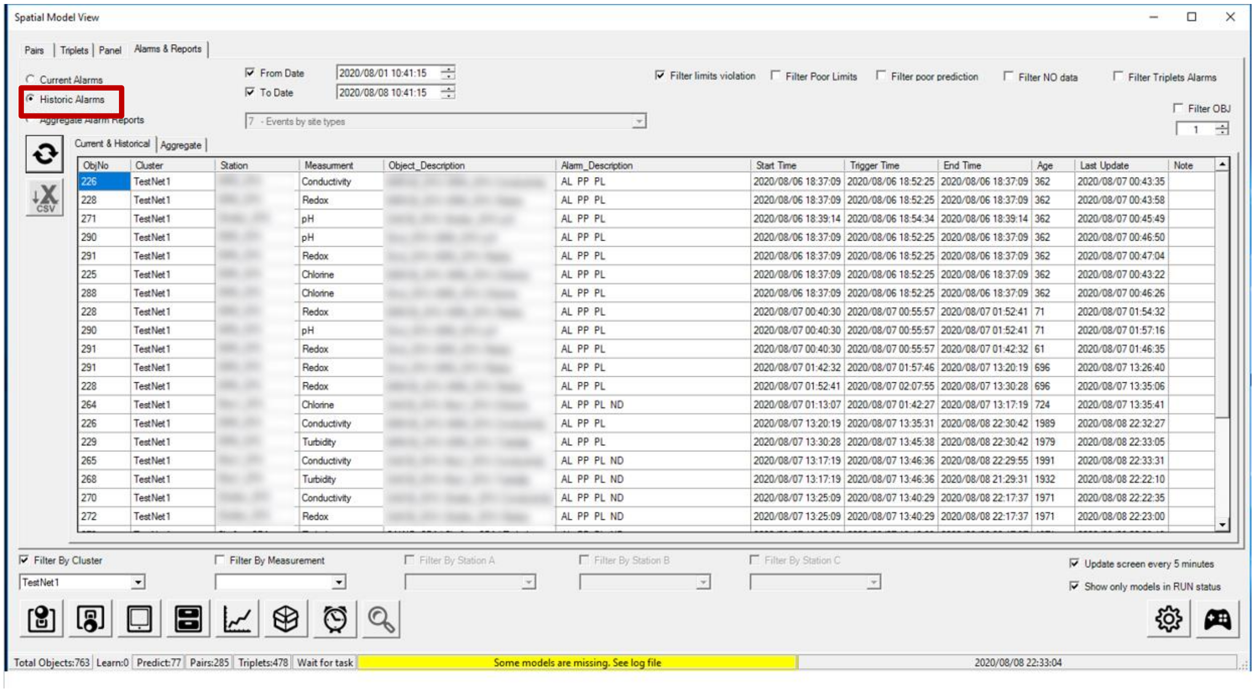
☑ Filter limits violation
□ Filter Poor Limits
□ Filter poor prediction
□ Filter NO data
□ Filter Triplets Alarms
□ Filter OBJ 1

ObjNo	Cluster	Station	Measurment	Object_Description	Alarm_Description	Start Time	Trigger Time	End Time	Age	Last Update	Note

☑ Filter By Cluster TestNet1
□ Filter By Measurement
□ Filter By Station A
□ Filter By Station B
□ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:763 | Learn:0 | Predict:77 | Pairs:285 | Triplets:478 | [165]Rom_SPA | Shap_SPA | Chlorine(Predict) | Some models are missing. See log file | 2020/08/08 22:32:04

Figure 81: The alarms view



○ Current Alarms
● Historic Alarms
○ Aggregate Alarm Reports

☑ From Date 2020/08/01 10:41:15
☑ To Date 2020/08/08 10:41:15

7 - Events by site types

☑ Filter limits violation ☐ Filter Poor Limits ☐ Filter poor prediction ☐ Filter NO data ☐ Filter Triplets Alarms

☐ Filter OBJ 1

ObjNo	Cluster	Station	Measurment	Object_Description	Alarm_Description	Start Time	Trigger Time	End Time	Age	Last Update	Note
226	TestNet1	[redacted]	Conductivity	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:43:35	
228	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:43:58	
271	TestNet1	[redacted]	pH	[redacted]	AL PP PL	2020/08/06 18:39:14	2020/08/06 18:54:34	2020/08/06 18:39:14	362	2020/08/07 00:45:49	
290	TestNet1	[redacted]	pH	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:46:50	
291	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:47:04	
225	TestNet1	[redacted]	Chlorine	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:43:22	
288	TestNet1	[redacted]	Chlorine	[redacted]	AL PP PL	2020/08/06 18:37:09	2020/08/06 18:52:25	2020/08/06 18:37:09	362	2020/08/07 00:46:26	
228	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/07 00:40:30	2020/08/07 00:55:57	2020/08/07 01:52:41	71	2020/08/07 01:54:32	
290	TestNet1	[redacted]	pH	[redacted]	AL PP PL	2020/08/07 00:40:30	2020/08/07 00:55:57	2020/08/07 01:52:41	71	2020/08/07 01:57:16	
291	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/07 00:40:30	2020/08/07 00:55:57	2020/08/07 01:42:32	61	2020/08/07 01:46:35	
291	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/07 01:42:32	2020/08/07 01:57:46	2020/08/07 13:20:19	696	2020/08/07 13:26:40	
228	TestNet1	[redacted]	Redox	[redacted]	AL PP PL	2020/08/07 01:52:41	2020/08/07 02:07:55	2020/08/07 13:30:28	696	2020/08/07 13:35:06	
264	TestNet1	[redacted]	Chlorine	[redacted]	AL PP PL ND	2020/08/07 01:13:07	2020/08/07 01:42:27	2020/08/07 13:17:19	724	2020/08/07 13:35:41	
226	TestNet1	[redacted]	Conductivity	[redacted]	AL PP PL	2020/08/07 13:20:19	2020/08/07 13:35:31	2020/08/08 22:30:42	1989	2020/08/08 22:32:27	
229	TestNet1	[redacted]	Turbidity	[redacted]	AL PP PL	2020/08/07 13:30:28	2020/08/07 13:45:38	2020/08/08 22:30:42	1979	2020/08/08 22:33:05	
265	TestNet1	[redacted]	Conductivity	[redacted]	AL PP PL ND	2020/08/07 13:17:19	2020/08/07 13:46:36	2020/08/08 22:29:55	1991	2020/08/08 22:33:31	
268	TestNet1	[redacted]	Turbidity	[redacted]	AL PP PL ND	2020/08/07 13:17:19	2020/08/07 13:46:36	2020/08/08 21:29:31	1932	2020/08/08 22:22:10	
270	TestNet1	[redacted]	Conductivity	[redacted]	AL PP PL ND	2020/08/07 13:25:09	2020/08/07 13:40:29	2020/08/08 22:17:37	1971	2020/08/08 22:22:35	
272	TestNet1	[redacted]	Redox	[redacted]	AL PP PL ND	2020/08/07 13:25:09	2020/08/07 13:40:29	2020/08/08 22:17:37	1971	2020/08/08 22:23:00	

☑ Filter By Cluster TestNet1
☐ Filter By Measurement
☐ Filter By Station A
☐ Filter By Station B
☐ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:763 | Learn:0 | Predict:77 | Pairs:285 | Triplets:478 | Wait for task | Some models are missing. See log file | 2020/08/08 22:33:04

Figure 82:Viewing Historic Alarms in the Alarms and Reports tab.

4.2 Analyzing and choosing flow regimes.

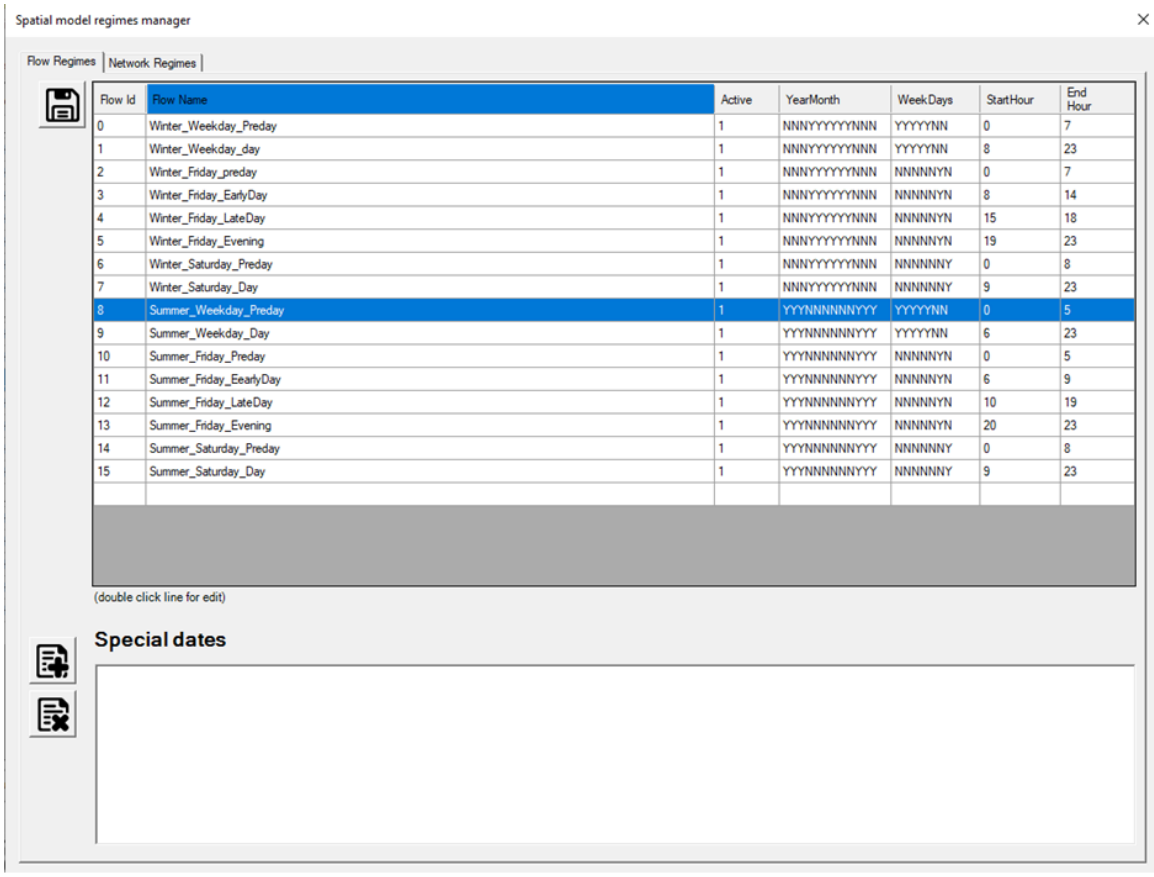
This section describes how flow regimes and network regimes are defined using the Spatial Model software user interface. To recap, flow regimes are recurring time periods, i.e. times of the day and day of the week, that exhibit similar flow times between neighboring stations. Therefore, flow time estimates are provided based on the different flow regimes. Network regimes are network states describing if water is supplied from network-wide water reservoirs, external water sources, or the network is transitioning between states. Network regimes are used to activate/ deactivate pair models, since some pair models may be relevant only for some of the network regimes. See additional details in Section 4.3.

4.2.1 Flow regime

The Spatial Model regimes manager can be accessed by pressing the 'clock' icon available under the toolbar in the main Spatial Model window. The Icon can be seen at Figure 83.

Figure 83: The icon for the Spatial Model regimes manager. Found in the main Spatial Model window toolbar

The Spatial Model regimes manager is shown in Figure 84. We describe the different on-screen elements on the next page.



Flow Id	Flow Name	Active	YearMonth	WeekDays	StartHour	End Hour
0	Winter_Weekday_Preday	1	NNNYYYYYYNNN	YYYYYNN	0	7
1	Winter_Weekday_day	1	NNNYYYYYYNNN	YYYYYNN	8	23
2	Winter_Friday_preday	1	NNNYYYYYYNNN	NNNNNYN	0	7
3	Winter_Friday_EarlyDay	1	NNNYYYYYYNNN	NNNNNYN	8	14
4	Winter_Friday_LateDay	1	NNNYYYYYYNNN	NNNNNYN	15	18
5	Winter_Friday_Evening	1	NNNYYYYYYNNN	NNNNNYN	19	23
6	Winter_Saturday_Preday	1	NNNYYYYYYNNN	NNNNNNY	0	8
7	Winter_Saturday_Day	1	NNNYYYYYYNNN	NNNNNNY	9	23
8	Summer_Weekday_Preday	1	YYYNNNNNNYYY	YYYYYNN	0	5
9	Summer_Weekday_Day	1	YYYNNNNNNYYY	YYYYYNN	6	23
10	Summer_Friday_Preday	1	YYYNNNNNNYYY	NNNNNYN	0	5
11	Summer_Friday_EearlyDay	1	YYYNNNNNNYYY	NNNNNYN	6	9
12	Summer_Friday_LateDay	1	YYYNNNNNNYYY	NNNNNYN	10	19
13	Summer_Friday_Evening	1	YYYNNNNNNYYY	NNNNNYN	20	23
14	Summer_Saturday_Preday	1	YYYNNNNNNYYY	NNNNNNY	0	8
15	Summer_Saturday_Day	1	YYYNNNNNNYYY	NNNNNNY	9	23

(double click line for edit)

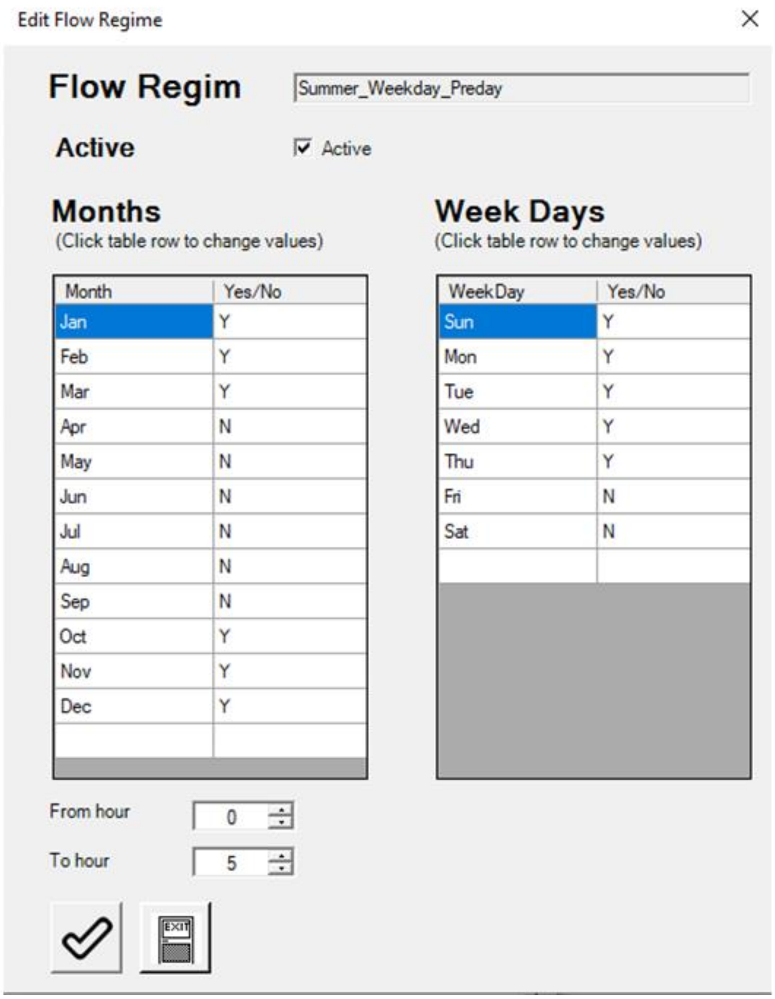
Special dates

Figure 84: The Spatial Model regimes manager

The top table on Figure 84 shows the available from regimes (indexed starting from 0), along with their names and recurring time periods. Recurring time periods for flow regimes are defined by the months of the year, days of the week and hours of the day. The Letters 'Y' and 'N' denote which days and months are active for the regime (Y) and inactive (N). It is up to the user to verify that inserted regimes cover all times and dates, and that each time-date combination is covered by a single flow regime. To edit a flow regime, double click on its row. This will bring up the Flow Regime editor shown in Figure 85. After done editing all flow regimes, click the save (diskette) icon.

Special dates - forcing calendrical dates of one type to become another type, e.g. to mark flow regimes for holidays, are defined using the 'special dates' list box at the bottom of the screen.

Figure 85 shows the Flow regime editor. This editor shows: the regime name; if it is activated (deactivated regimes are not considered for flow time estimation); and the months, week days and hours the regime is active. Activated months and days are shown in two separate tables. Click on the table rows to change table values (Y/N). The starting and ending hours for the regime are defined by the selection boxes at the bottom of the screen. Click the OK or cancel buttons at the bottom of the screen to approve or cancel your choice.



Edit Flow Regime

Flow Regim: Summer_Weekday_Preday

Active: ☑ Active

Months
(Click table row to change values)

Month	Yes/No
Jan	Y
Feb	Y
Mar	Y
Apr	N
May	N
Jun	N
Jul	N
Aug	N
Sep	N
Oct	Y
Nov	Y
Dec	Y

Week Days
(Click table row to change values)

WeekDay	Yes/No
Sun	Y
Mon	Y
Tue	Y
Wed	Y
Thu	Y
Fri	N
Sat	N

From hour: 0

To hour: 5

Figure 85: Flow regime editor.

4.2.2 Network regimes

The second tab in the Spatial Model regimes manager shows the different Network regimes (seen in Figure 86). The network regimes, also listed at the start of Unit C, are:

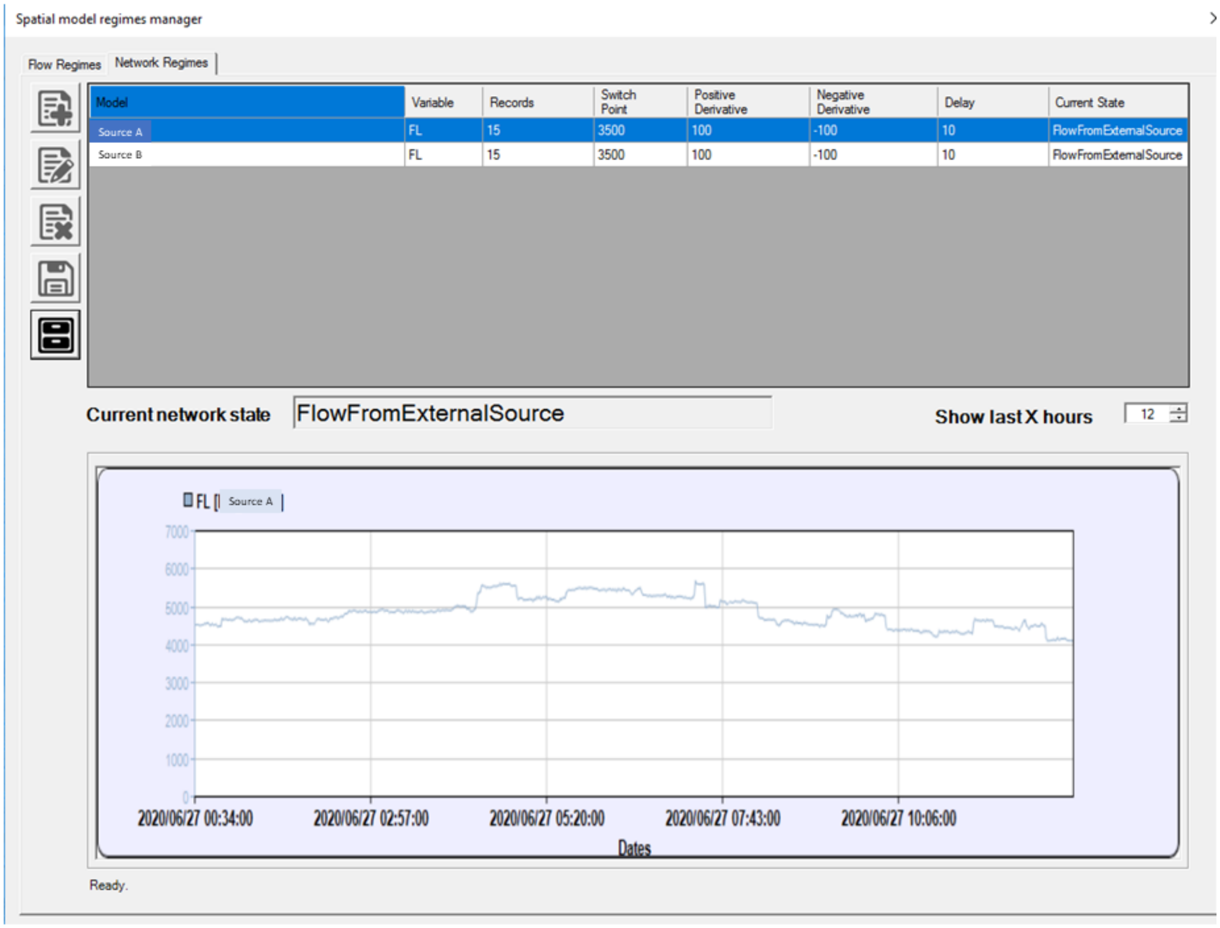
State E: Water supplied from external source

State R: Water supplied from reservoirs

State E→R: Network is transitioning from state E to R.

State R→E: Network is transitioning from state R to E.

Transitioning between network states: The algorithm by which the SM decides on the current network regime is as follows. The Spatial Model tracks flow variables defined by the user. For each flow variable, the SM evaluates a state machine, for its current state. The per-flow-variable states are seen in the last column of Figure 86, with the list of possible per-variable states being identical to the list of possible network states (regimes). When all monitored flow variables reach the same state, the network regime is changed to this state/regime. In other words, it is a voting mechanism requiring a unanimous vote. We move to describe the per-flow-variable state machine.



Spatial model regimes manager

Flow Regimes | Network Regimes

Model	Variable	Records	Switch Point	Positive Derivative	Negative Derivative	Delay	Current State
Source A	FL	15	3500	100	-100	10	FlowFromExternalSource
Source B	FL	15	3500	100	-100	10	FlowFromExternalSource

Current network state: FlowFromExternalSource

Show last X hours: 12

Ready.

Figure 86: Network regimes definitions tab

4.2.3 The per-flow-variable state machine

For each variable, the user defines:

- A switch-point (abbreviated to **threshold SP**)
- A positive and negative derivative thresholds (abbreviated to **threshold PD**, and **threshold ND**).
- The number of records used to compute the derivatives, or average the flow time raw values, abbreviated to **N**.
- The delay time, in minutes, giving the minimal time the state machine must spend in each state (to avoid jitter).

For each flow variable defined, the SM checks:

- Is the current flow value is above or below the current set point?
- Is the current derivate, given by the change in the flow value across the last N records (N being the number of records defined by the 'Records' field) above or below the PD and ND thresholds?

The transition rules for the state machine are described below. IF/THEN statements are marked with **bold/underline**, conditions on flow values are marked in red, and conditions on flow value derivatives are marked in blue.

Rule 1: If the network is in state 'R' and

all monitored flow values, averaged across N records are above their threshold SP and

all monitored flow value derivatives are smaller than their respective threshold ND values **then**:

transition to state 'R->E'.

Rule 2: If (the network is in state 'R' or state 'R->E') and

all monitored flow values, averaged across N records are below their threshold SP **then**:

transition to state 'E'.

Rule 3: If the network is in state 'E' and

all monitored flow values, averaged across N records are smaller their threshold SP and

all monitored flow value derivatives are higher than their respective threshold PD values **then**:

transition to state 'E->R'.

Rule 4: If (the network is in state 'E' or state 'E->R') and

all monitored flow values, averaged across N records are above their threshold SP **then**:

transition to state 'R'.

A diagram for the state machine, along with the possible transitions between states is given in Figure 87.

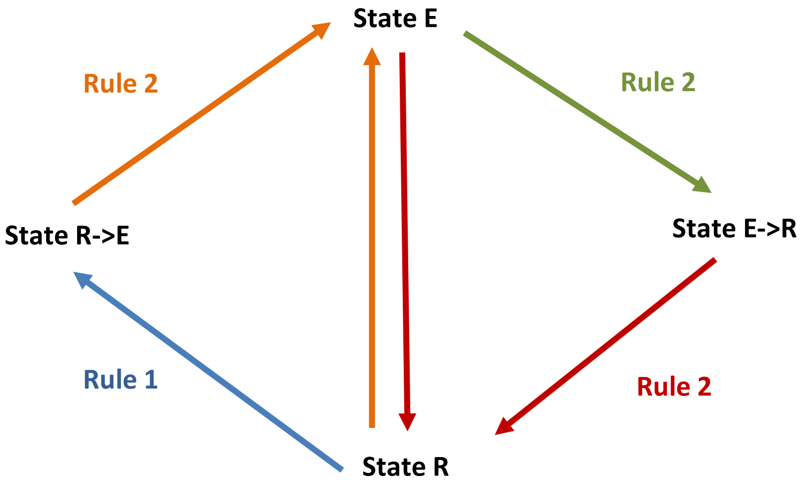


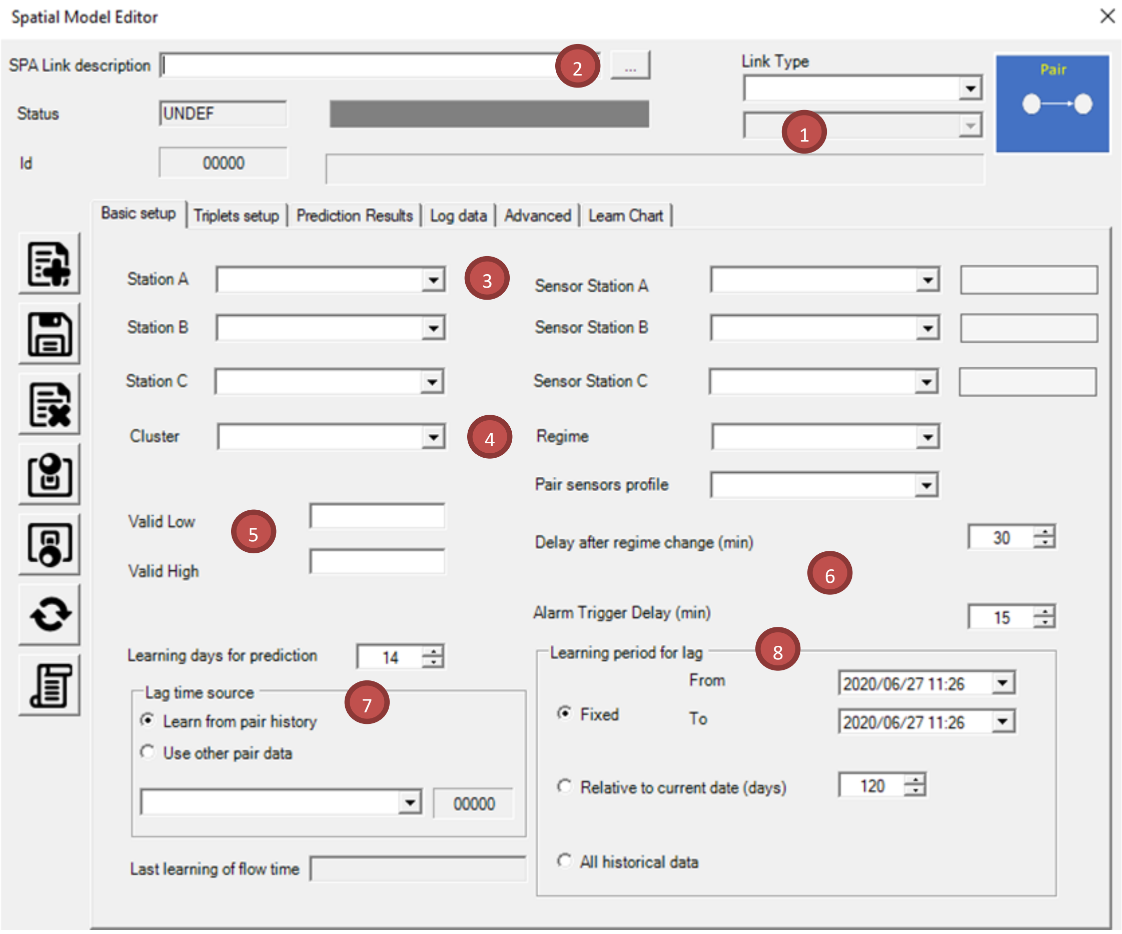
Figure 87: Network regimes state machine and transition rules

4.3 Pair Model – Definition

This section describes the spatial model editor, allowing the user to edit existing or new models. Sections 4.3-4.5 of this manual describe how to use the editor to edit pair models. Section 4.9, discussing triplet models describes how to create and edit triplet models. The model editor can be accessed by:

- clicking or a (pair) model in the spatial model main window, described in Section 4.1, and then clicking the model editor button ()
- clicking the 'create a new model' button () on the spatial model main window.

The spatial model window is shown in Figure 88. When creating a new model, the fields shown in the figure will be blank. When editing an existing model, the fields will be filled with model's details. The different field types (and field groups) in Figure 88 are numbered 1-8, in red. In the next page, we describe the buttons on the left side of the editor, along with all fields found under the basic setup tab. Additional tabs, explaining how to review estimated flow times, predictions performed and edit advanced model parameters are described in the next sections.



Status: UNDEF

Id: 00000

Delay after regime change (min): 30

Alarm Trigger Delay (min): 15

Learning days for prediction: 14

Lag time source: ● Learn from pair history ○ Use other pair data 00000

Learning period for lag: ● Fixed From 2020/06/27 11:26 To 2020/06/27 11:26 ○ Relative to current date (days) 120 ○ All historical data

Figure 88: The Spatial Model editor. Numbers corresponds to steps when building models, described in Section 4.3.

The buttons found on the left of the Spatial Model Editor and their corresponding actions are:

Save model

New Model

Delete model

Set model to 'Run state'

Set model to 'stop' state

Refresh fields

Request model flow time estimation

The **'New Model'** button will clear all fields, and allow a new model to be defined. The **'Delete Model'** button will delete the current model. The **'Save Model'** button will save the current model to the EDS/SM database, after changes have been made to model definitions. The **'lever-up'** and **'lever-down'** buttons will move the model between 'Run' and 'Stop' states, respectively. The **'Refresh'** button will reload all existing fields, as they are defined in the EDS/SM database. The **'Request model flow time estimation'** button will inform the SM that this model requires estimation of flow times, even if the last estimation of flow times for this models was made less than 24 hours ago.

Next, we describe the different fields required for the basic setup of a pair model. The steps, numbered 1-8 below, correspond to the numbering found in Figure 88.

Step 1 – select model type

Select the model type to be 'Pair' in the 'Link Type' selection box on Figure 89.

The illustration on the right of the selection box should display a pair model, with a source and target station.

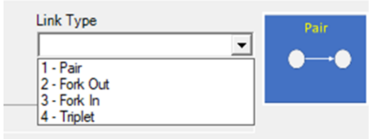
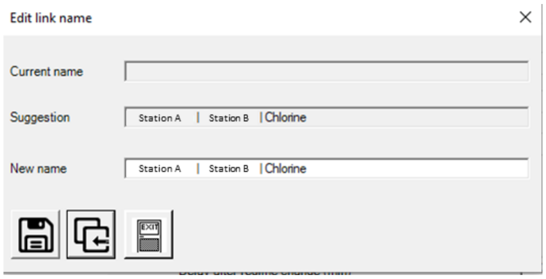


Figure 89: Model type selection box

Step 2 – select model name

Click the button with the label '…' to edit the model's name. The 'edit link name' window, shown in Figure 90 will appear. The window shows the current model name, a suggested model name and an editable field with the new name selected.



Edit link name

Current name:

Suggestion: Station A | Station B |Chlorine

New name: Station A | Station B |Chlorine

Figure 90: Model name selection box

Click this button to exit the window without selecting a new name

Click this button to copy the name suggested by the SM to the new name field

Click this button to save the currently selected name, fund under the 'new name' field

Step 3 – select model stations

Select the model stations, denoted by A and B for the source station and target station, respectively. For station, select the station name, along with the sensor name (4 selection boxes total). The measurement for the selected sensor will be displayed to the right of the sensor name for each sensor.

Station C is used in Triplet models alone.

Step 4 – select clusters, regime and sensors profile.

Step 4a: select cluster – each pair is part of a cluster. Clusters are groups of models, usually defined by geographical region. Clusters are used to filter models in the SM main window, so that only models of a selected cluster are viewed, see details in section 4.1. The user may type a new cluster name under the 'cluster' field, or select one of the existing cluster names

Step 4b: select network regimes -

Use the 'Regime' drop down menu (Figure 91) to select on which network regimes the model will be activated – all hours/ water supplied from storage/ water supplied from an external source.

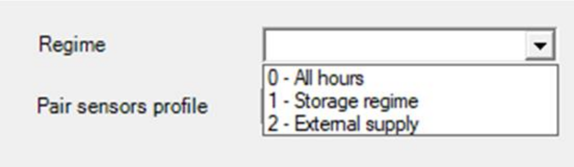
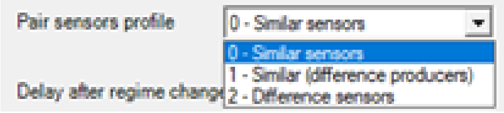


Figure 91: Selection box for the network regimes in which the model will be enabled.

Step 4c: select sensors profile -

Each pair has a sensors profile selected, as shown on Figure 92. The available profiles are:

- Similar sensors – sensors are for the same measurement and from the same manufacturer.
- Similar sensors, different producers – sensors are for the same measurement but from different manufacturers.
- Different sensors producers – Sensors are for different measurements



Pair sensors profile: 0 - Similar sensors

0 - Similar sensors
1 - Similar (difference producers)
2 - Difference sensors

Delay after regime change

Figure 92: Select pair sensors profile

The sensor profiles are used to categorize pair models when producing reports in the SM software.

Step 5 – select valid limits

Select the upper and lower limits for both the source station and target station, in terms of measurements. Measurements above or below these values will be regarded as technical faults and will be removed from statistical analysis.

Step 6- Select delay times

The basic model definition has two delay times defined:

Delay after regime change- following this time (in minutes) after a network regime change, the model will not generate alarms. Note that this condition is additional to the active regime selection in step 4b (pair active in network states E/R/E&R).

Alarm trigger delay- this delay time is the amount of time the target measurement must exceed the spatial model limits (shown in Sections 4.1 and 4.6 for each model) for an alarm to be trigged. Before this time, the alarm will be considered as pending.

Step 7- select the lag time source, and number of data days used for prediction

Learning days for prediction – this field sets the number of days, prior to the current EDS monitoring date, that will be used to construct the prediction model. Recommended values are 14-30 days, with 30 days required for general sensors, and 14 days required for high-quality physical measurement sensors.

Lag time source – sets from where flow time estimates are taken:

- Learn from pair history – this pair model estimates its own flow time.
- Use other pair – this pair model will copy its flow time estimates from another pair model, between the same source and target stations. When selecting this option, the user must also select from which pair model should the flow time estimates be copied.

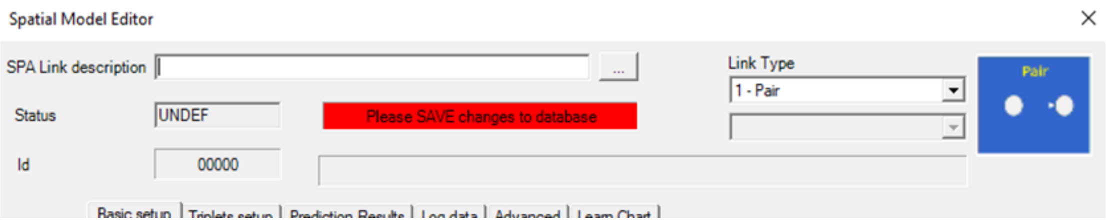
The field **'Last learning of flow time'** will show the last time, by server time, that flow time estimates were computed for this pair model.

<u>Step 8 - select the learning period for lag time</u>

The user must select the learning period for the lag time, is the pair model will estimate its own flow time (first option in Lag time source, step 7). The options are:

- Fixed – the user may select fixed dates from which data will be taken for flow time estimation
- Relative to current date – the user may select the number of days, prior to the current date, used for estimating flow times.
- All historical data – all available data will be used to obtain flow time estimates.

Following steps 1-8, the user must save the selected configuration. The Spatial Model editor will request the user to save the configuration, as in Figure 93.



Spatial Model Editor

SPA Link description:

Status: UNDEF

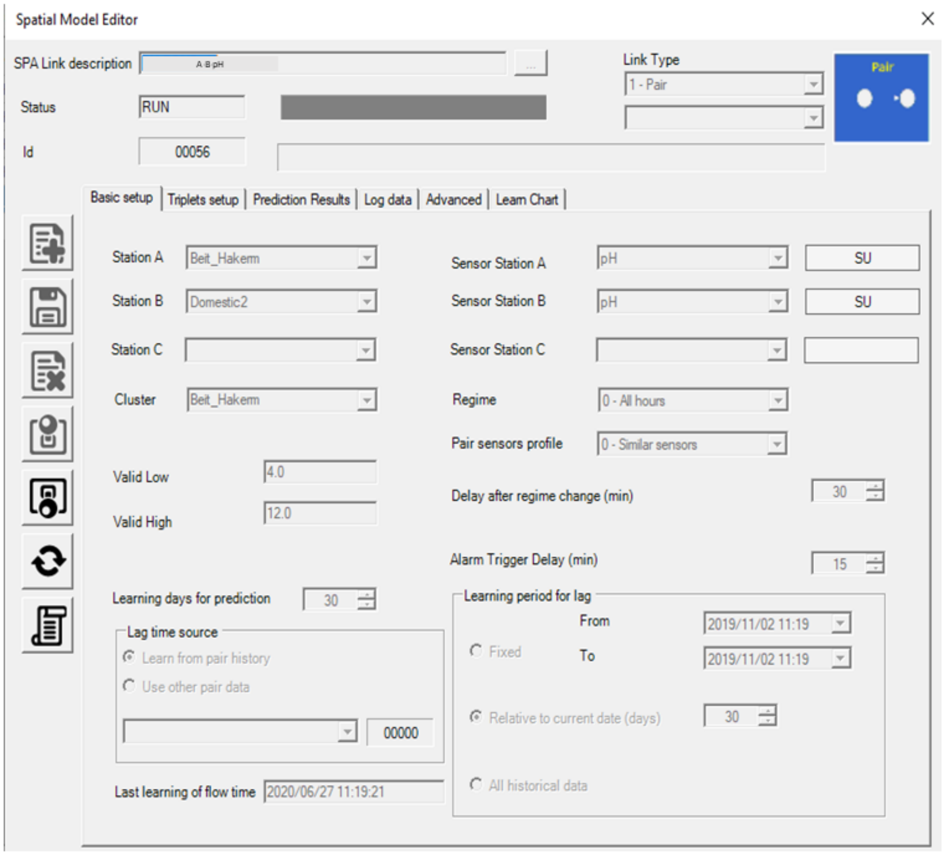
Please SAVE changes to database

Id: 00000

Link Type: 1 - Pair

Figure 93: message requesting the user to save the current model configuration.

Following the configuration, set the model to run using the 'lever-up button'. On screen fields will be grayed out, as in Figure 94, except for the 'Stop', 'Refresh' and 'Learn' buttons. Stopping the model will make all fields editable again.



Spatial Model Editor

SPA Link description: A B pH

Status: RUN

Id: 00056

Link Type: 1 - Pair

Station A: Beit_Hakerm

Station B: Domestic2

Station C:

Cluster: Beit_Hakerm

Sensor Station A: pH SU

Sensor Station B: pH SU

Sensor Station C:

Regime: 0 - All hours

Pair sensors profile: 0 - Similar sensors

Valid Low: 4.0

Valid High: 12.0

Delay after regime change (min): 30

Alarm Trigger Delay (min): 15

Learning days for prediction: 30

Lag time source

● Learn from pair history

○ Use other pair data

00000

Last learning of flow time: 2020/06/27 11:19:21

Learning period for lag

○ Fixed From: 2019/11/02 11:19 To: 2019/11/02 11:19

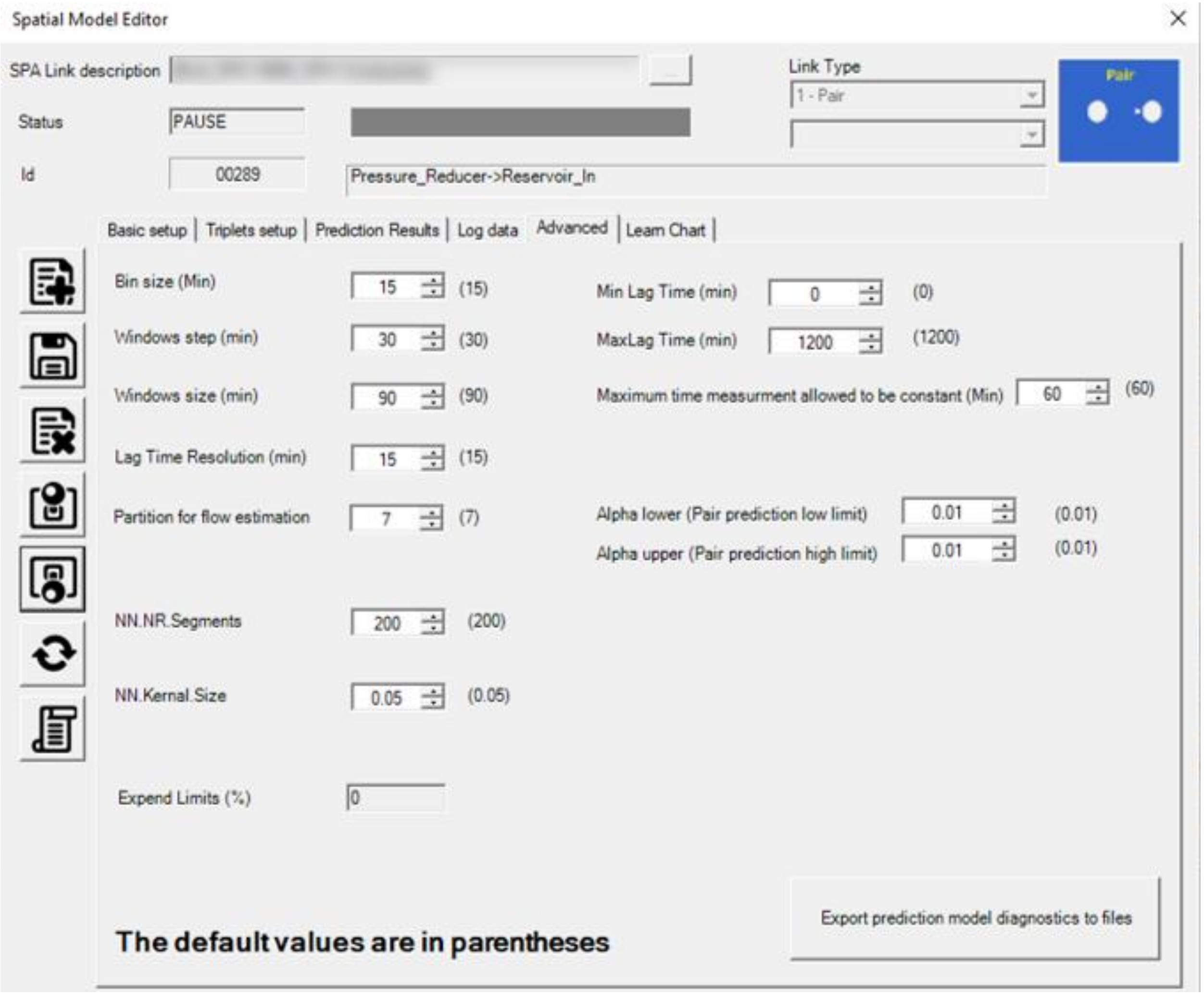
● Relative to current date (days): 30

○ All historical data

Figure 94: The model editor for a running model

4.4 Pair Model – Advanced Parameters

This section describes the advanced parameters tab, containing additional model configuration parameters. The tab is shown in Figure 95.



SPA Link description [redacted]

Status PAUSE

Id 00289 Pressure_Reducer->Reservoir_In

Link Type 1 - Pair

Bin size (Min) 15 (15)

Windows step (min) 30 (30)

Windows size (min) 90 (90)

Lag Time Resolution (min) 15 (15)

Partition for flow estimation 7 (7)

NN.NR.Segments 200 (200)

NN.Kernal.Size 0.05 (0.05)

Expend Limits (%) 0

Min Lag Time (min) 0 (0)

MaxLag Time (min) 1200 (1200)

Maximum time measurment allowed to be constant (Min) 60 (60)

Alpha lower (Pair prediction low limit) 0.01 (0.01)

Alpha upper (Pair prediction high limit) 0.01 (0.01)

The default values are in parentheses

Figure 95: The advanced parameters tab

The following parameters are found in the tab:

Bin size – measurements are binned across time, with the median value in each bin taken as the representative value, as discussed in Section 2.1. This value sets the size of the temporal bin, in minutes.

Window step – When computing the sum of the correlation statistic, across all window pairs, this will give the step size, in minutes, when incrementing the pointer across the two times series. Since each of the two time series is highly correlated across time, the correlation statistics for pairs of windows changes very little, when incrementing window positions by a small value in time. Hence, it makes sense to increment windows by a more than a single measurement, e.g. , increment window positions by 30 minutes, each increment, when iterating over all pairs of windows. See Section 3.4 for a detailed description.

Window size – The size of the temporal window , in terms of minutes, for computation of the correlation between the source and target measurement series. See Section 3.4 for a detailed description.

Lag time resolution – when estimating the score function across lags, this will give the resolution of the candidate lag times, in terms of minutes, e.g., estimate the possible lags at 15 minute steps: a flow time of 0 minutes, 15 minutes, 30 minutes, 45 minutes and so on.

Partition for flow time estimation – The number of partitioning units, when using a leave-one-out estimation method to assess the sampling error of flow time estimate. The full algorithm is described in Section 3.4, Unit B.

NN Nr segments – The total number of segments used for the kernel estimator regression model, as explained in Sections 2.5 and 3.5. Note that this is the number of segments used for the entire dataset (X-axis), not just for computing the local prediction estimate.

NN Kernal Size - The relative portion of all segments used for computing the kernel model's predictions, as explained in Sections 2.5 and 3.5.

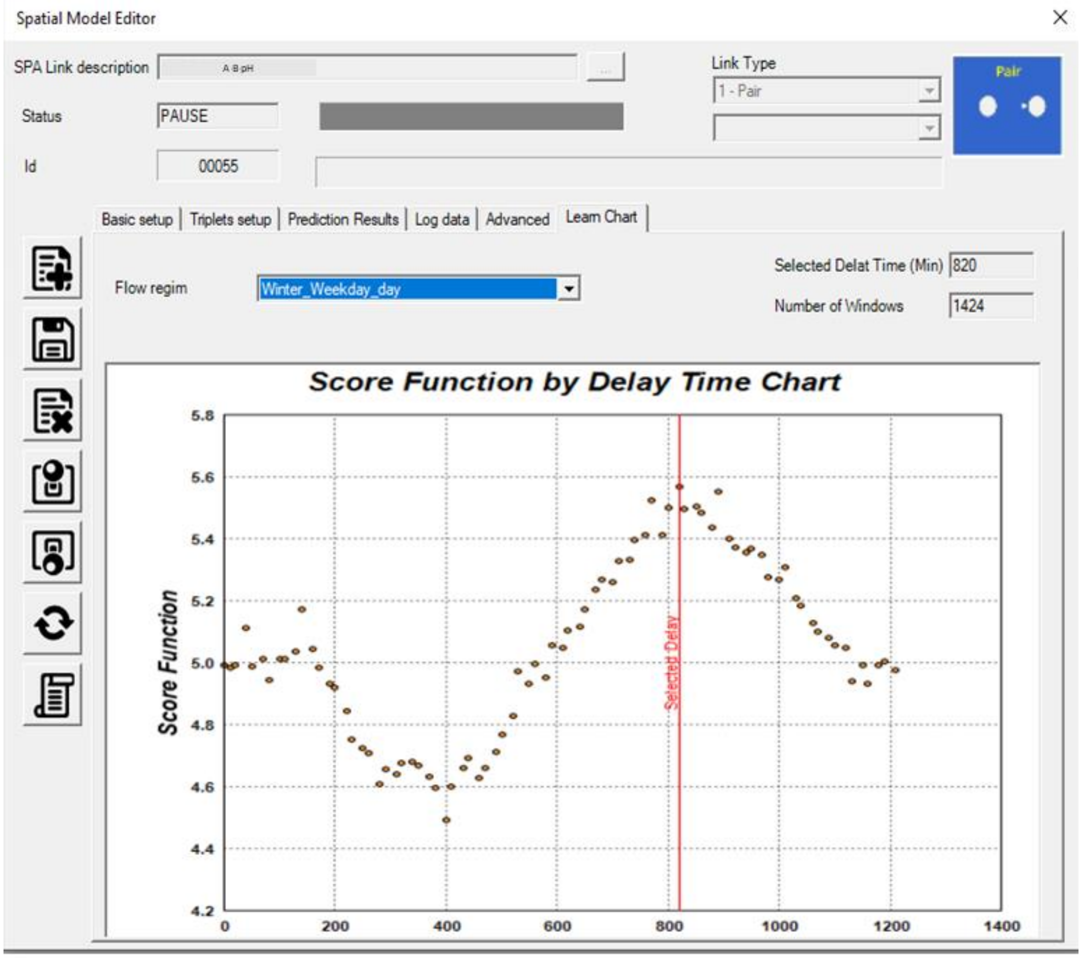
Minimum and Maximum lag time – The minimum and maximum times for flow time estimate candidates, in minutes.

Maximum time measurements is allowed to be constant – If a measurement is constant for more than this time, additional constant measurements will be considered invalid.

Alpha lower limit and Alpha upper limit – The upper and lower non-coverage probabilities for the prediction interval constructed by the prediction model. For example, by setting these values to be both 0.01, the upper and lower limits computed by the pair's prediction model will provide a 98% prediction interval. The 98% prediction interval will has a 98% probability of having the actual measurement inside it, for each measurement.

4.5 Pair model – Flow time estimation and diagnostics for flow time estimation

In Unit B, Section 3.2, we reviewed the algorithms and methods used for flow time estimation in Pair models. **The 'Learn Chart' tab is used to view the score function**, by flow regime, after flow times have been estimated for each flow regime. Figure 96 shows the 'Learn Chart' for an example model. The score function is plotted for each candidate lag value. The selected lag\ selected delay time is marked by a red line on the graph, and is written under the label "Selected Delay time". The number of window pairs used for flow time estimation is also presented on screen. The Flow regime to be presented may be can be selected using the drop-down menu on the left.



Spatial Model Editor

SPA Link description: A B pH

Status: PAUSE

Id: 00055

Link Type: 1 - Pair

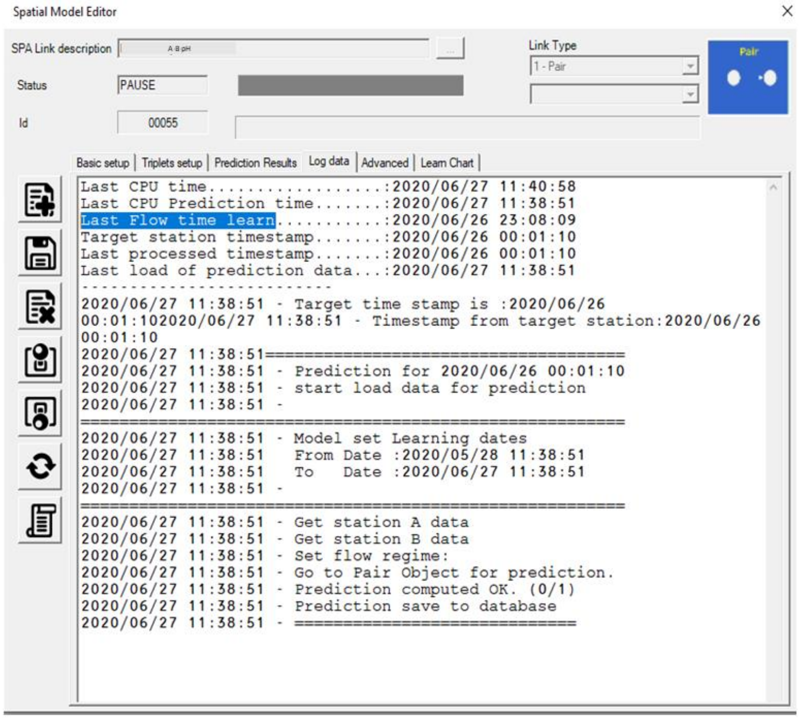
Flow regim: Winter_Weekday_day

Selected Delat Time (Min): 820

Number of Windows: 1424

Figure 96: The 'Learn Chart' tab of the model editor, showing the score (objective) function for each flow regime.

The 'Log Data' tab shown in Figure 97 lists the messages and error descriptions generated by the pair model. Figure HHH shows a possible output, where the model estimated flow times and preformed predictions correctly. The log messages detail the Last time CPU times were given to the model, the last target station timestamp processed, the dates used for flow time estimation, and other descriptive fields.



SPA Link description: A B pH

Status: PAUSE

Id: 00055

Link Type: 1 - Pair

```
Last CPU time...................:2020/06/27 11:40:58
Last CPU Prediction time........:2020/06/27 11:38:51
Last Flow time learn............:2020/06/26 23:08:09
Target station timestamp........:2020/06/26 00:01:10
Last processed timestamp........:2020/06/26 00:01:10
Last load of prediction data....:2020/06/27 11:38:51
--------------------------
2020/06/27 11:38:51 - Target time stamp is :2020/06/26
00:01:102020/06/27 11:38:51 - Timestamp from target station:2020/06/26
00:01:10
2020/06/27 11:38:51=====================================
2020/06/27 11:38:51 - Prediction for 2020/06/26 00:01:10
2020/06/27 11:38:51 - start load data for prediction
2020/06/27 11:38:51 -
========================================================
2020/06/27 11:38:51 - Model set Learning dates
2020/06/27 11:38:51   From Date :2020/05/28 11:38:51
2020/06/27 11:38:51   To   Date :2020/06/27 11:38:51
2020/06/27 11:38:51 -
========================================================
2020/06/27 11:38:51 - Get station A data
2020/06/27 11:38:51 - Get station B data
2020/06/27 11:38:51 - Set flow regime:
2020/06/27 11:38:51 - Go to Pair Object for prediction.
2020/06/27 11:38:51 - Prediction computed OK. (0/1)
2020/06/27 11:38:51 - Prediction save to database
2020/06/27 11:38:51 - =============================
```

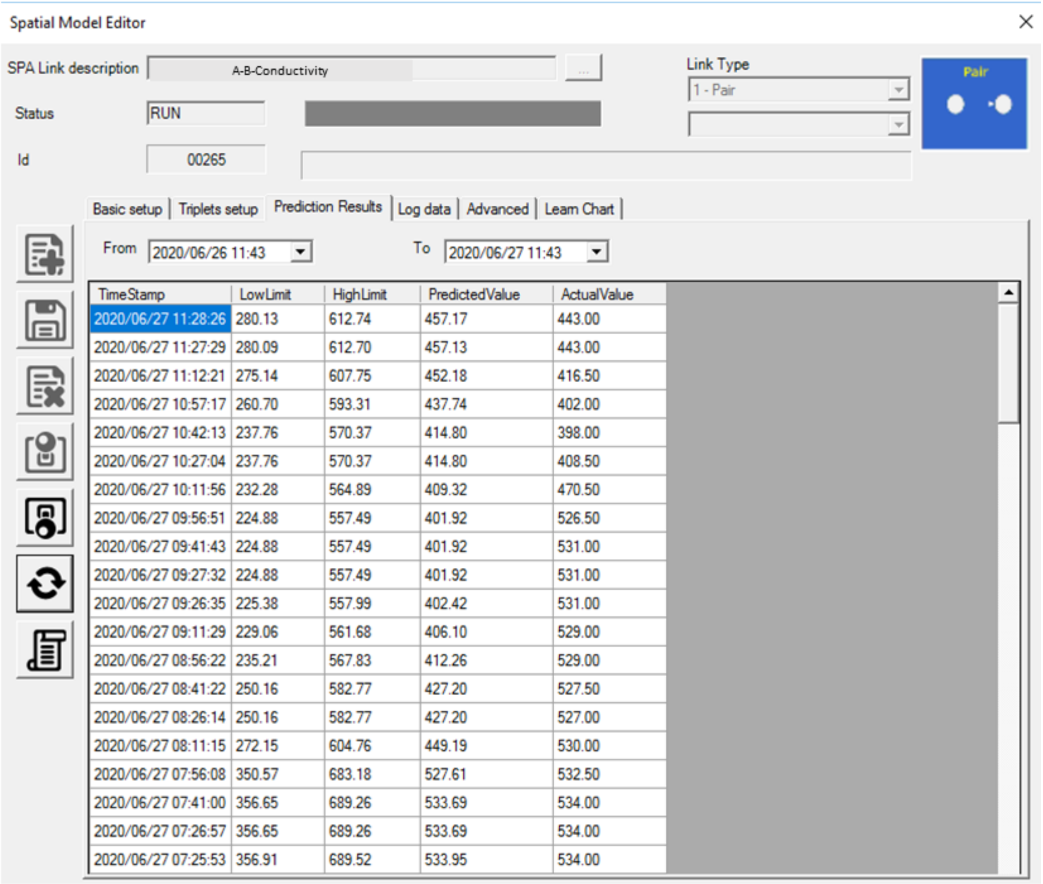
Figure 97: The 'Log Data' tab shows the error and message log for the pair model.

4.6 Pair model – Predictions, charts and diagnostics for prediction models

The 'Prediction Results' tab in the Spatial Model editor describes for each pair model:

- Model prediction for the target node by timestamp
- Actual values for the target node, by timestamp
- Spatial Model limits (both lower and upper limit) used for detecting abnormal events in the target node.

Figure 98 below shows the Prediction Results tab for a conductivity-based pair model.



SPA Link description: A-B-Conductivity

Status: RUN

Id: 00265

Link Type: 1 - Pair

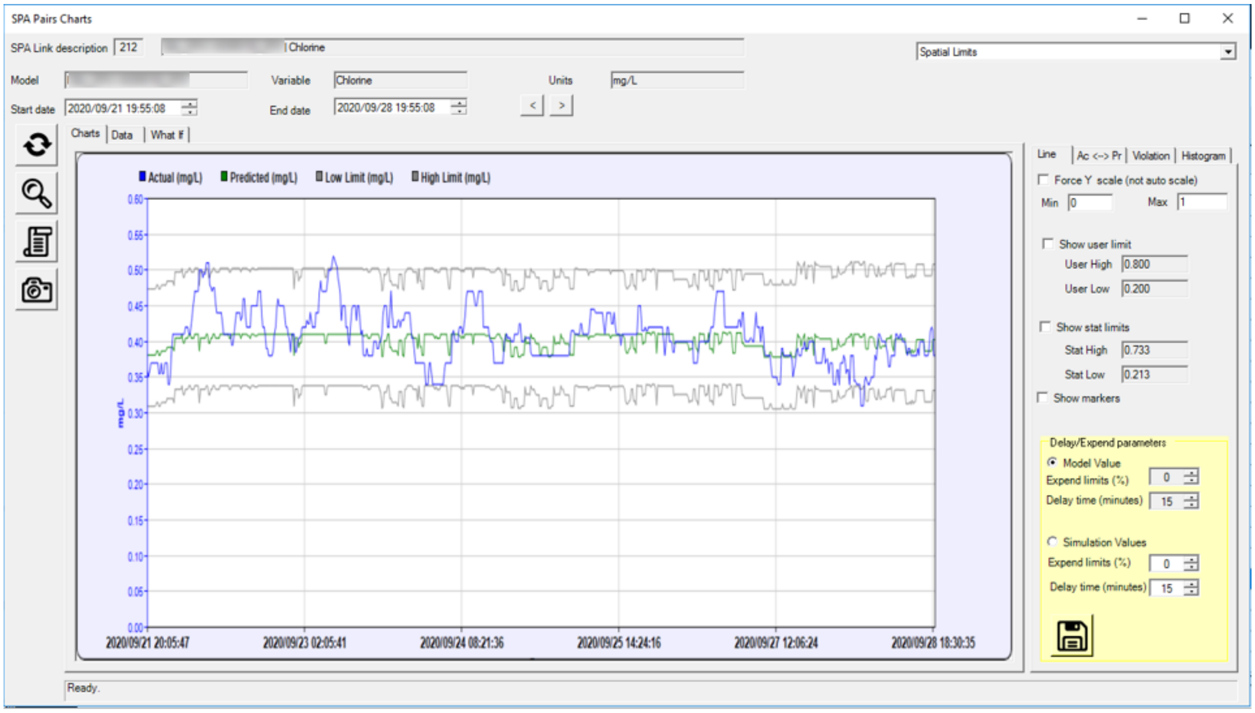
From 2020/06/26 11:43 To 2020/06/27 11:43

TimeStamp	LowLimit	HighLimit	PredictedValue	ActualValue
2020/06/27 11:28:26	280.13	612.74	457.17	443.00
2020/06/27 11:27:29	280.09	612.70	457.13	443.00
2020/06/27 11:12:21	275.14	607.75	452.18	416.50
2020/06/27 10:57:17	260.70	593.31	437.74	402.00
2020/06/27 10:42:13	237.76	570.37	414.80	398.00
2020/06/27 10:27:04	237.76	570.37	414.80	408.50
2020/06/27 10:11:56	232.28	564.89	409.32	470.50
2020/06/27 09:56:51	224.88	557.49	401.92	526.50
2020/06/27 09:41:43	224.88	557.49	401.92	531.00
2020/06/27 09:27:32	224.88	557.49	401.92	531.00
2020/06/27 09:26:35	225.38	557.99	402.42	531.00
2020/06/27 09:11:29	229.06	561.68	406.10	529.00
2020/06/27 08:56:22	235.21	567.83	412.26	529.00
2020/06/27 08:41:22	250.16	582.77	427.20	527.50
2020/06/27 08:26:14	250.16	582.77	427.20	527.00
2020/06/27 08:11:15	272.15	604.76	449.19	530.00
2020/06/27 07:56:08	350.57	683.18	527.61	532.50
2020/06/27 07:41:00	356.65	689.26	533.69	534.00
2020/06/27 07:26:57	356.65	689.26	533.69	534.00
2020/06/27 07:25:53	356.91	689.52	533.95	534.00

Figure 98: The 'Prediction Results' tab shows the predictions performed for the target node, along with the upper and lower limits and the actual values measured in the target node.

Next, we review the different charts available for each pair model. The charts are accessible through the Spatial Model main window, by selecting a pair model in the SM pairs tab, and clicking on the 'graphs' button in the main menu toolbar (📈).

After opening the charts screen for a pair model, the Spatial Limits chart for the model will appear, as in Figure 99. The Spatial Limits charts presented the actual, predicted and spatial limit values for the pair model's target node, in a time series chart format. A legend explaining the color coding is featured above the graph. The user may select different dates to display values for, and click the refresh button (⟳) to view an updated graph.



SPA Pairs Charts

SPA Link description: 212 [redacted] Chlorine

Model: [redacted]

Variable: Chlorine

Units: mg/L

Start date: 2020/09/21 19:55:08

End date: 2020/09/28 19:55:08

Spatial Limits

☐ Force Y scale (not auto scale)

Min: 0 Max: 1

☐ Show user limit

User High: 0.800

User Low: 0.200

☐ Show stat limits

Stat High: 0.733

Stat Low: 0.213

☐ Show markers

Delay/Expend parameters

● Model Value

Expend limits (%): 0

Delay time (minutes): 15

○ Simulation Values

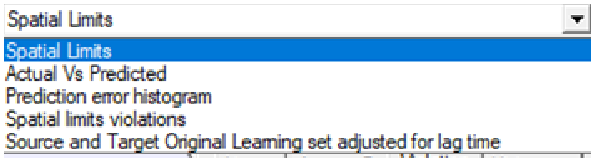
Expend limits (%): 0

Delay time (minutes): 15

Ready.

Figure 99: The SPA pairs chart window show the actual, predicted and limit values for the target node in a pair model, as a time series.

The user may select additional graphs, using the drop down menu, on the top right of the screen. The different charts options available are shown in Figure 100.



Spatial Limits

- Spatial Limits
- Actual Vs Predicted
- Prediction error histogram
- Spatial limits violations
- Source and Target Original Learning set adjusted for lag time

Figure 100: Type of charts available in the 'SPA Pairs Charts' window.

We proceed to reviewing the different types of charts available.

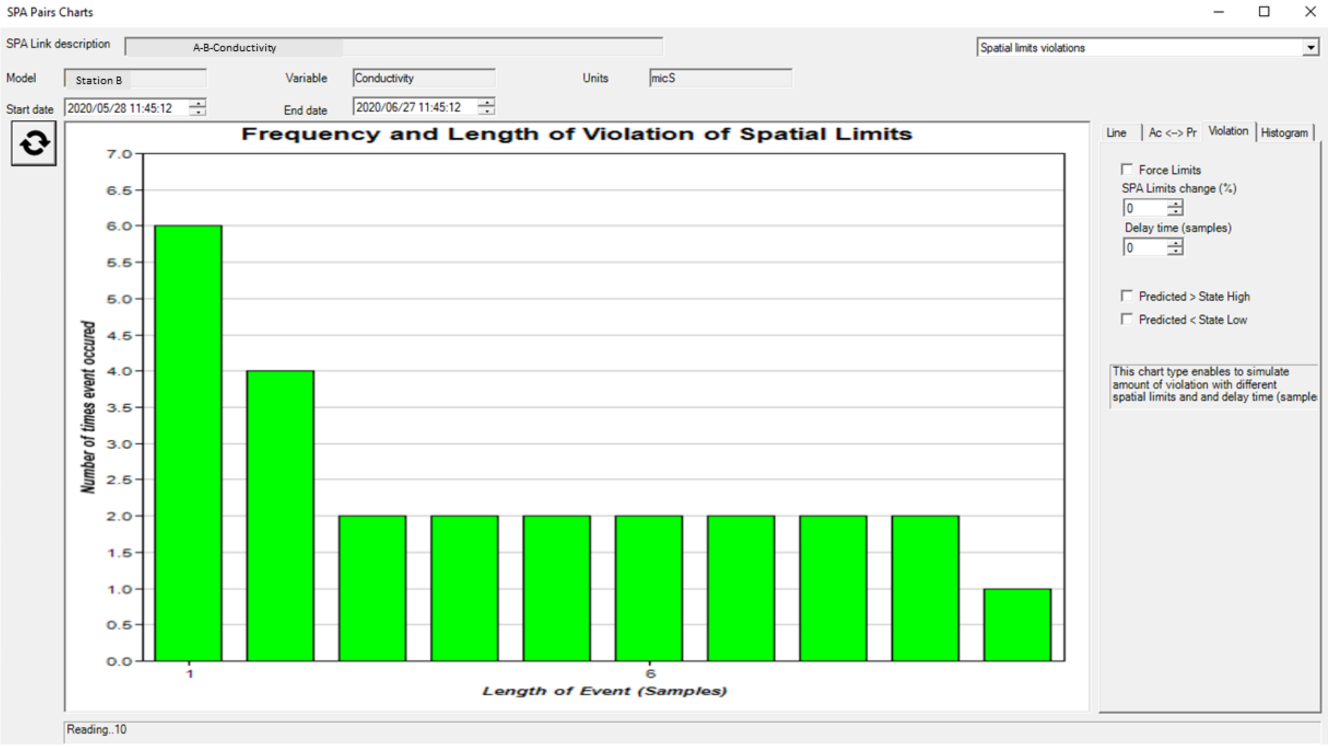
The 'Spatial Limits Violations' chart is a bar plot stating the number of measurements that featured a violation of the spatial model limits, and the number of consecutive violations of the spatial model limits up to that measurement. For example, the graph depicted in Figure 101 shows:

- 6 measurements that featured a violation of the spatial model limits, with no previous violation of the limits in the time prior to the measurement
- 4 measurements that featured a second violation (in-a-row) of the Spatial Model limits.
- 2 measurements that featured a third violation (in-a-row) of the Spatial Model limits.
- 2 measurements that featured a fourth violation (in-a-row) of the Spatial Model limits.
- …
- 2 measurements that featured a ninth violation (in-a-row) of the Spatial Model limits.
- 1 measurement that featured a tenth violation (in-a-row) of the Spatial Model limits.

This graph can be interpreted as a:

- **Two** events of Spatial Model limits violations lasting for a **single** measurements, together with,
- **Two** events of Spatial Model limits violations lasting for **two** consecutive measurements, together with,
- A **single** Spatial Model limits violation event lasting for **9** consecutive measurements, and
- A **single** Spatial Model limits violation event lasting for **10** consecutive measurements

(6 events total).


SPA Pairs Charts

SPA Link description: A-B-Conductivity

Spatial limits violations

Model: Station B
Variable: Conductivity
Units: micS

Start date: 2020/05/28 11:45:12
End date: 2020/06/27 11:45:12

☐ Force Limits
SPA Limits change (%): 0
Delay time (samples): 0

☐ Predicted > State High
☐ Predicted < State Low

This chart type enables to simulate amount of violation with different spatial limits and and delay time (sample

Reading..10

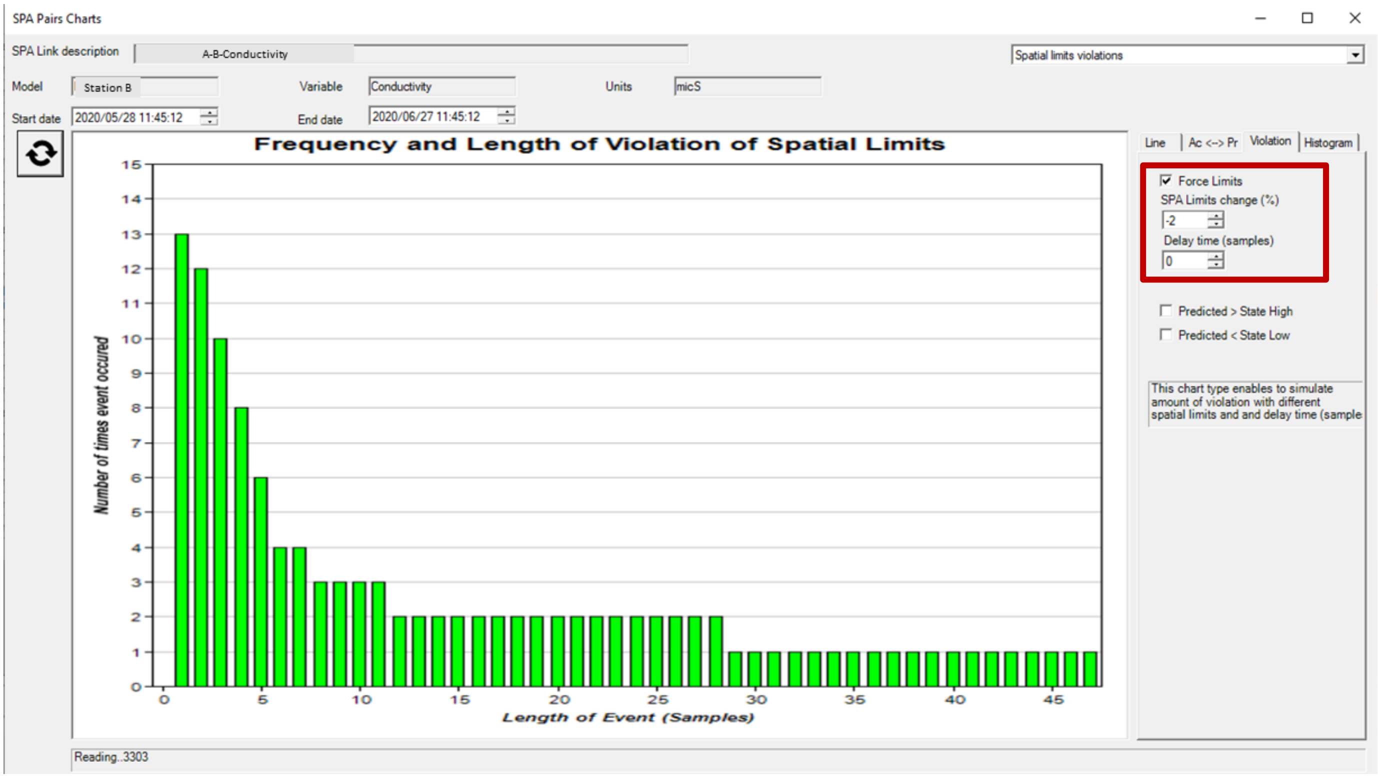
Frequency and Length of Violation of Spatial Limits
Number of times event occured
Length of Event (Samples)
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
6
Line
Ac <--> Pr
Violation
Histogram

Figure 101: Bar chart showing the number of timestamps with X consecutive violations of Spatial limits, X being a number of samples. For example, in the above graph, 2 events violated the spatial limits for 9 consecutive measurements, and 1 of the 2 events violated the spatial limits for an additional, 10th consecutive measurement.

The 'Spatial Limits violations' chart allows the user to run a 'what-if' analysis, testing how would the number of events change if:

- Spatial model limits were increased or decreased by a constant percent
- The delay time for raising an alarm was different, in terms of the number of observations.

In order to run a 'what-if' analysis, click the 'force-limits' option as shown in Figure 79 and click the refresh button (⟳).



SPA Pairs Charts

SPA Link description: A-B-Conductivity

Spatial limits violations

Model: Station B

Variable: Conductivity

Units: micS

Start date: 2020/05/28 11:45:12

End date: 2020/06/27 11:45:12

☑ Force Limits

SPA Limits change (%): -2

Delay time (samples): 0

☐ Predicted > State High

☐ Predicted < State Low

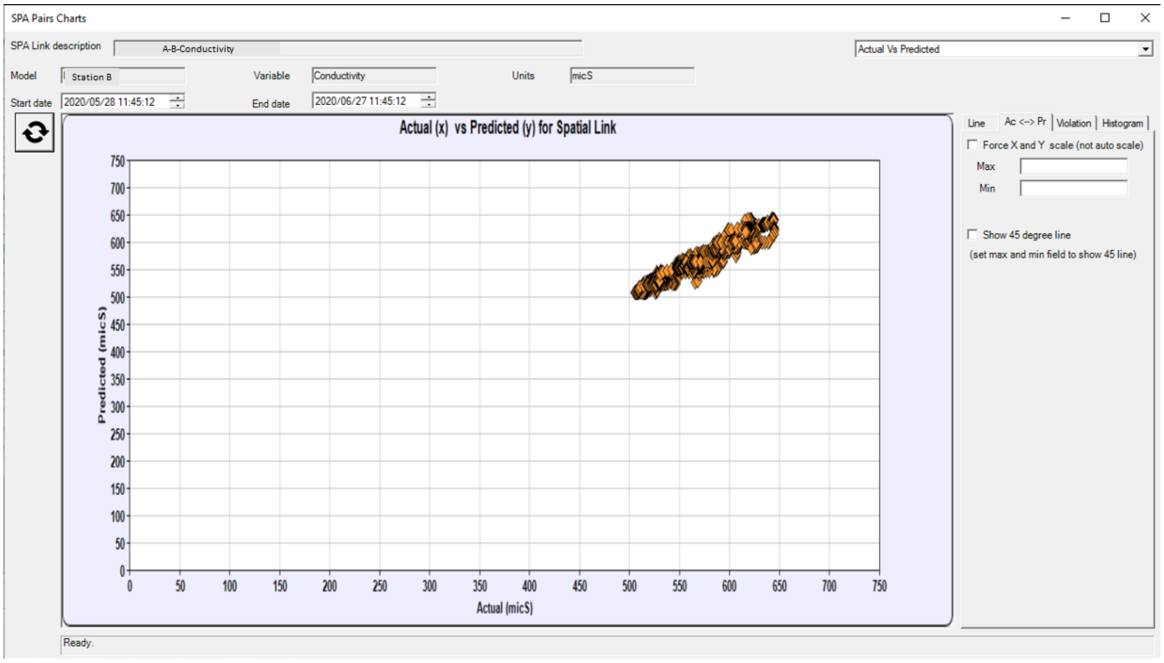
This chart type enables to simulate amount of violation with different spatial limits and and delay time (sample

Reading..3303

Frequency and Length of Violation of Spatial Limits
Number of times event occured
Length of Event (Samples)
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
5
10
15
20
25
30
35
40
45
Line
Ac <-> Pr
Violation
Histogram

Figure 102: The user may run a "what-if" analysis, increasing or decreasing the SPA limits by a fixed percentage, or changing the delay time. The "what-if" scenario can be run using the "force limits" option on the right.

The 'Actual vs. Predicted' chart (shown in Figure 103) allows the user to view the actual values measured in the target station along with their predicted values. Each point represents a single data point (time-stamp). The user may scale the axis, as shown in Figure 104.


SPA Pairs Charts

SPA Link description: A-B-Conductivity

Actual Vs Predicted

Model: Station B

Variable: Conductivity

Units: micS

Start date: 2020/05/28 11:45:12

End date: 2020/06/27 11:45:12

□ Force X and Y scale (not auto scale)

Max:

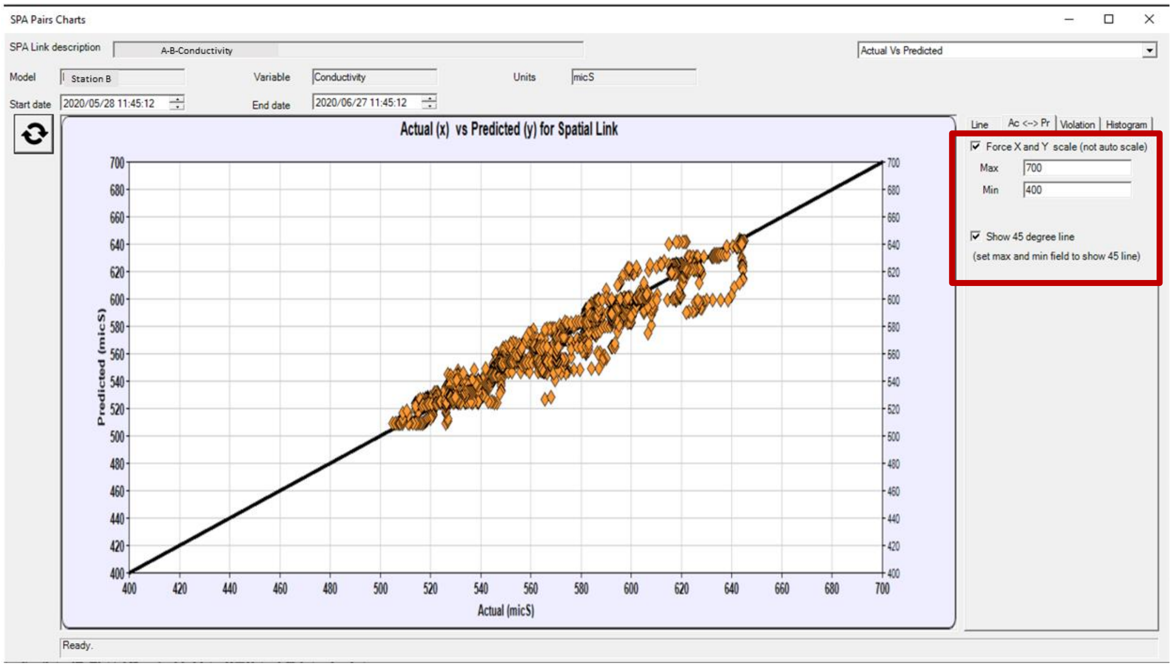
Min:

□ Show 45 degree line

(set max and min field to show 45 line)

Ready.

Figure 103: Actual Vs. Predicted graph for the target node in a pair


SPA Pairs Charts

SPA Link description: A-B-Conductivity

Actual Vs Predicted

Model: Station B

Variable: Conductivity

Units: micS

Start date: 2020/05/28 11:45:12

End date: 2020/06/27 11:45:12

☑ Force X and Y scale (not auto scale)

Max: 700

Min: 400

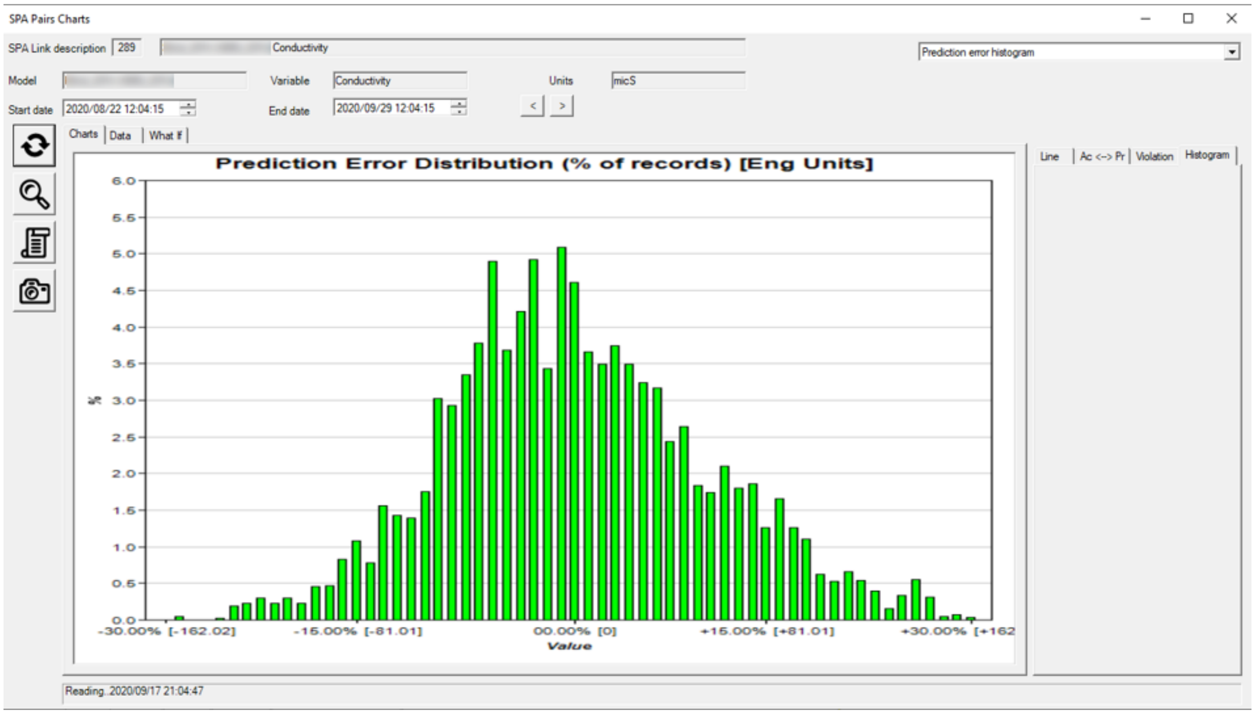
☑ Show 45 degree line

(set max and min field to show 45 line)

Ready.

Figure 104: The user may set the X and Y scales for the Actual vs. Predicted graph using the "Force X and Y scale" option on the right (marked in red).

The 'Prediction Error Histogram' chart, shown in Figure 105, shows the distribution of prediction errors, for the pair model using two scales: engineering units (in brackets), and relative part of the range defined by the user limits (next to the percentage sign).



SPA Pairs Charts

SPA Link description: 289 | [redacted] Conductivity

Model: [redacted]

Variable: Conductivity

Units: micS

Start date: 2020/08/22 12:04:15

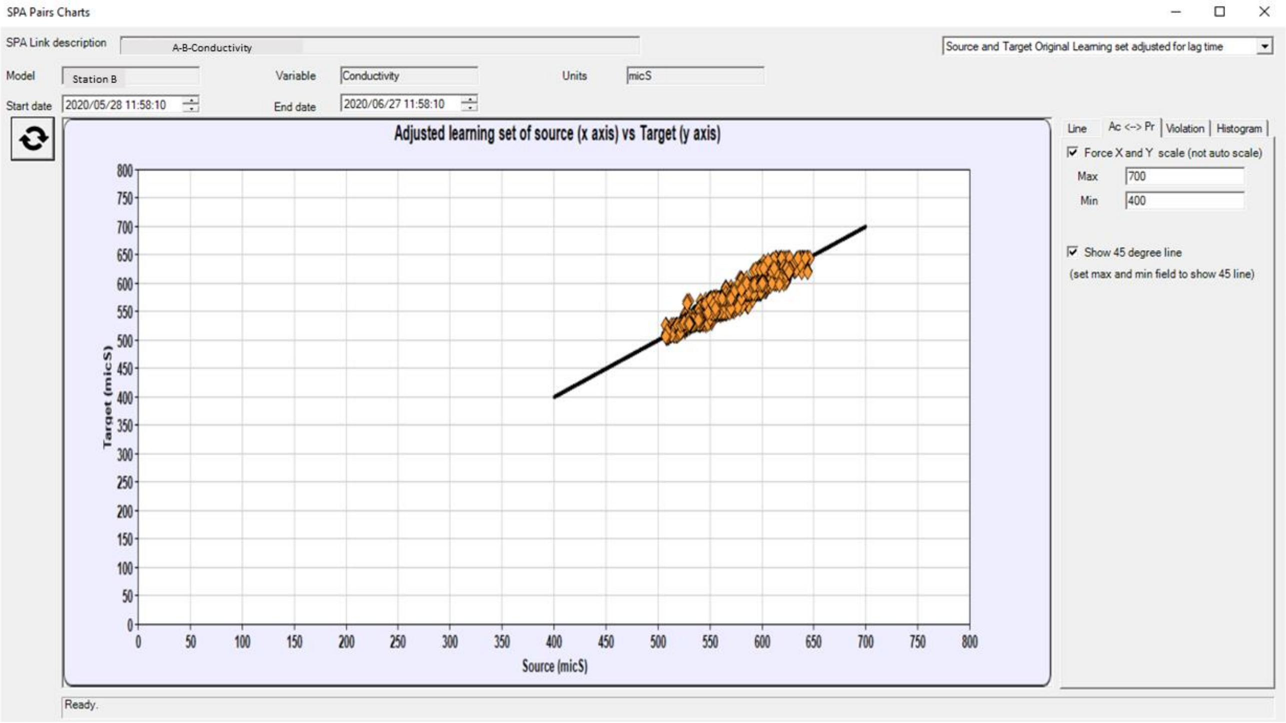
End date: 2020/09/29 12:04:15

Prediction error histogram

Reading..2020/09/17 21:04:47

Figure 105: A histogram for the distribution of prediction errors.

The 'Source and Target Original Learning Set (adjusted for flow-time)' chart presented the data points used to build the prediction model, as explain in Sections 2.5 and 3.5. The pairs of points represent target node measurements, along with their *flow time adjusted* source station values. A pair of points in the chart, (X,Y), represent values measured at the source and target stations. The two measurements are not taken at the same time, but at a time gap according to the estimated flow time between stations. Note that not all points in the graph use the same flow time (lag time)- the time difference used for associating X and Y measurements is taken based on the time stamp and flow regime associated with the target node/ Y value.



SPA Pairs Charts

SPA Link description: A-B-Conductivity

Source and Target Original Learning set adjusted for lag time

Model: Station B

Variable: Conductivity

Units: micS

Start date: 2020/05/28 11:58:10

End date: 2020/06/27 11:58:10

☑ Force X and Y scale (not auto scale)

Max: 700

Min: 400

☑ Show 45 degree line

(set max and min field to show 45 line)

Ready.

Figure 106: A scatterplot showing the actual target station values vs. the lag-time adjusted values from the source station.

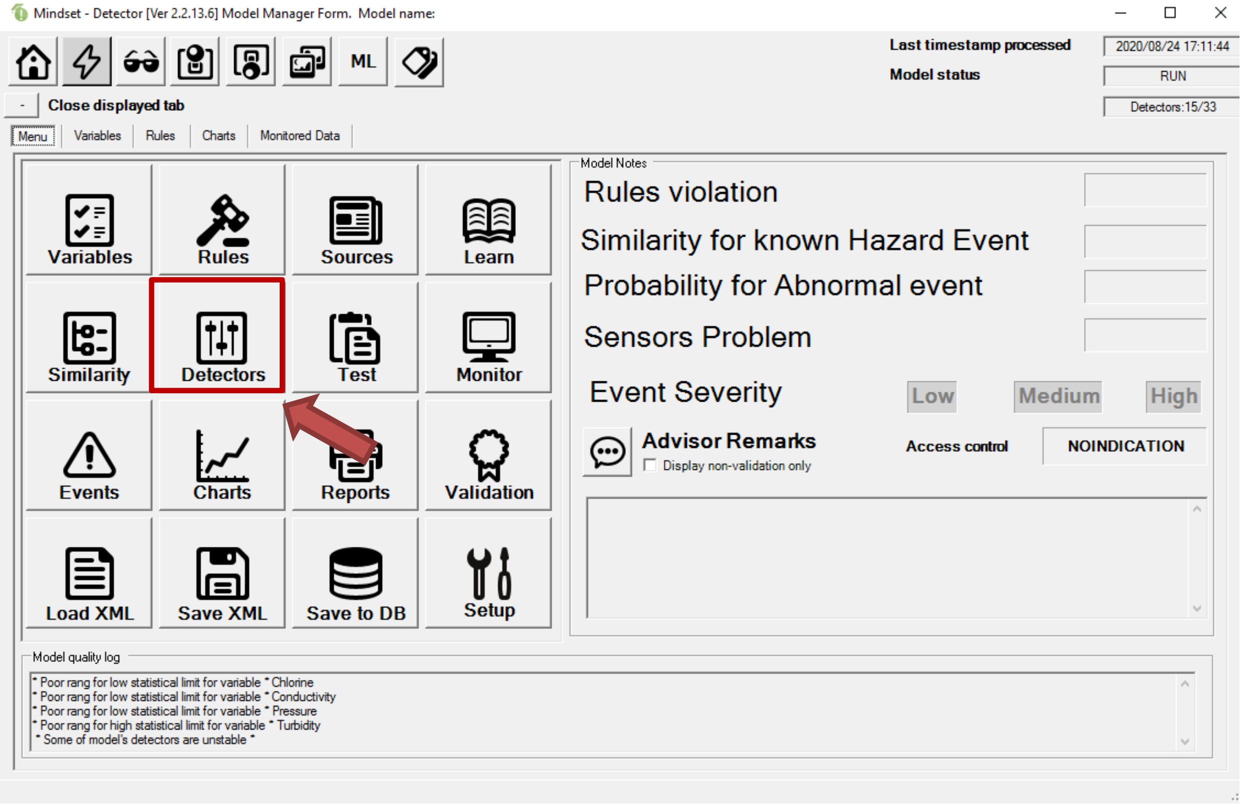
4.7 Pair model – Connecting to EDS, defining detectors and delay time

This section discusses how to connect the SM to the alarm logging and notification management in the EDS system. Users are encouraged to review the types of alarms, their meaning and trigger conditions, and their indicators, at the start of Unit C and at Section 4.1. Note that only SM limit violation alarms trigger an EDS alarm. All other alarm types (poor limit/ poor violation/ no data/ Triplet alarm) will be added as secondary alarms (descriptive), to the limits violation alarm.

To connect a alarms from SM model to an EDS model perform the following steps.

Step I: Open the EDS model manager form for the model to which you want to connect SM alarms. The model manager window can be opened through the EDS main window. Figure 107 shows such a form after opening.

Step II: Select the "Detectors" window, by clicking the the "Detectors" button marked by a red rectangle in Figure 107.



Mindset - Detector [Ver 2.2.13.6] Model Manager Form. Model name:

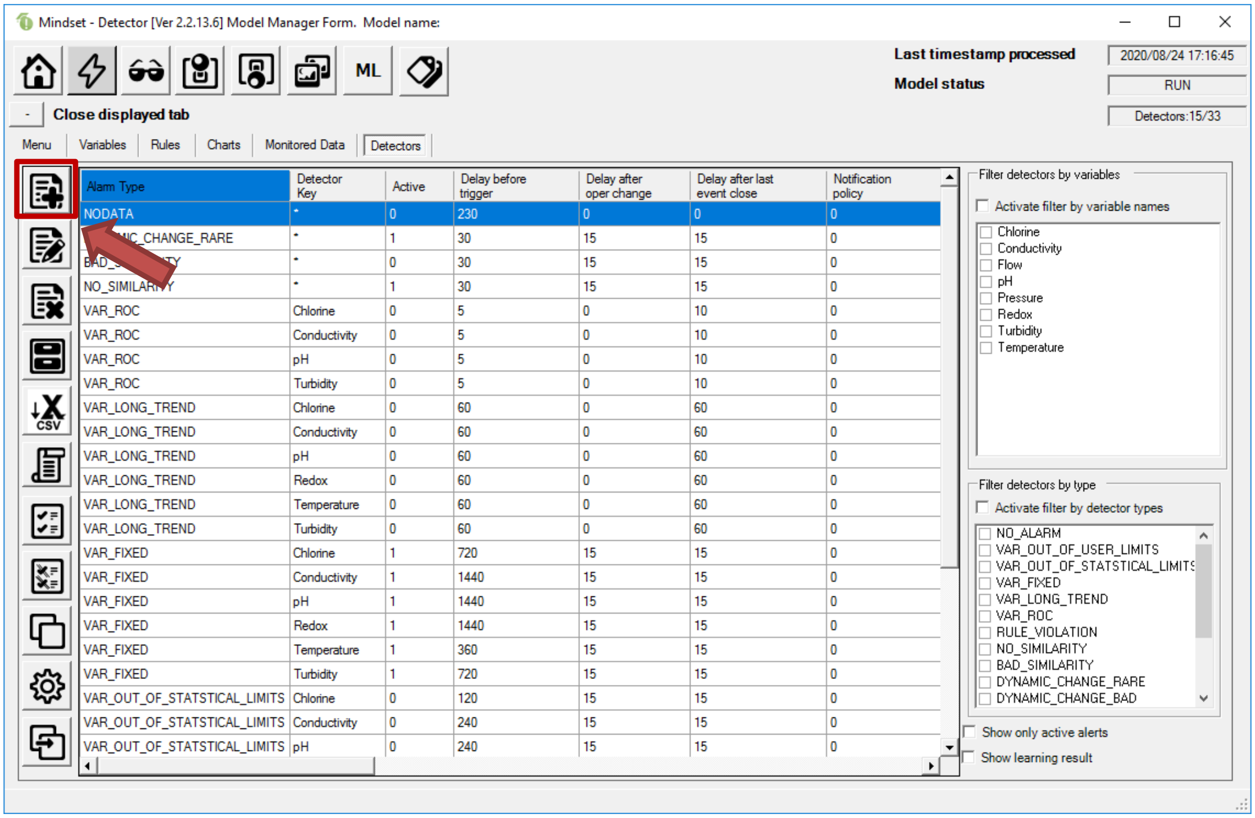
Last timestamp processed: 2020/08/24 17:11:44
Model status: RUN
Detectors:15/33

Model Notes
Rules violation
Similarity for known Hazard Event
Probability for Abnormal event
Sensors Problem
Event Severity: Low Medium High
Advisor Remarks
□ Display non-validation only
Access control: NOINDICATION

Model quality log
```
* Poor rang for low statistical limit for variable * Chlorine
* Poor rang for low statistical limit for variable * Conductivity
* Poor rang for low statistical limit for variable * Pressure
* Poor rang for high statistical limit for variable * Turbidity
 * Some of model's detectors are unstable *
```

Figure 107: The EDS model manager window for a specific EDS model. The button used for accessing the Detectors screen is marked in red.

Step III: Figure 108 shows the Detectors window for a given station. The "add new detector" button is marked in red. Click on this button to add a new detector, connecting SM models to the EDS.



Last timestamp processed: 2020/08/24 17:16:45

Model status: RUN

Detectors:15/33

Alarm Type	Detector Key	Active	Delay before trigger	Delay after oper change	Delay after last event close	Notification policy
NODATA	*	0	230	0	0	0
DYNAMIC_CHANGE_RARE	*	1	30	15	15	0
BAD_SIMILARITY	*	0	30	15	15	0
NO_SIMILARITY	*	1	30	15	15	0
VAR_ROC	Chlorine	0	5	0	10	0
VAR_ROC	Conductivity	0	5	0	10	0
VAR_ROC	pH	0	5	0	10	0
VAR_ROC	Turbidity	0	5	0	10	0
VAR_LONG_TREND	Chlorine	0	60	0	60	0
VAR_LONG_TREND	Conductivity	0	60	0	60	0
VAR_LONG_TREND	pH	0	60	0	60	0
VAR_LONG_TREND	Redox	0	60	0	60	0
VAR_LONG_TREND	Temperature	0	60	0	60	0
VAR_LONG_TREND	Turbidity	0	60	0	60	0
VAR_FIXED	Chlorine	1	720	15	15	0
VAR_FIXED	Conductivity	1	1440	15	15	0
VAR_FIXED	pH	1	1440	15	15	0
VAR_FIXED	Redox	1	1440	15	15	0
VAR_FIXED	Temperature	1	360	15	15	0
VAR_FIXED	Turbidity	1	720	15	15	0
VAR_OUT_OF_STATSTICAL_LIMITS	Chlorine	0	120	15	15	0
VAR_OUT_OF_STATSTICAL_LIMITS	Conductivity	0	240	15	15	0
VAR_OUT_OF_STATSTICAL_LIMITS	pH	0	240	15	15	0

Filter detectors by variables

☐ Activate filter by variable names

☐ Chlorine
☐ Conductivity
☐ Flow
☐ pH
☐ Pressure
☐ Redox
☐ Turbidity
☐ Temperature

Filter detectors by type

☐ Activate filter by detector types

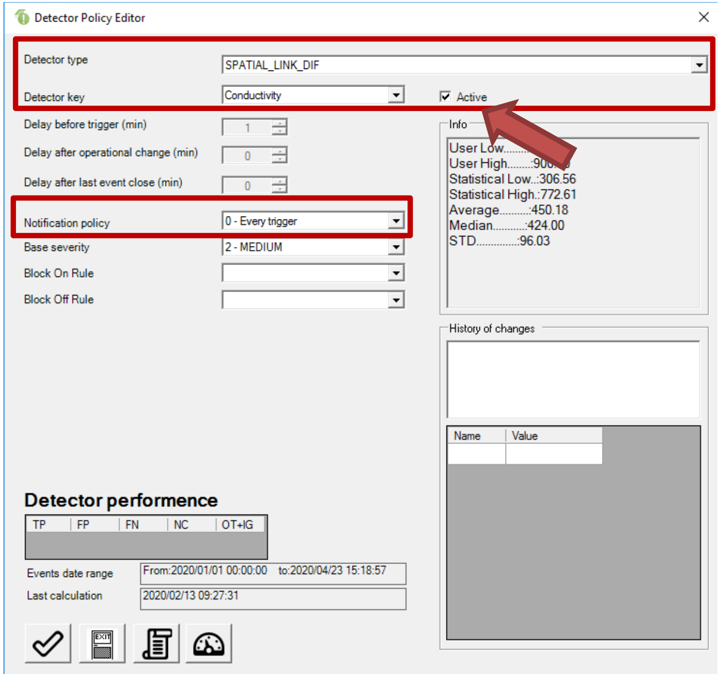
☐ NO_ALARM
☐ VAR_OUT_OF_USER_LIMITS
☐ VAR_OUT_OF_STATSTICAL_LIMITS
☐ VAR_FIXED
☐ VAR_LONG_TREND
☐ VAR_ROC
☐ RULE_VIOLATION
☐ NO_SIMILARITY
☐ BAD_SIMILARITY
☐ DYNAMIC_CHANGE_RARE
☐ DYNAMIC_CHANGE_BAD

☐ Show only active alerts

☐ Show learning result

Figure 108: The Detectors window. The button for adding a new detector in marked in red.

Step IV: Figure 109 shows the Detector Policy Editor Window. Select the Detector type to be "SPATIAL_LINK_DIF" and then select as the detector key the water quality measurement monitored by the SM model you wish to connect. The EDS will track all pair model with this station as their target node, and with this measurement, and will trigger EDS events based on their alarms.
Also make sure the **click the "Active" check box indicating the detector is active**. See the red rectangle at the top of Figure 109, and the red arrow marking the check box.



Detector Policy Editor

Detector type: SPATIAL_LINK_DIF

Detector key: Conductivity

☑ Active

Delay before trigger (min): 1

Delay after operational change (min): 0

Delay after last event close (min): 0

Notification policy: 0 - Every trigger

Base severity: 2 - MEDIUM

Block On Rule:

Block Off Rule:

Info

User Low........
User High........:900.
Statistical Low..:306.56
Statistical High.:772.61
Average..........:450.18
Median...........:424.00
STD.............:96.03

History of changes

Name	Value

Detector performence

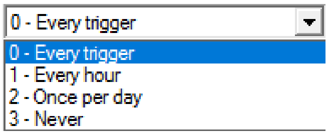
TP	FP	FN	NC	OT+IG

Events date range: From:2020/01/01 00:00:00 to:2020/04/23 15:18:57

Last calculation: 2020/02/13 09:27:31

Figure 109:Detector Policy Editor Window

Step V: As a notification policy (marked by the lower rectangular red box) in Figure 109, select "Every trigger", so the EDS will provide a notification for every alarm raised by the SM.



0 - Every trigger
0 - Every trigger
1 - Every hour
2 - Once per day
3 - Never

Step VI: click the approve button to finalize.

✓ - Approve selection

- Quit (disregard changes)

- Calculate detector performance

- Calibrate Delay times for alarms

4.8 Suggested Parameters, Case studies and examples

This section discusses the "off-the-shelf" suggested parameter configuration for SM models, as well as a variety of examples that are used to better explain the behavior of the spatial model under various settings. Graphs demonstrating the behavior of different models is used to serve as a yardstick for assessing the performance of user models

4.8.1 Suggested Parameters

For pair models, we suggest the following parameter configuration:

Flow time estimation:

- Preferably estimate flow times using all available data (unless network operation is substantially different when looking more than 2 years back).
- At least 2 months of data for typically behaved sensor. More data may be needed for noisy sensors.
- Split to no more than 15 flow regimes.
- Make sure regimes are 6-7 hours in length, for each regime.
- Set the maximum flow time to be considered to 1200 minutes (and less than 1 day at all times).
- Set the flow time resolution to 10-15 minutes.
- Set the window size for estimation of flow time to be 60-90 minutes (with 60 minutes for a sensor with little noise, and a very distinct time-varying signal in the data). For battery powered sensors – 120-150 minutes.
- Set window step size to be 1/3 of window size.
- Filter out measurements constant for more than 1 hour.

Prediction:

- Train using 30 days of data.
- Use a 95% prediction interval, i.e. $\alpha_L = \alpha_U = 0.025$
- Set the delay time to be 1 hour
- **Hard-coded as of November 2020:** prediction interval will be set to be no less than 10% of measurement user limits.
- **Hard-coded as of November 2020:** alarms will not be raised if distance between actual and predicted is not more than 20% of measurement statistical limits.
- **Hard coded as of November 2020:** the 5% most recent errors (residuals) will not be used when estimating the distribution of errors.
- Set the delay time to 30 minutes.

Triplets:

- $\gamma = 4$
- For fork-out triplets, set IsPivot to FALSE.
- $\alpha_L = \alpha_U = 0.025$
- A delay time of one hour.
- Set M to be the same as in pair models (see description in Section 3.7).

4.8.2 Case of events detected both by EDS and Spatial Model

We examine a case of a drop in Free Chlrone event identified by both the EDS and the Spatial Model. Figure 110 shows the drop in Free Chlorine, falling below a value of 0.2 mg/L, as marked by a red cicle. Figure 111 shows this events, built of several small "drops" in F-CL value was identified by the EDS. Figure 112 shows this event was also identified by the Spatial Model.

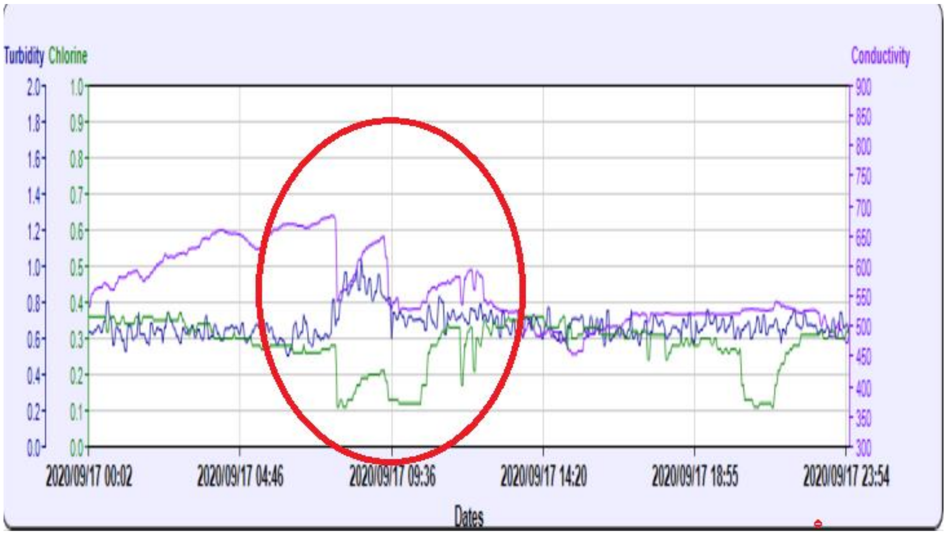
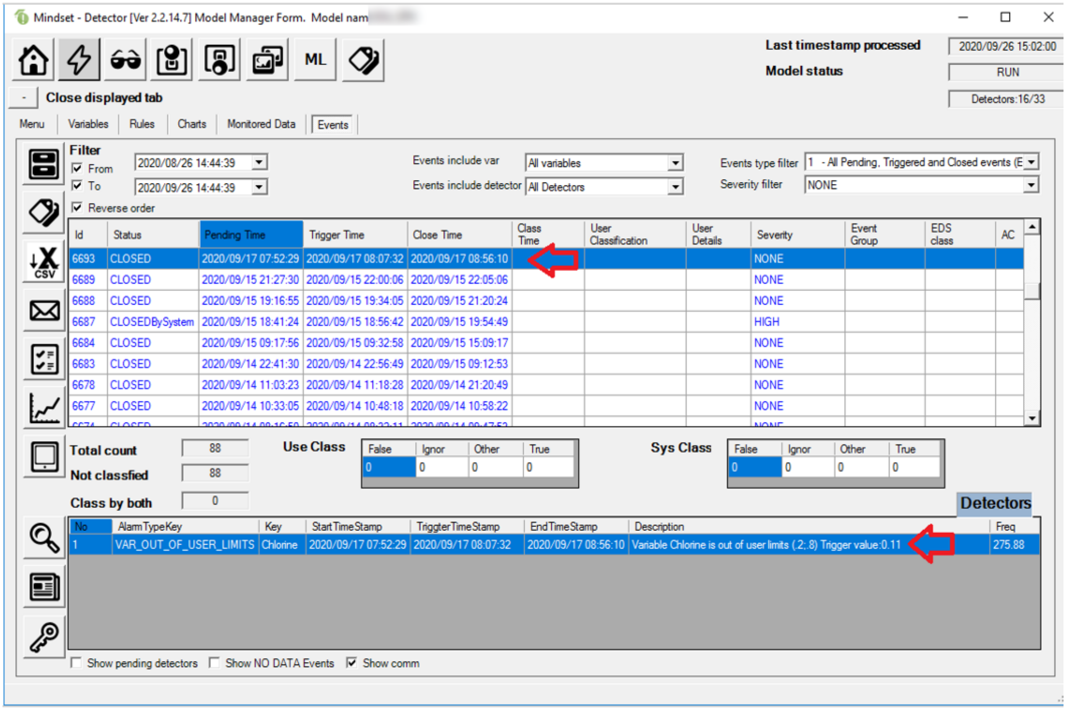


Figure 110: An example for a drop in Chlorine Event



Last timestamp processed 2020/09/26 15:02:00

Model status RUN

Detectors:16/33

Filter

☑ From 2020/08/26 14:44:39

☑ To 2020/09/26 14:44:39

☑ Reverse order

Events include var: All variables

Events include detector: All Detectors

Events type filter: 1 - All Pending, Triggered and Closed events (E

Severity filter: NONE

Id	Status	Pending Time	Trigger Time	Close Time	Class Time	User Classification	User Details	Severity	Event Group	EDS class	AC
6693	CLOSED	2020/09/17 07:52:29	2020/09/17 08:07:32	2020/09/17 08:56:10				NONE			
6689	CLOSED	2020/09/15 21:27:30	2020/09/15 22:00:06	2020/09/15 22:05:06				NONE			
6688	CLOSED	2020/09/15 19:16:55	2020/09/15 19:34:05	2020/09/15 21:20:24				NONE			
6687	CLOSEDBySystem	2020/09/15 18:41:24	2020/09/15 18:56:42	2020/09/15 19:54:49				HIGH			
6684	CLOSED	2020/09/15 09:17:56	2020/09/15 09:32:58	2020/09/15 15:09:17				NONE			
6683	CLOSED	2020/09/14 22:41:30	2020/09/14 22:56:49	2020/09/15 09:12:53				NONE			
6678	CLOSED	2020/09/14 11:03:23	2020/09/14 11:18:28	2020/09/14 21:20:49				NONE			
6677	CLOSED	2020/09/14 10:33:05	2020/09/14 10:48:18	2020/09/14 10:58:22				NONE			

Total count 88

Not classfied 88

Class by both 0

Use Class

False	Ignor	Other	True
0	0	0	0

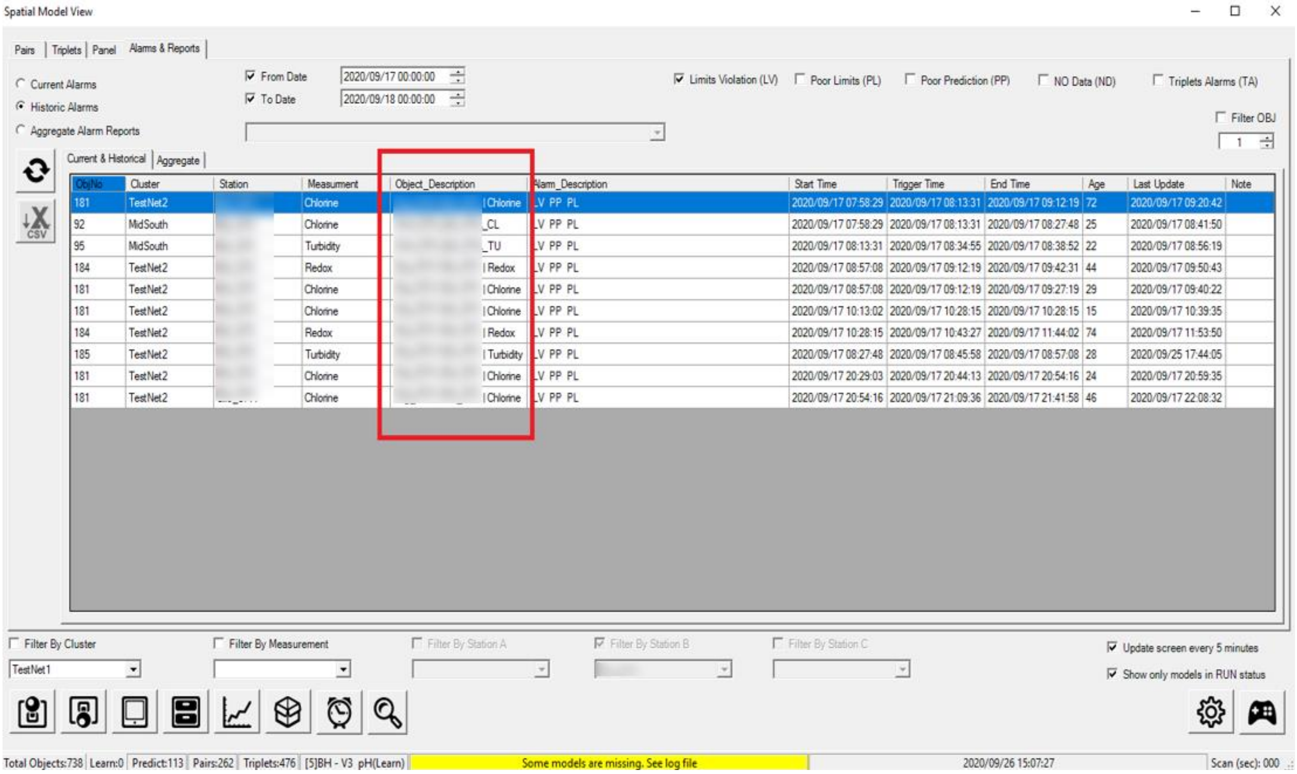
Sys Class

False	Ignor	Other	True
0	0	0	0

No	AlarmTypeKey	Key	StartTimeStamp	TriggterTimeStamp	EndTimeStamp	Description	Freq
1	VAR_OUT_OF_USER_LIMITS	Chlorine	2020/09/17 07:52:29	2020/09/17 08:07:32	2020/09/17 08:56:10	Variable Chlorine is out of user limits (.2;.8) Trigger value:0.11	275.88

☐ Show pending detectors ☐ Show NO DATA Events ☑ Show comm

Figure 111: EDS alarms for the drop in F-CL event showed in Figure 110.



○ Current Alarms
● Historic Alarms
○ Aggregate Alarm Reports

☑ From Date 2020/09/17 00:00:00
☑ To Date 2020/09/18 00:00:00

☑ Limits Violation (LV) □ Poor Limits (PL) □ Poor Prediction (PP) □ NO Data (ND) □ Triplets Alarms (TA)

□ Filter OBJ 1

ObjNo	Cluster	Station	Measurment	Object_Description	Alarm_Description	Start Time	Trigger Time	End Time	Age	Last Update	Note
181	TestNet2	[redacted]	Chlorine	[redacted] | Chlorine	LV PP PL	2020/09/17 07:58:29	2020/09/17 08:13:31	2020/09/17 09:12:19	72	2020/09/17 09:20:42	
92	MidSouth	[redacted]	Chlorine	[redacted] _CL	LV PP PL	2020/09/17 07:58:29	2020/09/17 08:13:31	2020/09/17 08:27:48	25	2020/09/17 08:41:50	
95	MidSouth	[redacted]	Turbidity	[redacted] _TU	LV PP PL	2020/09/17 08:13:31	2020/09/17 08:34:55	2020/09/17 08:38:52	22	2020/09/17 08:56:19	
184	TestNet2	[redacted]	Redox	[redacted] | Redox	LV PP PL	2020/09/17 08:57:08	2020/09/17 09:12:19	2020/09/17 09:42:31	44	2020/09/17 09:50:43	
181	TestNet2	[redacted]	Chlorine	[redacted] | Chlorine	LV PP PL	2020/09/17 08:57:08	2020/09/17 09:12:19	2020/09/17 09:27:19	29	2020/09/17 09:40:22	
181	TestNet2	[redacted]	Chlorine	[redacted] | Chlorine	LV PP PL	2020/09/17 10:13:02	2020/09/17 10:28:15	2020/09/17 10:28:15	15	2020/09/17 10:39:35	
184	TestNet2	[redacted]	Redox	[redacted] | Redox	LV PP PL	2020/09/17 10:28:15	2020/09/17 10:43:27	2020/09/17 11:44:02	74	2020/09/17 11:53:50	
185	TestNet2	[redacted]	Turbidity	[redacted] | Turbidity	LV PP PL	2020/09/17 08:27:48	2020/09/17 08:45:58	2020/09/17 08:57:08	28	2020/09/25 17:44:05	
181	TestNet2	[redacted]	Chlorine	[redacted] | Chlorine	LV PP PL	2020/09/17 20:29:03	2020/09/17 20:44:13	2020/09/17 20:54:16	24	2020/09/17 20:59:35	
181	TestNet2	[redacted]	Chlorine	[redacted] | Chlorine	LV PP PL	2020/09/17 20:54:16	2020/09/17 21:09:36	2020/09/17 21:41:58	46	2020/09/17 22:08:32	

□ Filter By Cluster TestNet1
□ Filter By Measurement
□ Filter By Station A
☑ Filter By Station B [redacted]
□ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:738 | Learn:0 | Predict:113 | Pairs:262 | Triplets:476 | [5]BH - V3 pH(Learn) | Some models are missing. See log file | 2020/09/26 15:07:27 | Scan (sec): 000

Figure 112: SM alarms for the drop in Chlorine event showed in Figure 110. The relevant rows are the ones with a measurement of Chlorine and a "TestNet2" cluster identity.

4.8.3 Case of an early warning

We show an example of the SM resulted in an early detection of an irregular event with increased conductivity. Figure 113 shows a case of an increased conductivity value reaching 950 $\mu S/cm$ at 3:00 AM 10th September 2020. This event triggered an EDS alarm. Figure 114 shows an Event was triggered by the SM, up to 36 hours before, due to a prior increase in conductivity. A look at the increases in conductivity in Figure HHH shows the increase experienced in Figure 114 is in fact a pending problem, that could have been identified before.

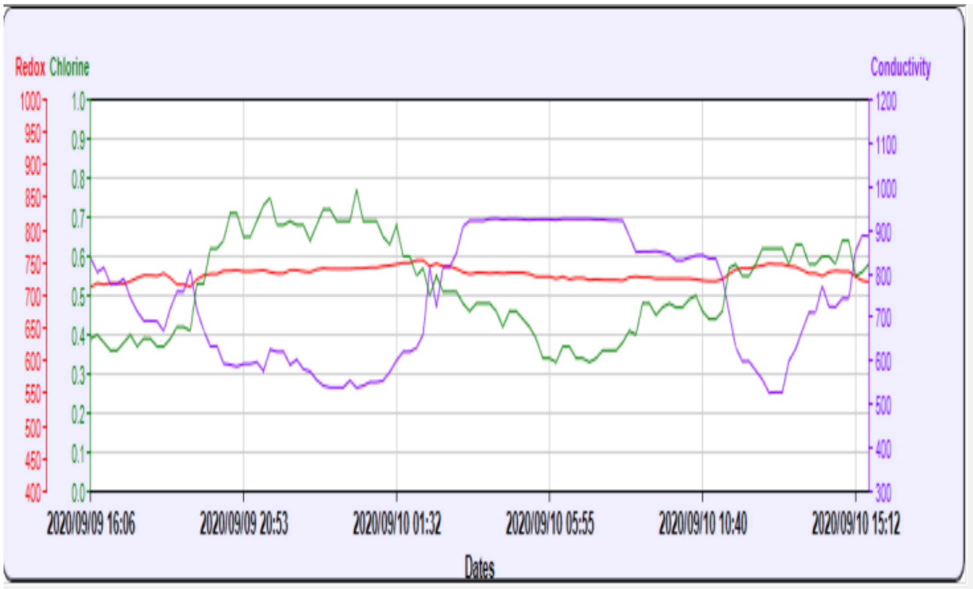


Figure 113: An increase in Conductivity identified by the EDS.

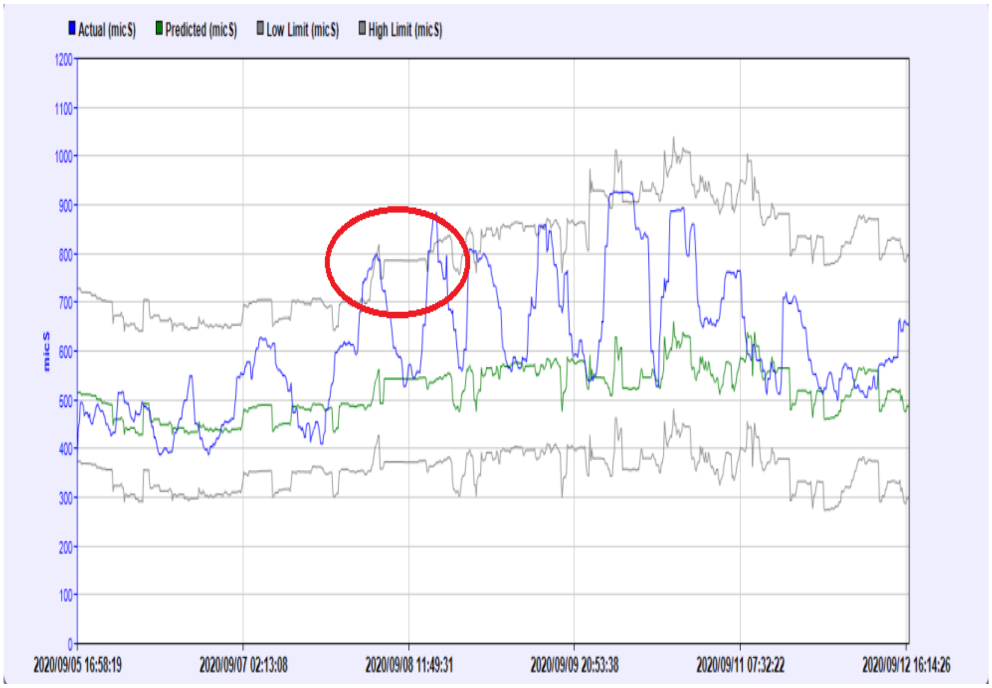


Figure 114: Detection of irregular Conductivity measurements by the SM, prior to the event in Figure HHH.

4.8.4 Additional Examples for real events in data

We show examples for additional events detected in data by the SM. Figure 115 shows an example of a spike in Turbidity detected by the SM. Figure 116 shows an example of a slow increase in conductivity detected by the SM.

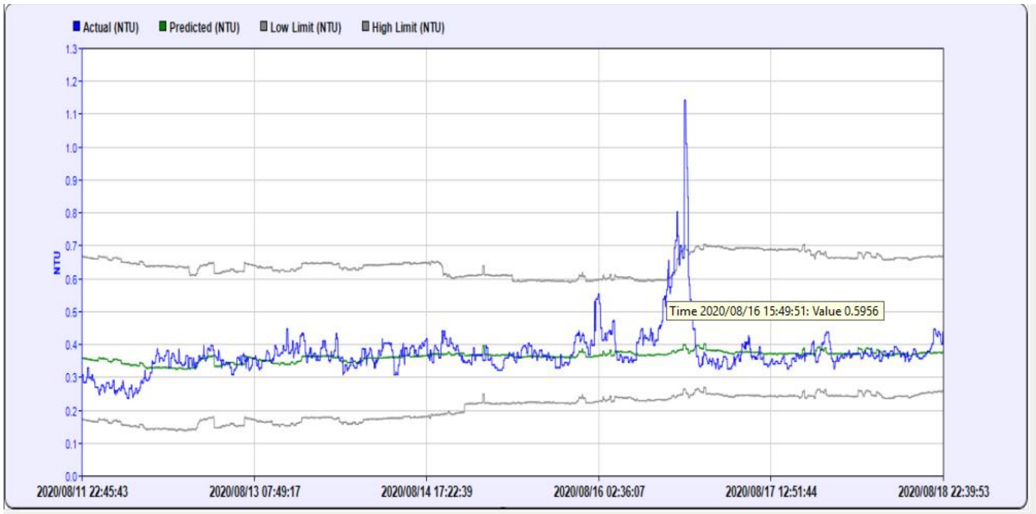


Figure 115: An event with a spike in conductivity detected by the SM.

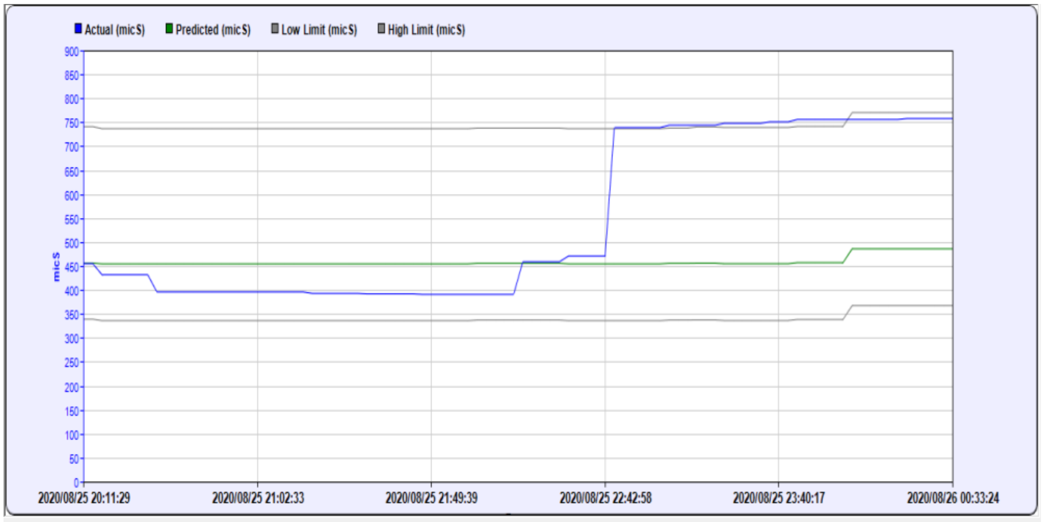


Figure 116: An event with a slow increase in conductivity detected by the SM.

4.8.5 Comparison of conductivity models

We show a comparison of two SM pair prediction models for conductivity measurements. The comparison is given in Figure 117. Subfigures A and B in Figure 117 show the actual vs. predicted plot and histogram for prediction errors, respectively, for the conductivity prediction model showing a good fit to the data. Subfigures C and D show the same plots of the same type for a conductivity model with a substantially worse fit to the data.

In subfigure C we see points in the actual vs. predicted line don't match a 45 degree angle. The errors in subfigure D are substantially more dispersed than in Figure B.

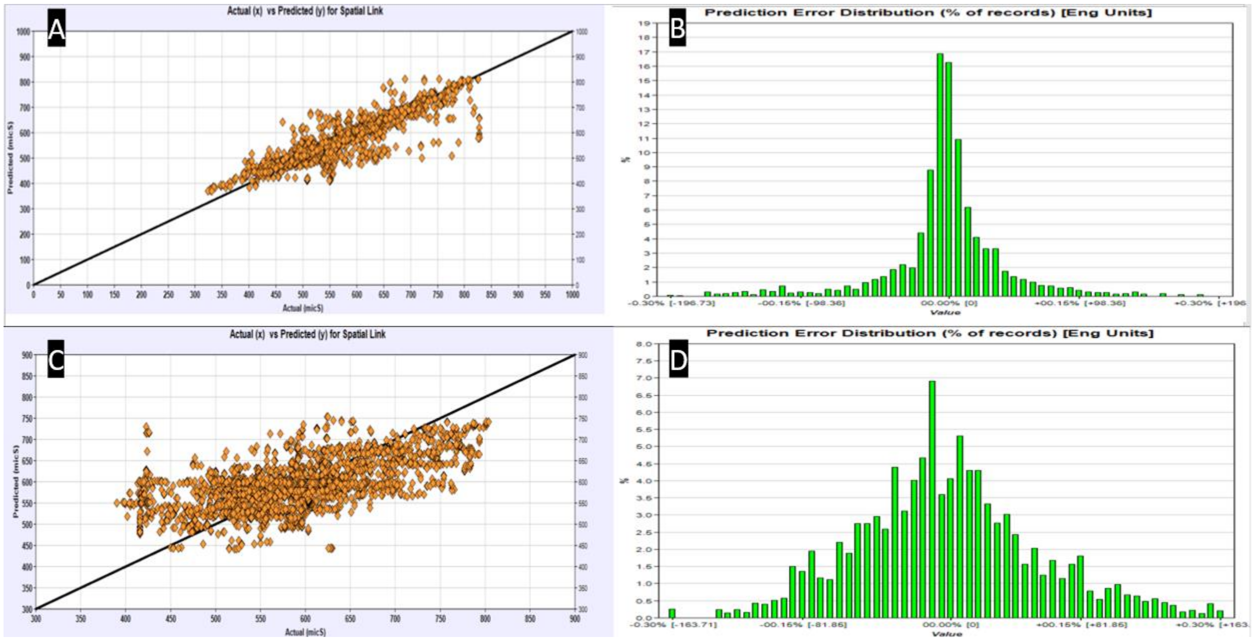


Figure 117: Comparison of the models for predicting conductivity in two SM pair models. Subfigures A and B show the actual vs. predicted plot and histrogram of prediction errors for a model showing good fit to the data. Subfigures C and D show the same charts for a model with worse fit to the data.

4.8.6 Comparison of Redox models

We show a comparison of two SM pair prediction models for redox measurements. The comparison is given in Figure 118. Subfigures A and B in Figure 118 show the actual vs. predicted plot and histogram for prediction errors, respectively, for the redox prediction model showing a good fit to the data. Subfigures C and D show the same plots of the same type for a redox model with a substantially worse fit to the data.

We observe points in the actual vs. predicted plot in Subfigure C to be substantially more dispersed than the points in Subfigure A.

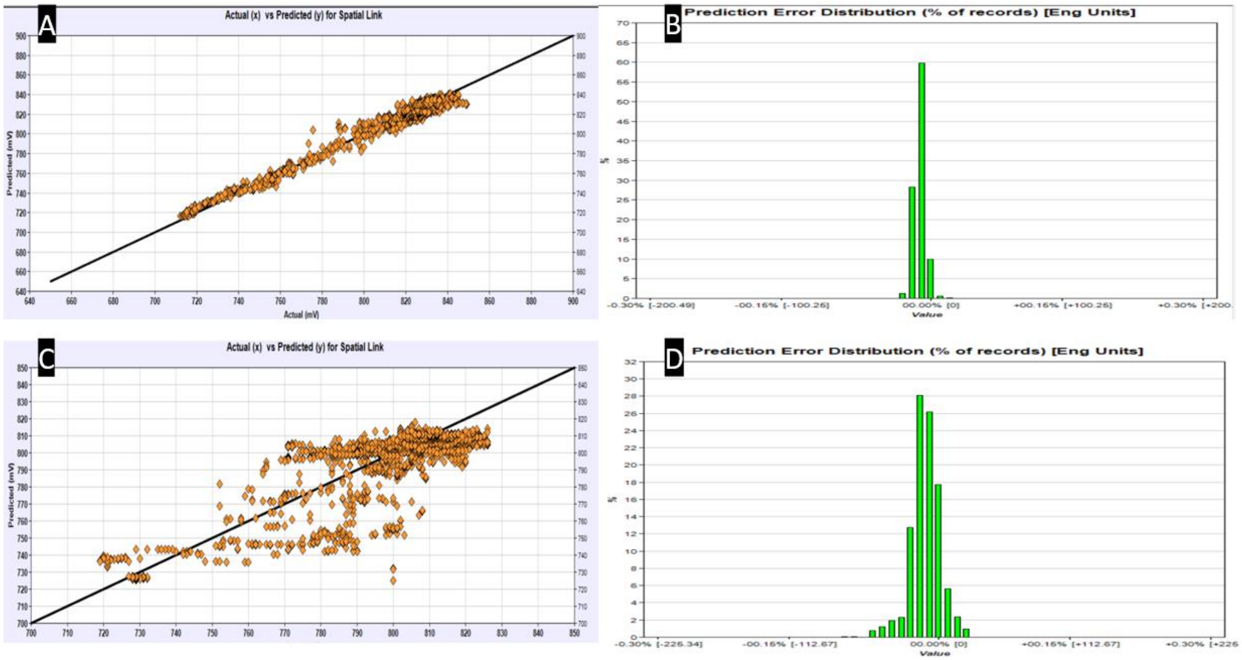


Figure 118: Comparison of the models for predicting redox in two SM pair models. Subfigures A and B show the actual vs. predicted plot and histrogram of prediction errors for a model showing good fit to the data. Subfigures C and D show the same charts for a model with worse fit to the data.

4.8.7 Comparison of Turbidity models

We show a comparison of two SM pair prediction models for turbidity measurements. The comparison is given in Figure 119. Subfigures A and B in Figure 119 show the actual vs. predicted plot and histogram for prediction errors, respectively, for the turbidity prediction model showing a good fit to the data. Subfigures C and D show the same plots of the same type for a turbidity model with a substantially worse fit to the data.

In subfigure C we see points in the actual vs. predicted line don't match a 45 degree angle. The errors in subfigure D are substantially more dispersed than in Figure B, and the distribution of prediction errors is not symmetric around zero.

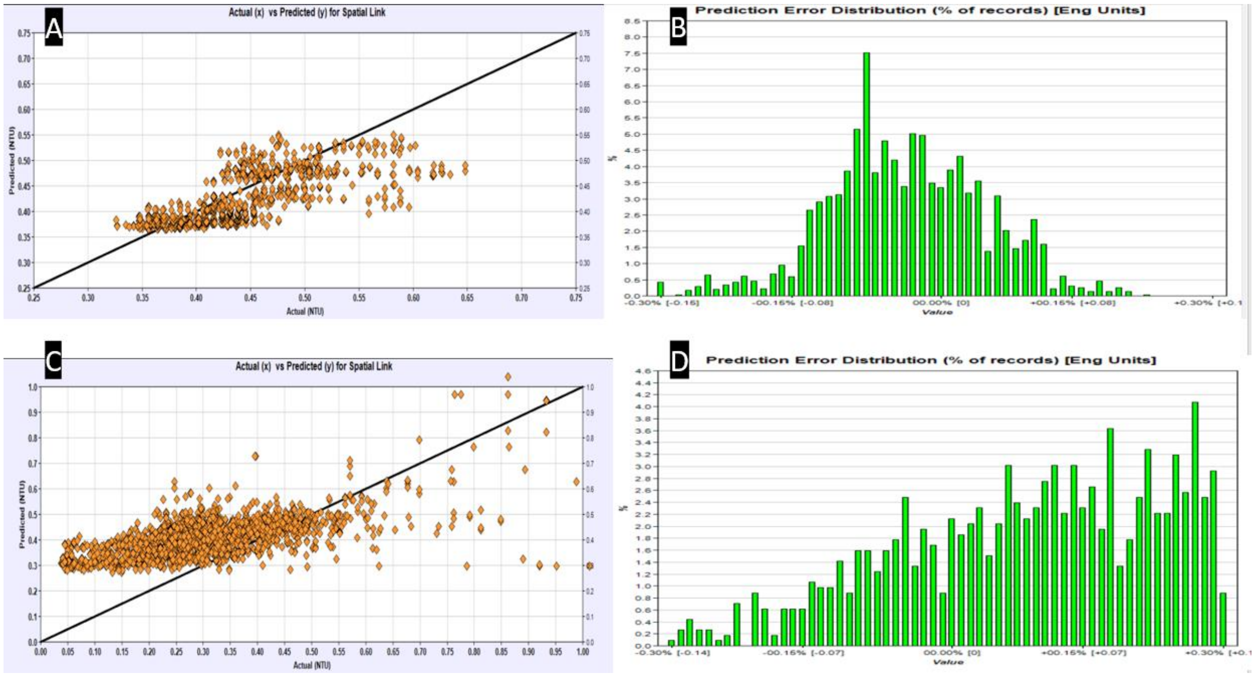


Figure 119: Comparison of the models for predicting Turbidity in two SM pair models. Subfigures A and B show the actual vs. predicted plot and histrogram of prediction errors for a model showing good fit to the data. Subfigures C and D show the same charts for a model with worse fit to the data.

4.8.8 Examples for poor efficiency of spatial model limits

We show two examples where spatial model limits are inefficient, and are larger than 50% of the range between user limits. The examples are shown in Figures 120 and 121.

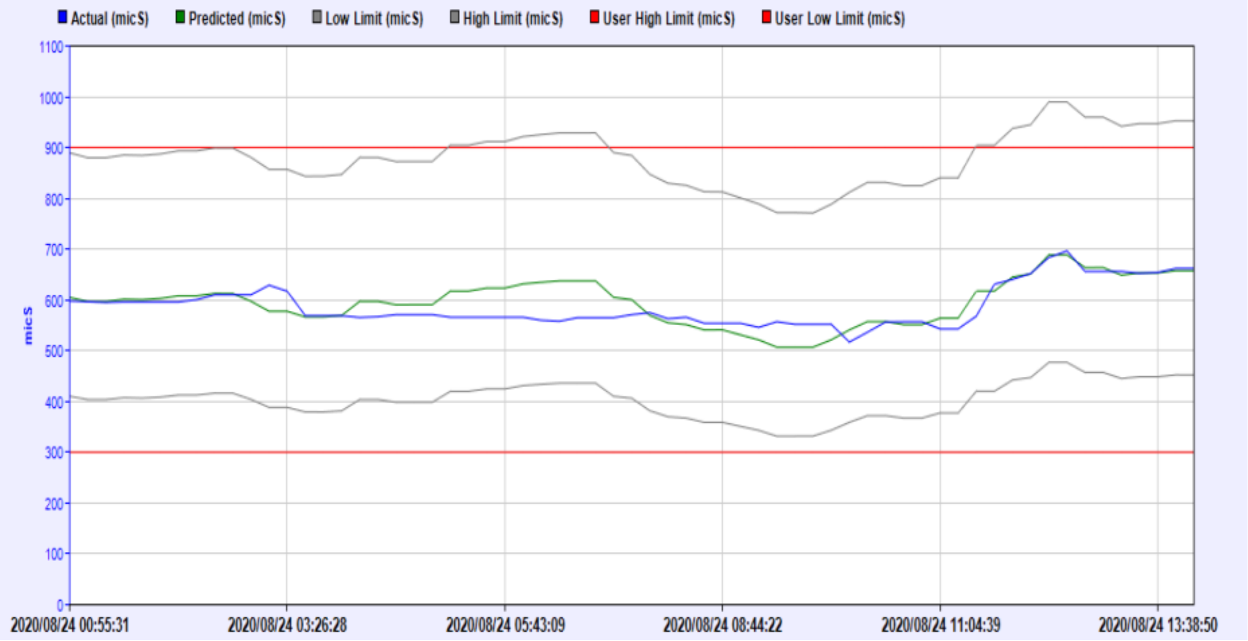


Figure 120: An example with inefficient Spatial Model limits when monitoring conductivity measurements.

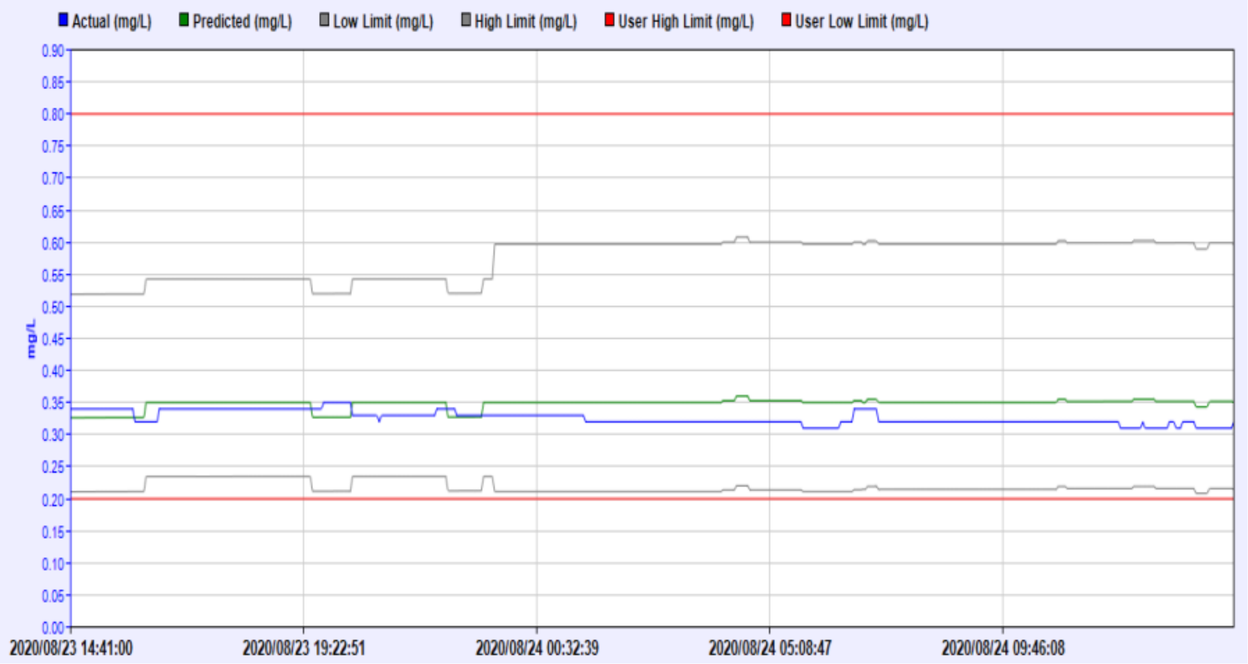


Figure 121: An example with inefficient Spatial Model limits when monitoring Free-CL measurements.

4.8.9 Limit time for an event

This example shows the importance of filtering the most recent 5% of prediction errors from the distribution used for computing the Spatial Model limits in Section 3.5.2. Figure 122 shows an example of an abnormal increase in pH, **marked by the red arrow**. The abnormal pH measurements continue for more than a day, with values returning to normal only at the timepoint indicated by the **purple arrow**. However, the abnormal prediction errors become a substantial portion of the training set already at the time point marked by the **green arrow**. Hence, the difference between the abnormal value of pH, and the current source value stops being irregular, and SM limits show the current value of target station pH is not abnormal. In order to avoid a case where prediction errors from a water quality event affect the detection thresholds, the most recent 5% of prediction errors are removed from the data.

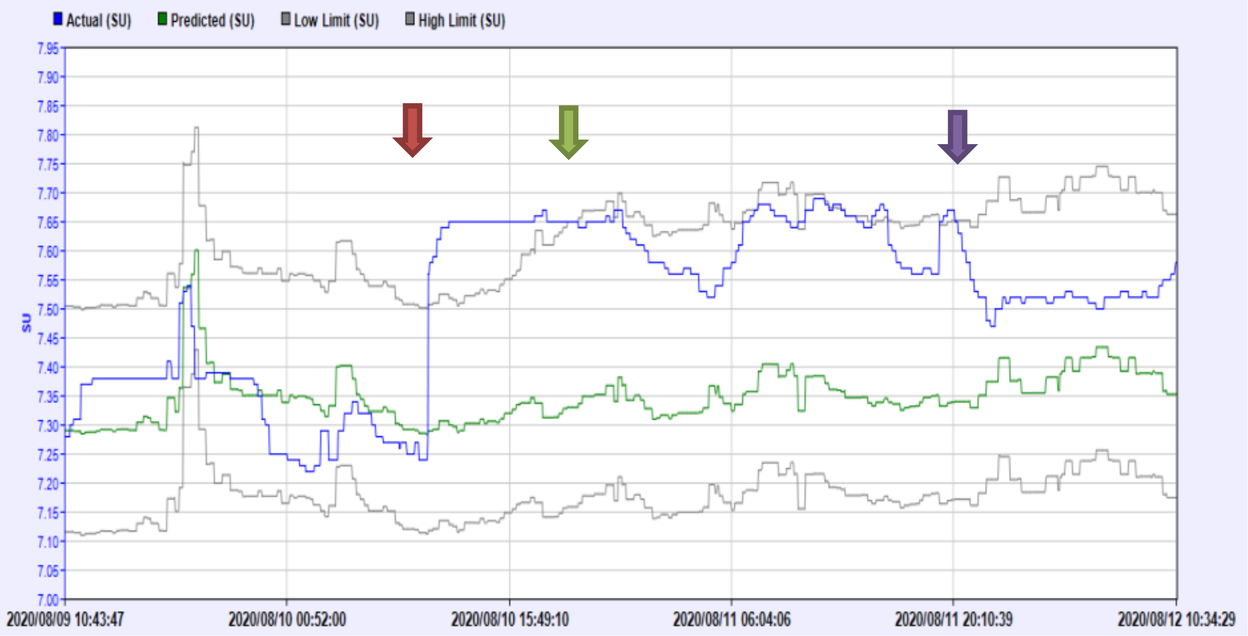


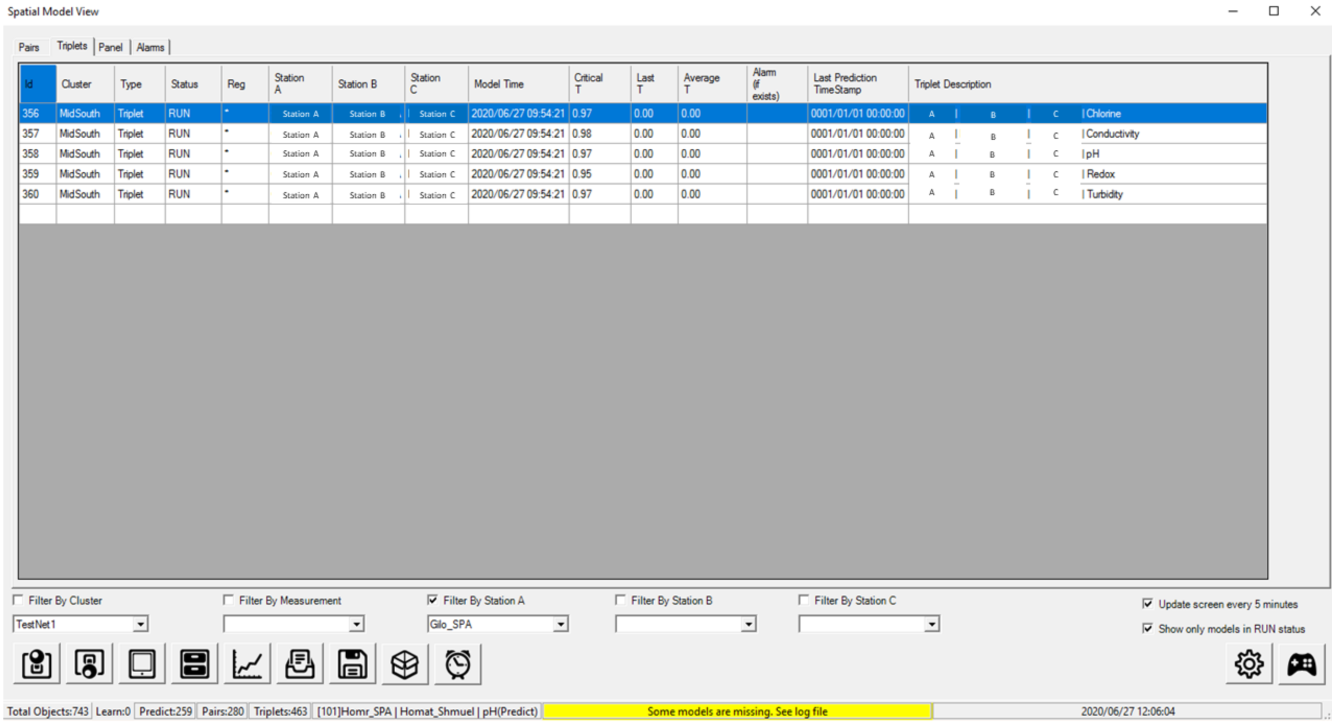
Figure 122: Example of why the most recent prediction errors should be not be used for setting the SM event limits. An abnormally high value of pH is observed starting from the time point indicated by the red arrow until the timepoint indicated by the purple arrow. Target station measurements are not declared as abnormal starting from the time point marked by the green arrow, since large prediction errors have entered the empirical distribution used for setting SM event limits.

4.9 Triplet model – definition

The triplets tab in the Spatial Model main window (Figure 123) lists all defined triplet models. The values listed for each mode are:

- The triplet ID (unique value)
- The Cluster identity for each triplet
- Triplet type – One of three values: triplet (line)/ Fork out/ Fork in
- Triplet status (Run/ Stop).
- Reg - The network regimes in which the model is active (user's choice, as in Section 4.3)
- The three stations featured in the triplet – Denoted by Station A/B/C.
- The model time for each triplet model – the most recent timestamp, by sensor clocks, that can be monitored by the triplet model. See Section 3.6 for description of the time adjustments between sensor times, used in order to determine this field.
- Indication if a triplet alarm exists
- The Last predicted timestamp. – the last time (by server clock) a triplet was monitored
- The triplet description – a descriptive string usually listing the three stations in the triplets, along with the monitored measurement name.

Cluster and station filtration options (using the toolbar) described in Section 4.1 are also applied to this screen.



Spatial Model View

Id	Cluster	Type	Status	Reg	Station A	Station B	Station C	Model Time	Critical T	Last T	Average T	Alarm (if exists)	Last Prediction TimeStamp	Triplet Description
356	MidSouth	Triplet	RUN	*	Station A	Station B	Station C	2020/06/27 09:54:21	0.97	0.00	0.00		0001/01/01 00:00:00	A | B | C |Chlorine
357	MidSouth	Triplet	RUN	*	Station A	Station B	Station C	2020/06/27 09:54:21	0.98	0.00	0.00		0001/01/01 00:00:00	A | B | C |Conductivity
358	MidSouth	Triplet	RUN	*	Station A	Station B	Station C	2020/06/27 09:54:21	0.97	0.00	0.00		0001/01/01 00:00:00	A | B | C |pH
359	MidSouth	Triplet	RUN	*	Station A	Station B	Station C	2020/06/27 09:54:21	0.95	0.00	0.00		0001/01/01 00:00:00	A | B | C |Redox
360	MidSouth	Triplet	RUN	*	Station A	Station B	Station C	2020/06/27 09:54:21	0.97	0.00	0.00		0001/01/01 00:00:00	A | B | C |Turbidity

□ Filter By Cluster: TestNet1
□ Filter By Measurement:
☑ Filter By Station A: Gilo_SPA
□ Filter By Station B:
□ Filter By Station C:
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:743 | Learn:0 | Predict:259 | Pairs:280 | Triplets:463 | [101]Homr_SPA | Homat_Shmuel | pH(Predict) | Some models are missing. See log file | 2020/06/27 12:06:04

Figure 123: The triplets tab

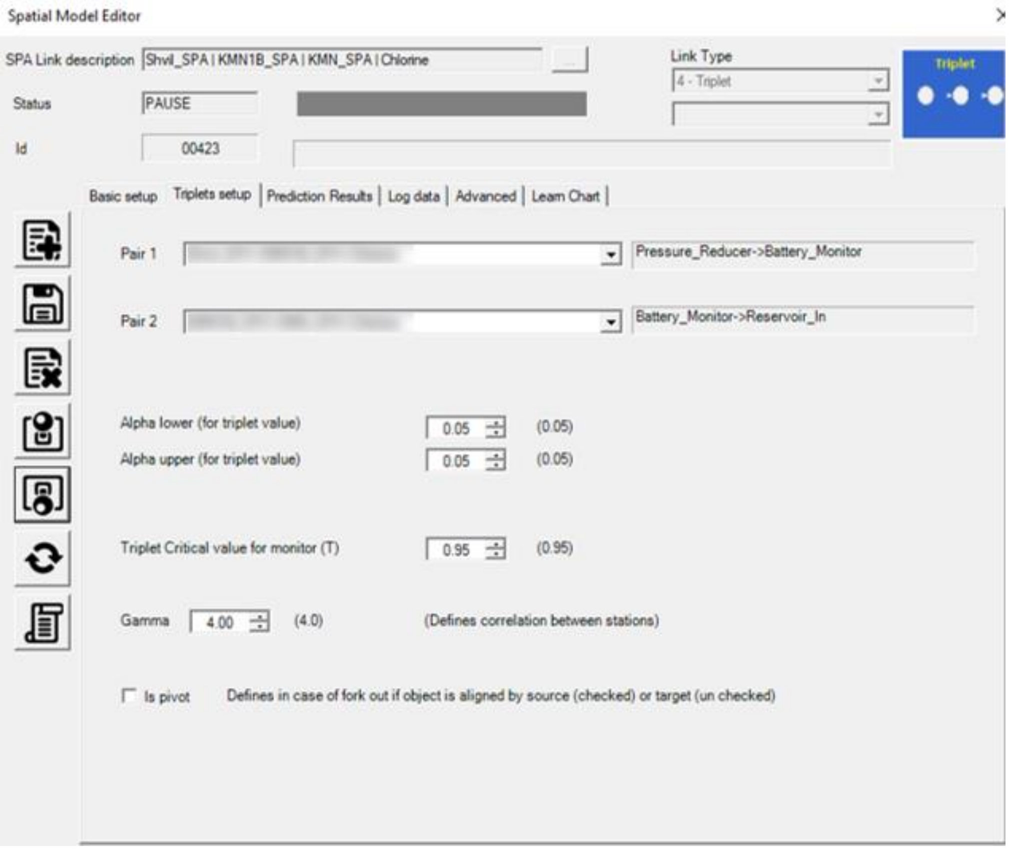
New triplet models may be defined using the Spatial Model Editor. Figure 124 shows the 'Triplets Setup' tab in the model editor. To define a triplet model :

1. select the type of triplet, as shown in Figure 124. Triplet types may be either: fork-in/ fork-out or triplet (line).
2. Select the model name ("Spa Link Description").
3. Select the two pairs for the triplet model.

    For a line/triplet model – Pair 1 should be the upstream pair, and pair 2 should be the downstream pair. The two pairs should have a shared node – the target node for pair 1 should be the source node for pair 2.

    For a fork out model, the two pairs should share the source node

    For a fork in model, the two pairs should share the target node.
4. Select the additional parameters, $\alpha_L, \alpha_U, \gamma, isPivot$ and the critical value (given by percentile) for calculating the alarm threshold. See Section 3.7 in Unit B for a complete description of the parameters.



Spatial Model Editor

SPA Link description: Shvil_SPA | KMN1B_SPA | KMN_SPA | Chlorine

Status: PAUSE

Id: 00423

Link Type: 4 - Triplet

Pair 1: [redacted] Pressure_Reducer->Battery_Monitor

Pair 2: [redacted] Battery_Monitor->Reservoir_In

Alpha lower (for triplet value) 0.05 (0.05)

Alpha upper (for triplet value) 0.05 (0.05)

Triplet Critical value for monitor (T) 0.95 (0.95)

Gamma 4.00 (4.0) (Defines correlation between stations)

☐ Is pivot Defines in case of fork out if object is aligned by source (checked) or target (un checked)

Figure 124: Defining a triplet model using the Spatial Model Editor window.

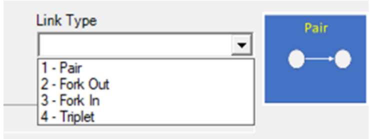
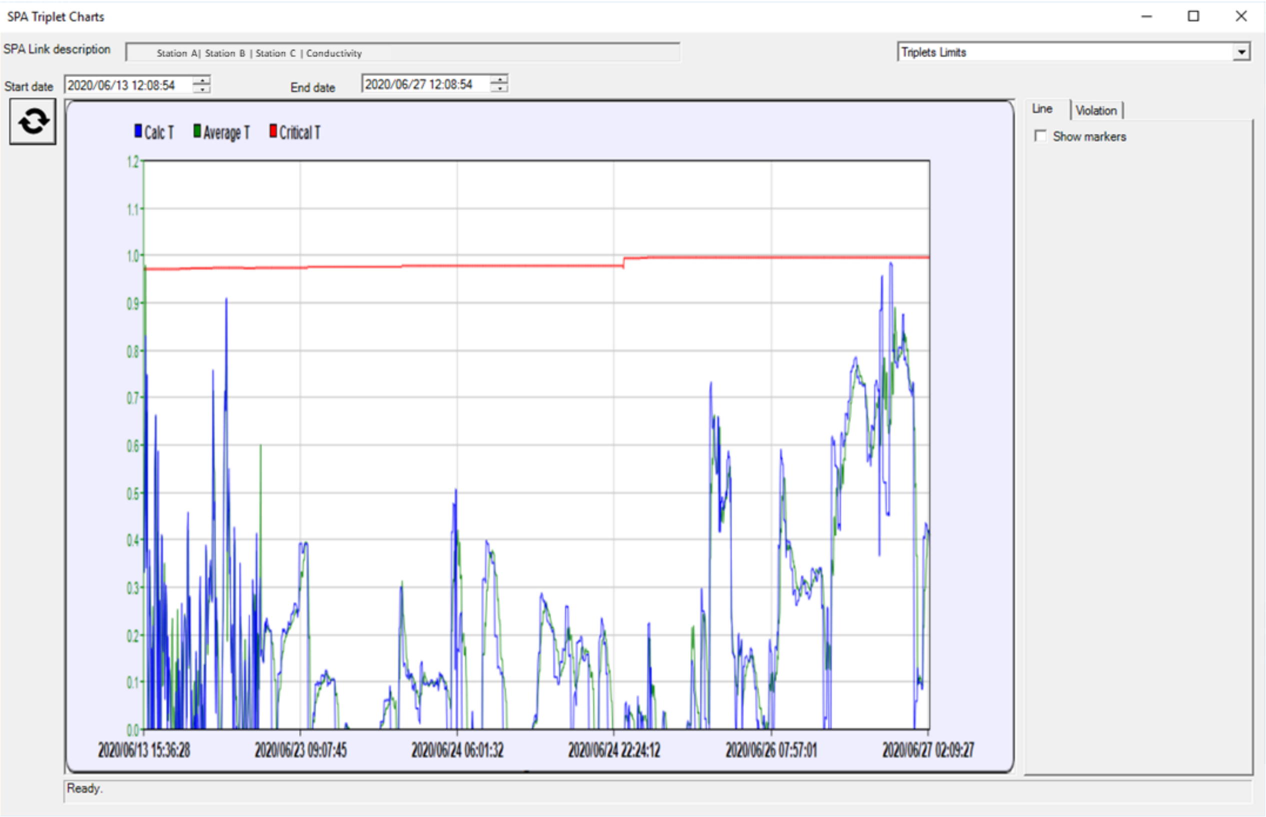


Figure 125: Selecting the type of triplet model using the 'Link Type' selection box

The charts window for a triplet shows the 'Triplets Limits graph'. This graph shows:

- The triplet's T value
- The T value averaged across the last 3 timestamps, in 10 minute increments (moving average)
- The critical T value, as computed by the algorithm described in Section 3.7.


SPA Triplet Charts

SPA Link description: Station A| Station B | Station C | Conductivity

Triplets Limits

Start date: 2020/06/13 12:08:54

End date: 2020/06/27 12:08:54

☐ Show markers

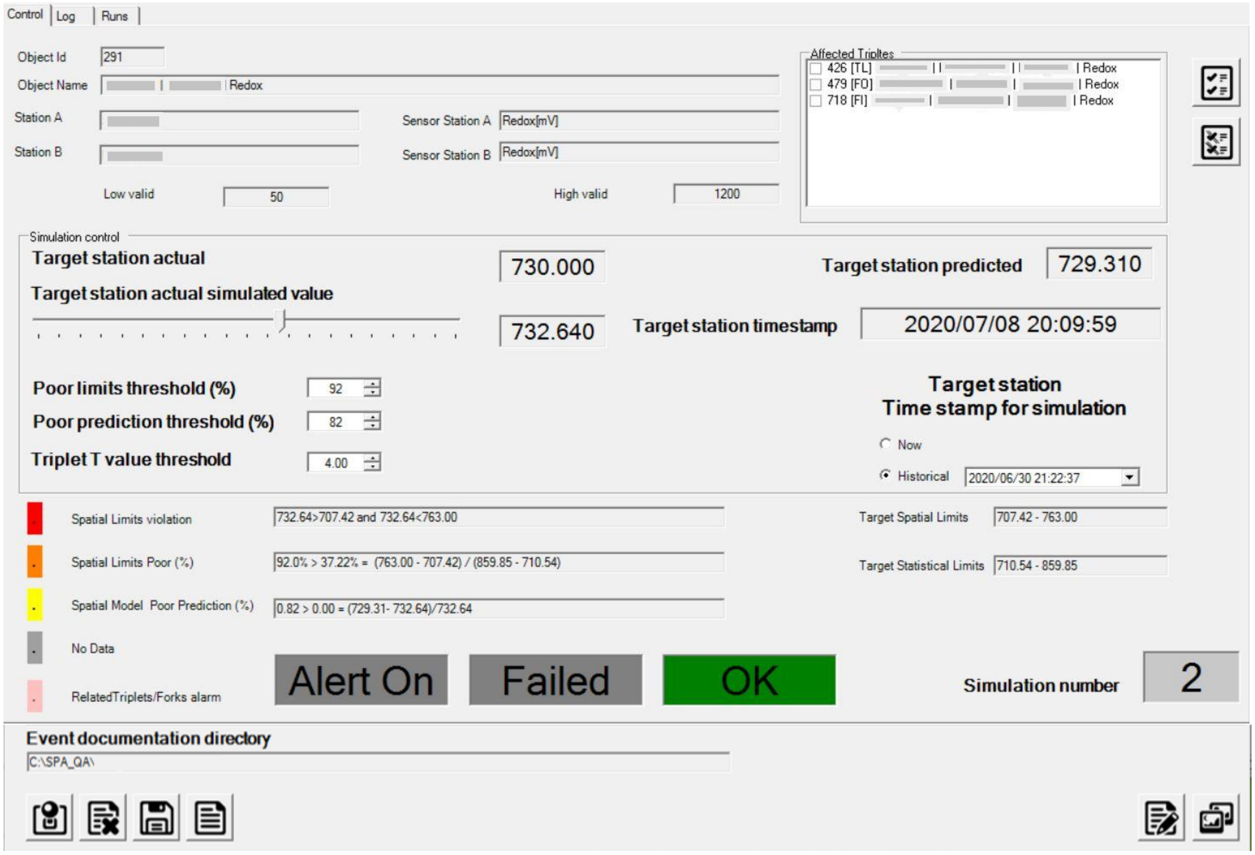
Ready.

Figure 126: Triplets limits chart for a triplet model.

4.10 Running simulation scenarios

The simulation screen allows the user to run 'what-if' scenarios and analyze the models' sensitivity to possible events in the data. The simulation screen is shown on Figure 127 and is accessible through the main Spatial Model screen toolbar, see details in Section 4.1. The user selects (highlights) a Pair model in the Pairs tab, and click the "simulations" icon (🎮) in the main toolbar.

In the next few pages, and using Figures 128-130, we will go over the different fields and options available in the simulations menu. See the next page for the detailed description.



Object Id 291

Object Name [redacted] | [redacted] | Redox

Station A [redacted]

Sensor Station A Redox[mV]

Station B [redacted]

Sensor Station B Redox[mV]

Low valid 50

High valid 1200

Affected Tripltes

- □ 426 [TL] [redacted] || [redacted] || [redacted] | Redox
- □ 479 [FO] [redacted] | [redacted] | [redacted] | Redox
- □ 718 [FI] [redacted] | [redacted] | [redacted] | Redox

Simulation control

Target station actual 730.000

Target station predicted 729.310

Target station actual simulated value 732.640

Target station timestamp 2020/07/08 20:09:59

Poor limits threshold (%) 92

Poor prediction threshold (%) 82

Triplet T value threshold 4.00

Target station Time stamp for simulation

○ Now

● Historical 2020/06/30 21:22:37

🟥 Spatial Limits violation 732.64>707.42 and 732.64<763.00

🟧 Spatial Limits Poor (%) 92.0% > 37.22% = (763.00 - 707.42) / (859.85 - 710.54)

🟨 Spatial Model Poor Prediction (%) 0.82 > 0.00 = (729.31- 732.64)/732.64

⬛ No Data

▨ RelatedTriplets/Forks alarm

Target Spatial Limits 707.42 - 763.00

Target Statistical Limits 710.54 - 859.85

Simulation number 2

Event documentation directory

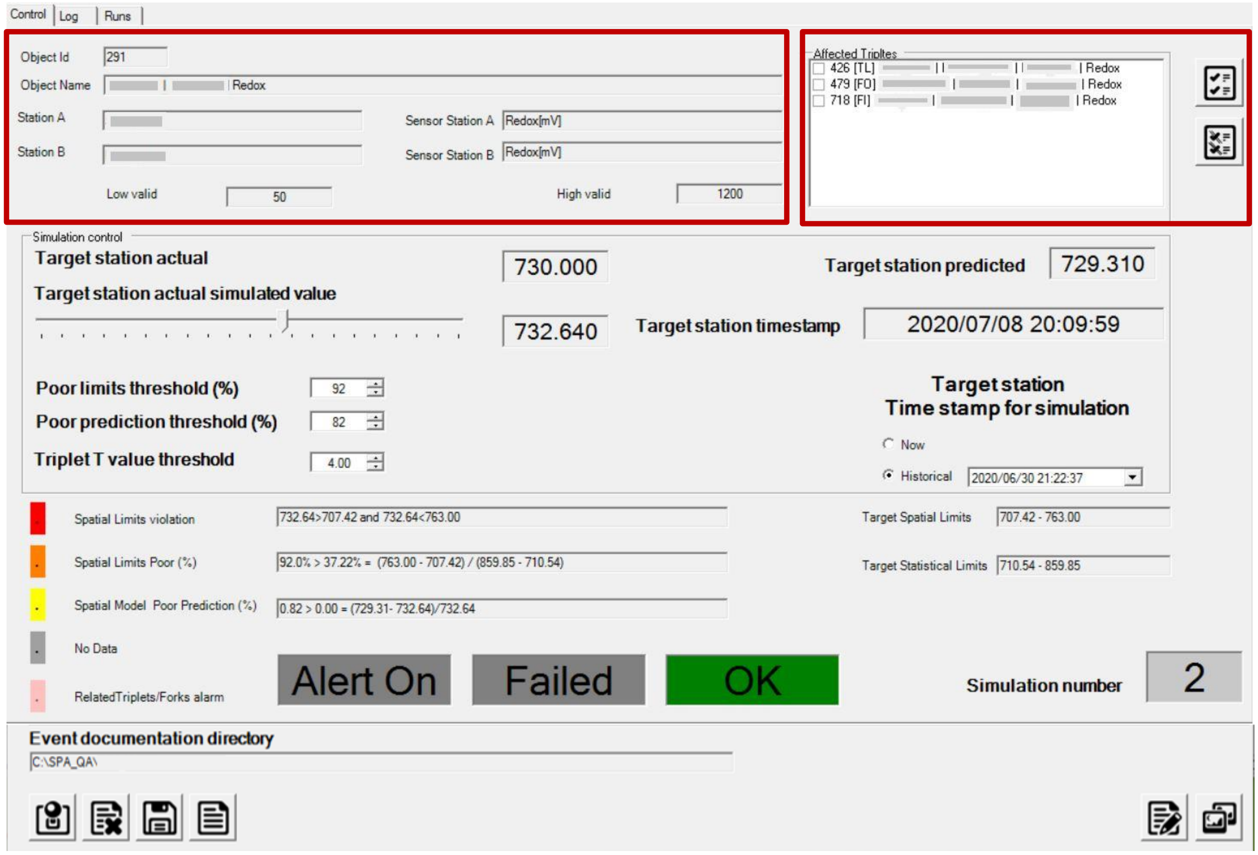
C:\SPA_QA\

Figure 127: The simulations screen

The upper left set of fields in the Simulations screen, marked by the left red rectangle in Figure 128, shows a description of the current selected Pair model for simulation. The different fields show the Object ID (unique identifier for each object), the object name, and the two stations lined to the object – the source and the target station. The type of measurement (and measurement units) for the source and target stations is depicted as well. The Low/High valid fields show the range of valid measurements for the two stations, as configured by the user in the Model Editor window, as in Section 4.3.

The selection pane and selection buttons on the upper right side of the Simulations screen, marked by the right red rectangle in Figure 128, list all triplet models the currently viewed pair model is part of. When simulating events for a pair model, events triggered by triplet models including the simulated pair will also be listed, if the corresponding triplet models are selected in the "Affected Triplets" selection pane.

On the next page, we discuss how to configure the simulated events.



Object Id: 291
Object Name: [redacted] | [redacted] | Redox
Station A: [redacted]
Sensor Station A: Redox[mV]
Station B: [redacted]
Sensor Station B: Redox[mV]
Low valid: 50
High valid: 1200

Affected Tripltes
☐ 426 [TL] [redacted] | | [redacted] | | [redacted] | Redox
☐ 479 [FO] [redacted] | [redacted] | [redacted] | Redox
☐ 718 [FI] [redacted] | [redacted] | [redacted] | Redox

Simulation control
Target station actual: 730.000
Target station predicted: 729.310
Target station actual simulated value: 732.640
Target station timestamp: 2020/07/08 20:09:59
Poor limits threshold (%): 92
Poor prediction threshold (%): 82
Triplet T value threshold: 4.00

Target station
Time stamp for simulation
○ Now
● Historical: 2020/06/30 21:22:37

🟥 Spatial Limits violation: 732.64>707.42 and 732.64<763.00
🟧 Spatial Limits Poor (%): 92.0% > 37.22% = (763.00 - 707.42) / (859.85 - 710.54)
🟨 Spatial Model Poor Prediction (%): 0.82 > 0.00 = (729.31- 732.64)/732.64
⬛ No Data
▨ RelatedTriplets/Forks alarm

Target Spatial Limits: 707.42 - 763.00
Target Statistical Limits: 710.54 - 859.85

Simulation number: 2

Event documentation directory: C:\SPA_QA\

Figure 128: Object identification fields and list of affected triplets in the Simulations screen.

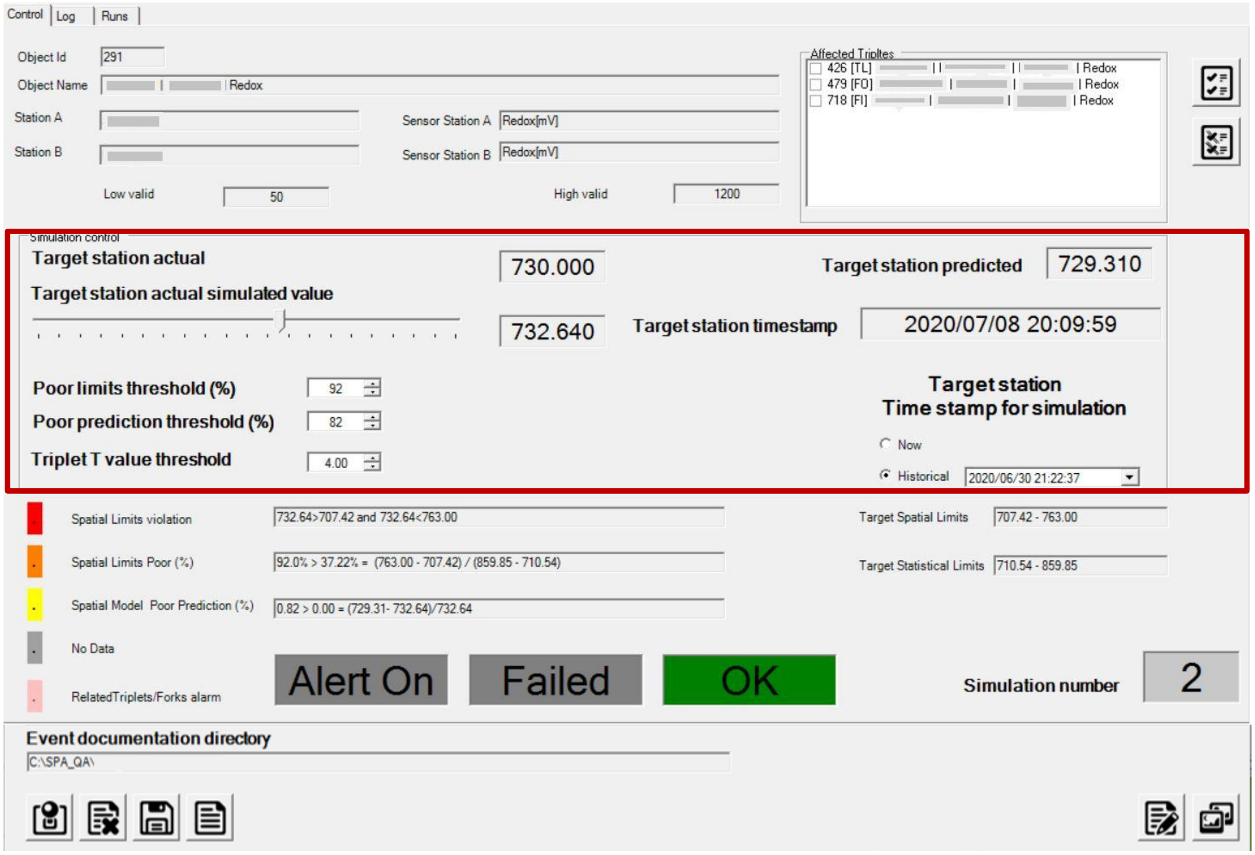
Figure 129 highlights the different simulation controls, along with additional fields informing the user on the models performance.

The controls and available to the user for configuring the simulated event are:

The fields shown to the user are:

- Info field: The current timestamp measured by the target station.
- Control field: The target station timestamp for simulation – either for the current time or for a historic timestamp.
- Info field: The actual value measured in the target station, for the timestamp selected for simulation.
- Control field: The value to be simulated for the target station at the chosen timestamp.
- The predicted value for the target station, computed after running the simulation (see remark below).
- Controls for setting the thresholds for the Poor limits, Poor prediction and Triplet alarms, as defined at the start of Unit C.

The user may run the simulation by setting the above control fields and clicking the "simulate" button, shown as a lever () at the bottom of Figure 129.



Object Id 291

Object Name [redacted] | [redacted] | Redox

Station A [redacted]

Sensor Station A Redox[mV]

Station B [redacted]

Sensor Station B Redox[mV]

Low valid 50

High valid 1200

Affected Tripltes
- □ 426 [TL] [redacted] | | [redacted] | | [redacted] | Redox
- □ 479 [FO] [redacted] | [redacted] | [redacted] | Redox
- □ 718 [FI] [redacted] | [redacted] | [redacted] | Redox

Simulation control

Target station actual 730.000

Target station actual simulated value 732.640

Target station predicted 729.310

Target station timestamp 2020/07/08 20:09:59

Poor limits threshold (%) 92

Poor prediction threshold (%) 82

Triplet T value threshold 4.00

Target station
Time stamp for simulation

○ Now

● Historical 2020/06/30 21:22:37

🟥 Spatial Limits violation 732.64>707.42 and 732.64<763.00

🟧 Spatial Limits Poor (%) 92.0% > 37.22% = (763.00 - 707.42) / (859.85 - 710.54)

🟨 Spatial Model Poor Prediction (%) 0.82 > 0.00 = (729.31- 732.64)/732.64

⬛ No Data

🟥 RelatedTriplets/Forks alarm

Target Spatial Limits 707.42 - 763.00

Target Statistical Limits 710.54 - 859.85

Simulation number 2

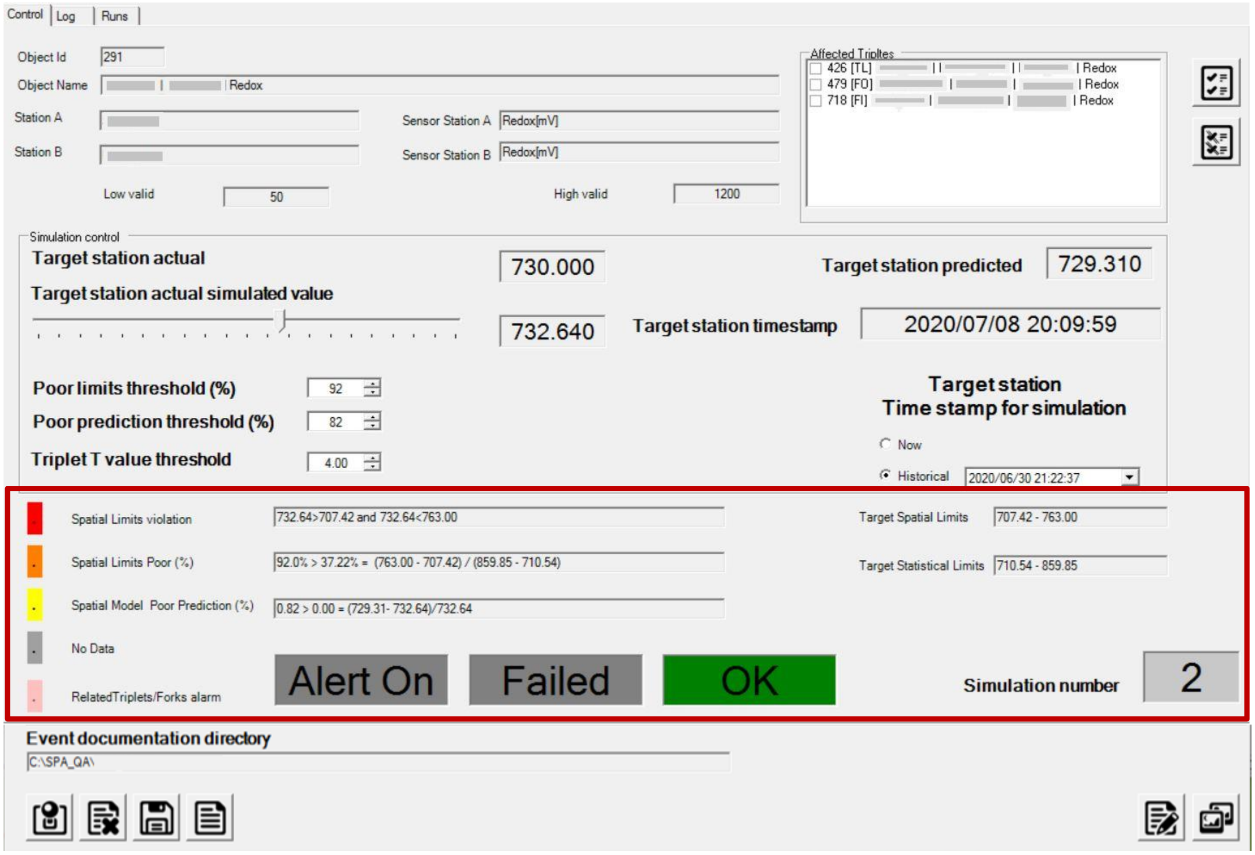
Event documentation directory
C:\SPA_QA\

Figure 129: Simulation control fields and alarm thresholds.

After running the simulated event, the fields highlighted by a red rectangle at the bottom of Figure 130 will show the simulated limits as well as alarm indicators. The following outputs are available:

- Spatial limits and statistical limits for the target measurement at the target node.
- A field describing the simulated measurement, with respect to the computed spatial limits. If the spatial limits are violated, the label will be marked in red.
- A field showing the ratio between the spatial limits range and the statistical limits range (along with the computation method), and stating if this ratio is below the "Spatial Limits Poor" limit. If the ratio is above the alarm limit, an alarm will be triggered and the label will turn red.
- A field showing the ratio between the prediction error and the actual measurement. If this ratio is above the "Spatial Model Poor Prediction" alarm threshold, an alarm will be triggered, and the field name will turn red.
- A label for a no data alarm- if no data was available for the source station.
- A label for a triplets alarm – if one of the monitored triplet models (selected in the Affected Triplets pane) raised in alarm.
- Colored fields for no alarm (OK, in green), a failed prediction or a triggered alarm.
- A counter for the number of simulations run.

The log file for the current simulation can be viewed under the "Log" tab. Previous simulation runs and their results can be viewed under the "Runs" tab.
Sections 4.10.1-4.10.5 exemplify possible outputs generated by simulated events.



Object Id: 291
Object Name: [redacted] | [redacted] | Redox
Station A: [redacted]
Sensor Station A: Redox[mV]
Station B: [redacted]
Sensor Station B: Redox[mV]
Low valid: 50
High valid: 1200

Affected Triples
□ 426 [TL] [redacted] | | [redacted] | | [redacted] | Redox
□ 479 [FO] [redacted] | [redacted] | [redacted] | Redox
□ 718 [FI] [redacted] | [redacted] | [redacted] | Redox

Simulation control
Target station actual: 730.000
Target station predicted: 729.310
Target station actual simulated value: 732.640
Target station timestamp: 2020/07/08 20:09:59
Poor limits threshold (%): 92
Poor prediction threshold (%): 82
Triplet T value threshold: 4.00

Target station
Time stamp for simulation
○ Now
● Historical 2020/06/30 21:22:37

🟥 Spatial Limits violation: 732.64>707.42 and 732.64<763.00
🟧 Spatial Limits Poor (%): 92.0% > 37.22% = (763.00 - 707.42) / (859.85 - 710.54)
🟨 Spatial Model Poor Prediction (%): 0.82 > 0.00 = (729.31- 732.64)/732.64
⬛ No Data
▨ RelatedTriplets/Forks alarm

Target Spatial Limits: 707.42 - 763.00
Target Statistical Limits: 710.54 - 859.85

Simulation number: 2

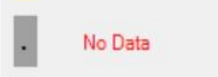
Event documentation directory
C:\SPA_QA\

Figure 130: Indicators for simulation results

- reset
- export def
- to html
- printscreen

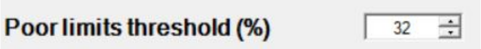
4.10.1 Simulating a No Data Alarm

- In case no data is available for the source station, the following indicator will appear:

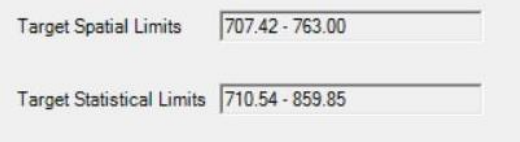

No Data

4.10.2 Simulating a Spatial Limits Poor Alarm

- Consider the case the poor limits alarm is set to 32%.

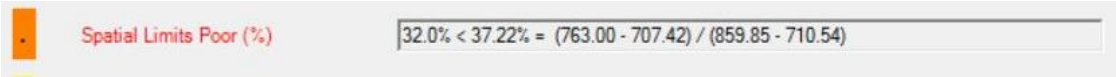

Poor limits threshold (%) 32

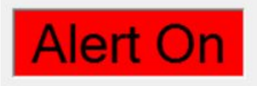
- For our simulation, the generated spatial and statistical limits are the following:


Target Spatial Limits 707.42 - 763.00

Target Statistical Limits 710.54 - 859.85

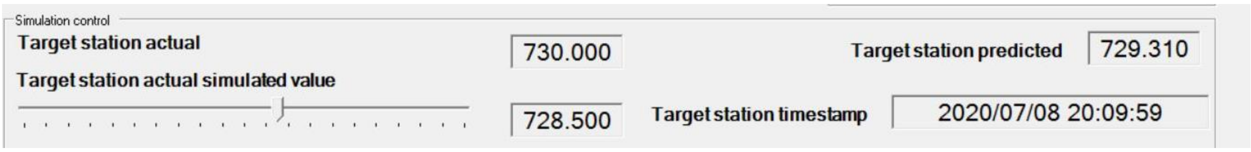
- The ratio between the two ranges is 37.22% (see computation below). Hence, an alarm is triggered:


Spatial Limits Poor (%) 32.0% < 37.22% = (763.00 - 707.42) / (859.85 - 710.54)

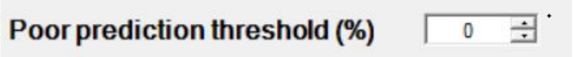

Alert On

4.10.3 Simulating a Spatial Limits Poor Prediction Alarm

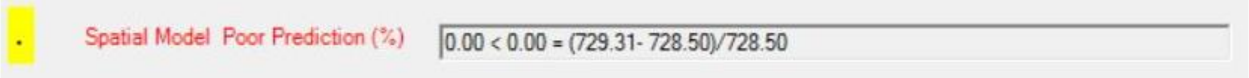
- Consider a simulated setting with the following actual and predicted values:


Simulation control
Target station actual 730.000
Target station predicted 729.310
Target station actual simulated value 728.500
Target station timestamp 2020/07/08 20:09:59

- We set the poor prediction error threshold to 0 %.

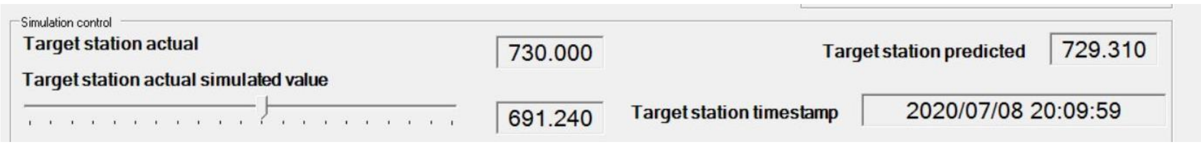

Poor prediction threshold (%) 0

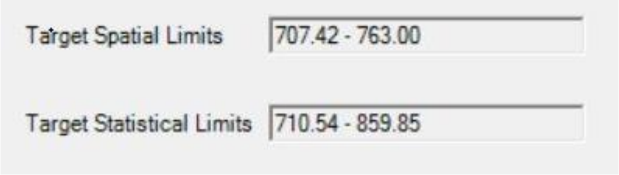
- The relative error, given by $(729.31 - 728.50)/728.50$, is larger than zero (by a small amount, but still). Hence, an alarm is triggered:


Spatial Model Poor Prediction (%) 0.00 < 0.00 = (729.31- 728.50)/728.50

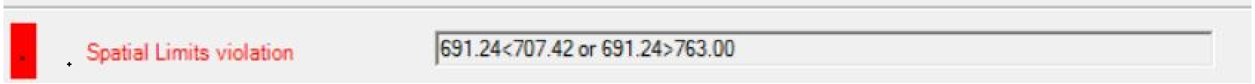
4.10.4 Simulating a Spatial Limit Violation Alarm

- Consider a simulated setting with the following actual and predicted values, and the following spatial limits:


Simulation control
Target station actual 730.000
Target station predicted 729.310
Target station actual simulated value 691.240
Target station timestamp 2020/07/08 20:09:59

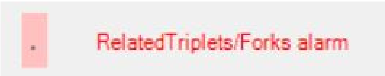

Target Spatial Limits 707.42 - 763.00
Target Statistical Limits 710.54 - 859.85

- The simulated "actual" measurement is found outside of the spatial limits, and therefore a "Spatial Limits violation" alarm is raised:


Spatial Limits violation 691.24<707.42 or 691.24>763.00

4.10.5 Simulating a Related Triplets Forks Alarms

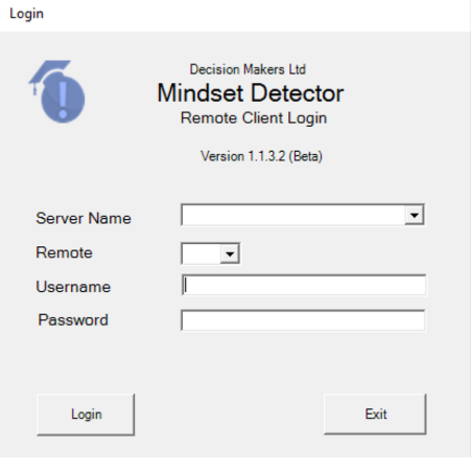
- In case alarms will be raised by any of the monitored Triplet models, the following field will turn red:


RelatedTriplets/Forks alarm

4.11 Viewing Spatial Model Predictions and Alarms in the Thin Client

This section discusses how the SM, its model prediction and alarms can be viewed in the EDS thin client, a desktop based application that can be used for monitoring the water network, without accessing the EDS server software directly.

Log-in: Start the EDS thin client using the desktop icon. Figure 131 shows the log-in screen. Type in the server name, username and password as supplied by Decision Makers LTD. For the remote option, click 'yes', if you are connecting to a remote server, and false otherwise. Click log-in.



Login

Decision Makers Ltd

Mindset Detector

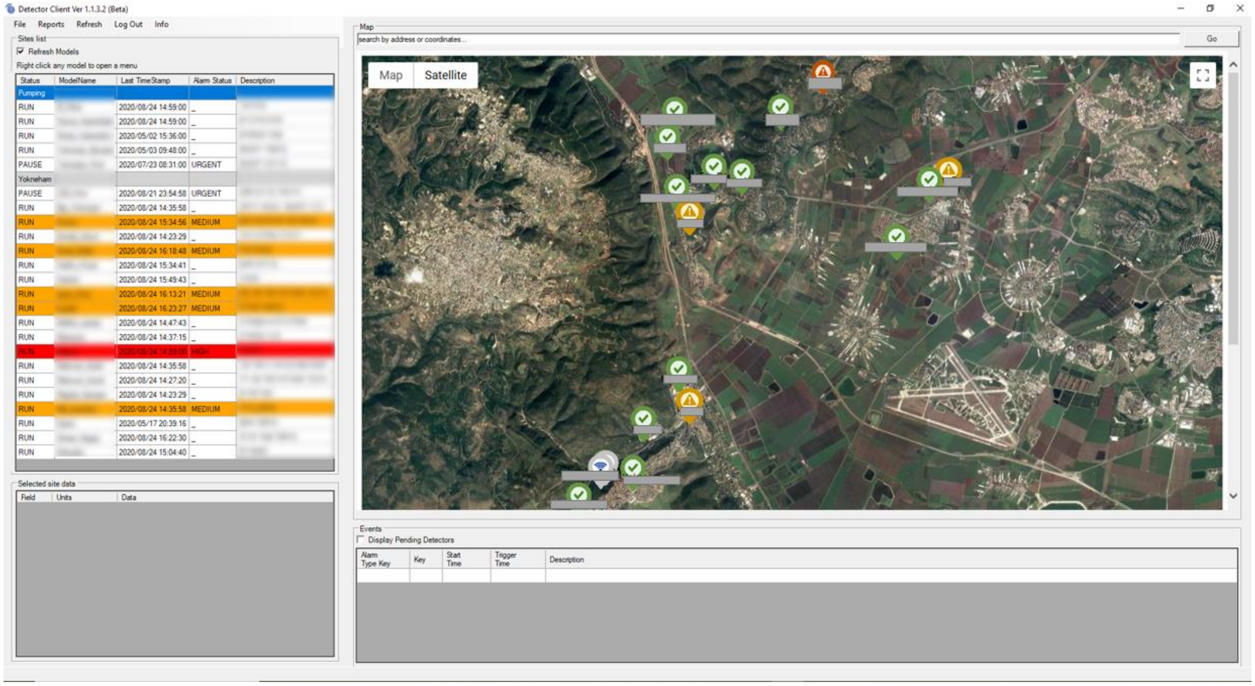
Remote Client Login

Version 1.1.3.2 (Beta)

Server Name	
Remote	
Username	
Password	

Figure 131: EDS thin client log-in screen.

The main screen and map: Figure 132 shows the main screen for the Thin Client. The map at the center of the screen shows the different station locations, with icons showing their type and status. A list of stations on the left shows the current nodes in the network, their status, alarms and description. The last timestamp monitored appears as well.



Detector Client Ver 1.1.3.2 (Beta)

Sites list

☑ Refresh Models

Right click any model to open a menu

Status	ModelName	Last TimeStamp	Alarm Status	Description
Pumping				
RUN	[redacted]	2020/08/24 14:59:00	_	[redacted]
RUN	[redacted]	2020/08/24 14:59:00	_	[redacted]
RUN	[redacted]	2020/05/02 15:36:00	_	[redacted]
RUN	[redacted]	2020/05/03 09:48:00	_	[redacted]
PAUSE	[redacted]	2020/07/23 08:31:00	URGENT	[redacted]
Yokneham				
PAUSE	[redacted]	2020/08/21 23:54:58	URGENT	[redacted]
RUN	[redacted]	2020/08/24 14:35:58	_	[redacted]
RUN	[redacted]	2020/08/24 15:34:56	MEDIUM	[redacted]
RUN	[redacted]	2020/08/24 14:23:29	_	[redacted]
RUN	[redacted]	2020/08/24 16:18:48	MEDIUM	[redacted]
RUN	[redacted]	2020/08/24 15:34:41	_	[redacted]
RUN	[redacted]	2020/08/24 15:49:43	_	[redacted]
RUN	[redacted]	2020/08/24 16:13:21	MEDIUM	[redacted]
RUN	[redacted]	2020/08/24 16:23:27	MEDIUM	[redacted]
RUN	[redacted]	2020/08/24 14:47:43	_	[redacted]
RUN	[redacted]	2020/08/24 14:37:15	_	[redacted]
RUN	[redacted]	2020/08/24 14:59:00	HIGH	[redacted]
RUN	[redacted]	2020/08/24 14:35:58	_	[redacted]
RUN	[redacted]	2020/08/24 14:27:20	_	[redacted]
RUN	[redacted]	2020/08/24 14:23:29	_	[redacted]
RUN	[redacted]	2020/08/24 14:35:58	MEDIUM	[redacted]
RUN	[redacted]	2020/05/17 20:39:16	_	[redacted]
RUN	[redacted]	2020/08/24 16:22:30	_	[redacted]
RUN	[redacted]	2020/08/24 15:04:40	_	[redacted]

Selected site data

Field	Units	Data

Map

search by address or coordinates...

Events

☐ Display Pending Detectors

Alarm Type Key	Key	Start Time	Trigger Time	Description

Figure 132:EDS thin client main screen and map.

Pressing the legend icon will show a window as in Figure 144. Positioning the mouse pointer will show a tool tip text. See Figure description for a description of icons.

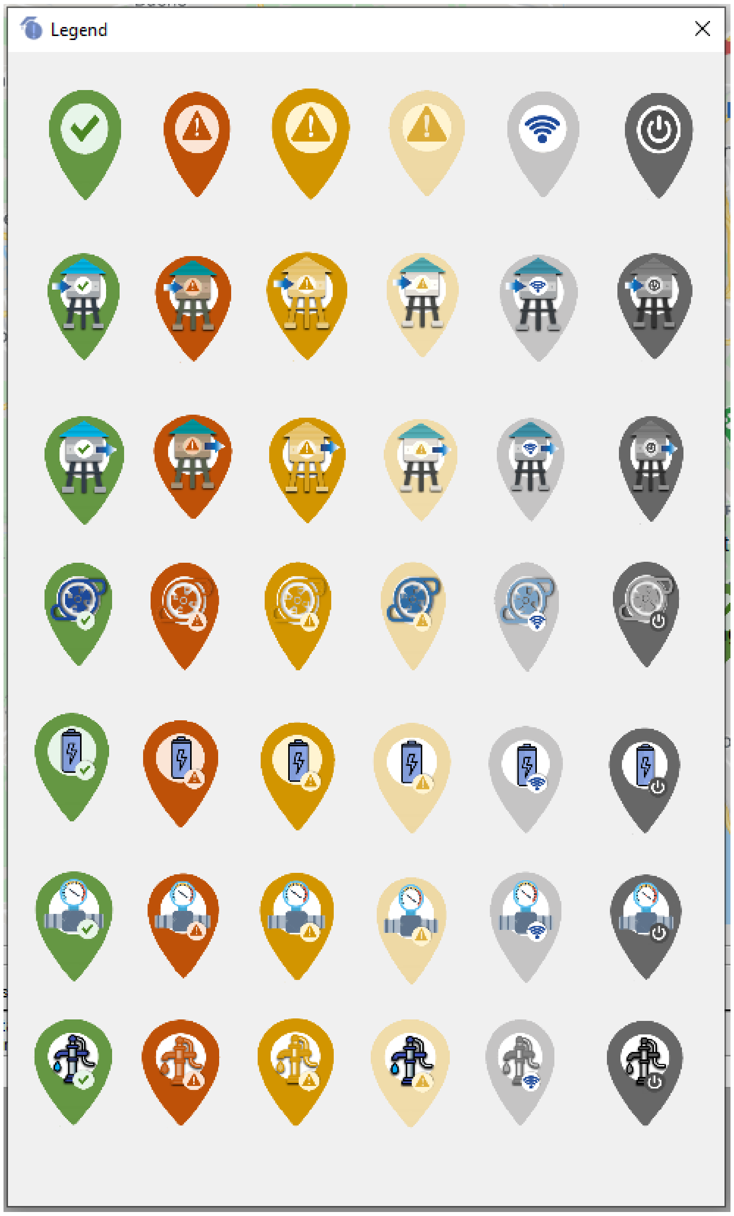
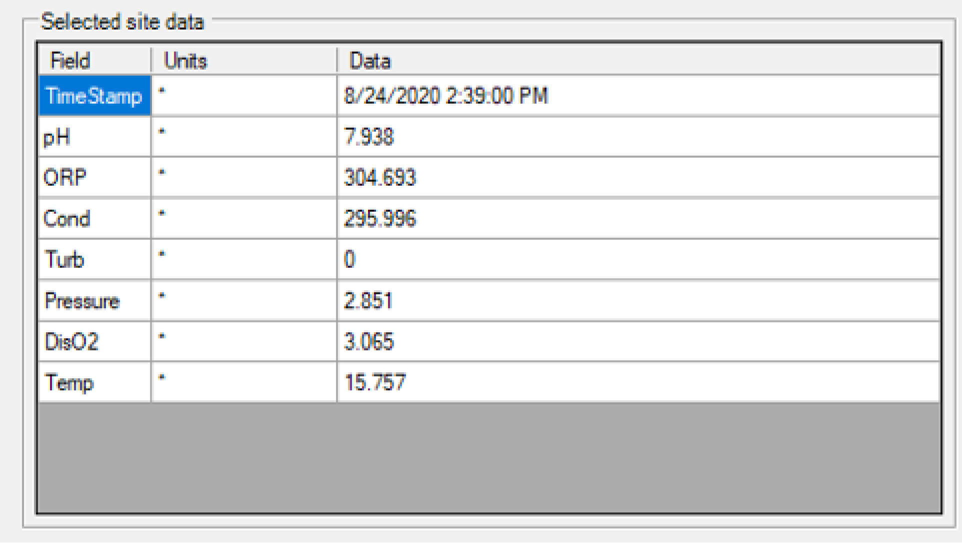


Figure 133:Legend for EDS thing client map icon. Object types, from top to button: General station, reservoir in, reservoir out, pump, battery-powered station, pressure reducer, general water exit from network. States are color coded, from left to right: status OK, High severity alarm, medium severity alarm, low severity alarm, no communication and model disabled (not monitored).

Site data: Clicking on a site will present it's current data under the "Selected site data" panel.

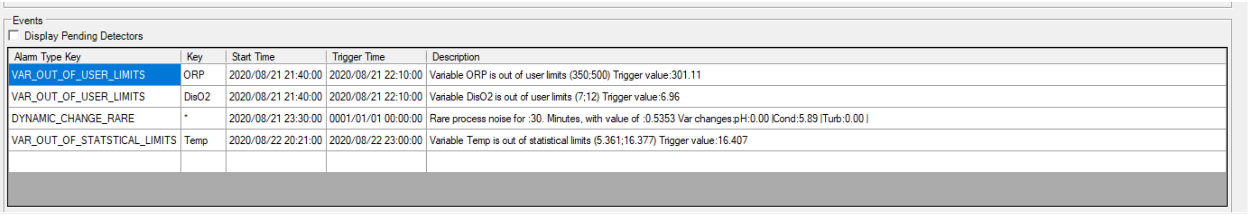


Selected site data

Field	Units	Data
TimeStamp	*	8/24/2020 2:39:00 PM
pH	*	7.938
ORP	*	304.693
Cond	*	295.996
Turb	*	0
Pressure	*	2.851
DisO2	*	3.065
Temp	*	15.757

Figure 134: EDS thin client "selected site data" panel

Site events: Clicking on a site will present it's current events under the "events" panel.



Events

□ Display Pending Detectors

Alarm Type Key	Key	Start Time	Trigger Time	Description
VAR_OUT_OF_USER_LIMITS	ORP	2020/08/21 21:40:00	2020/08/21 22:10:00	Variable ORP is out of user limits (350;500) Trigger value:301.11
VAR_OUT_OF_USER_LIMITS	DisO2	2020/08/21 21:40:00	2020/08/21 22:10:00	Variable DisO2 is out of user limits (7;12) Trigger value:6.96
DYNAMIC_CHANGE_RARE	*	2020/08/21 23:30:00	0001/01/01 00:00:00	Rare process noise for :30. Minutes, with value of :0.5353 Var changes:pH:0.00 |Cond:5.89 |Turb:0.00 |
VAR_OUT_OF_STATSTICAL_LIMITS	Temp	2020/08/22 20:21:00	2020/08/22 23:00:00	Variable Temp is out of statistical limits (5.361;16.377) Trigger value:16.407

Figure 135: EDS thin client "selected site events" panel

File Menu: Clicking on the menu titled "file" will show the options depicted in Figure HHH.

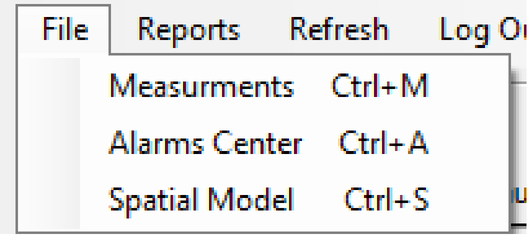
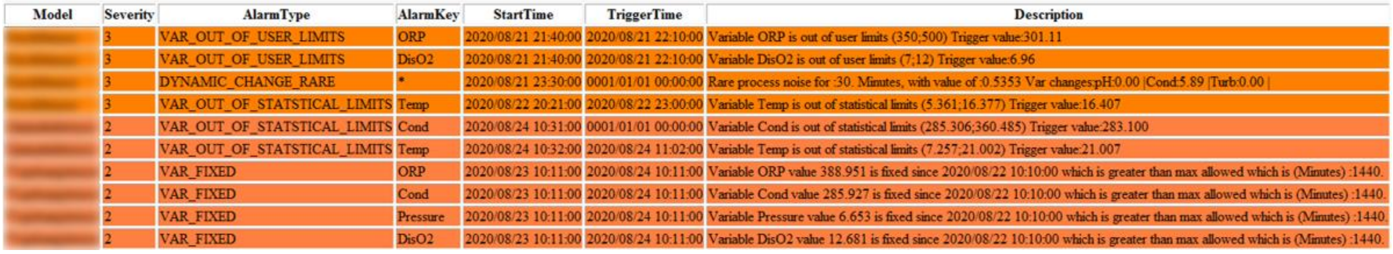


Figure 136: EDS thin client file menu

Alarms Center: By Selecting the alarms center, a window as in Figure 137 will appear, showing all available alarms.



Model	Severity	AlarmType	AlarmKey	StartTime	TriggerTime	Description
[redacted]	3	VAR_OUT_OF_USER_LIMITS	ORP	2020/08/21 21:40:00	2020/08/21 22:10:00	Variable ORP is out of user limits (350;500) Trigger value:301.11
[redacted]	3	VAR_OUT_OF_USER_LIMITS	DisO2	2020/08/21 21:40:00	2020/08/21 22:10:00	Variable DisO2 is out of user limits (7;12) Trigger value:6.96
[redacted]	3	DYNAMIC_CHANGE_RARE	*	2020/08/21 23:30:00	0001/01/01 00:00:00	Rare process noise for :30. Minutes, with value of :0.5353 Var changes:pH:0.00 |Cond:5.89 |Turb:0.00 |
[redacted]	3	VAR_OUT_OF_STATSTICAL_LIMITS	Temp	2020/08/22 20:21:00	2020/08/22 23:00:00	Variable Temp is out of statistical limits (5.361;16.377) Trigger value:16.407
[redacted]	2	VAR_OUT_OF_STATSTICAL_LIMITS	Cond	2020/08/24 10:31:00	0001/01/01 00:00:00	Variable Cond is out of statistical limits (285.306;360.485) Trigger value:283.100
[redacted]	2	VAR_OUT_OF_STATSTICAL_LIMITS	Temp	2020/08/24 10:32:00	2020/08/24 11:02:00	Variable Temp is out of statistical limits (7.257;21.002) Trigger value:21.007
[redacted]	2	VAR_FIXED	ORP	2020/08/23 10:11:00	2020/08/24 10:11:00	Variable ORP value 388.951 is fixed since 2020/08/22 10:10:00 which is greater than max allowed which is (Minutes) :1440.
[redacted]	2	VAR_FIXED	Cond	2020/08/23 10:11:00	2020/08/24 10:11:00	Variable Cond value 285.927 is fixed since 2020/08/22 10:10:00 which is greater than max allowed which is (Minutes) :1440.
[redacted]	2	VAR_FIXED	Pressure	2020/08/23 10:11:00	2020/08/24 10:11:00	Variable Pressure value 6.653 is fixed since 2020/08/22 10:10:00 which is greater than max allowed which is (Minutes) :1440.
[redacted]	2	VAR_FIXED	DisO2	2020/08/23 10:11:00	2020/08/24 10:11:00	Variable DisO2 value 12.681 is fixed since 2020/08/22 10:10:00 which is greater than max allowed which is (Minutes) :1440.

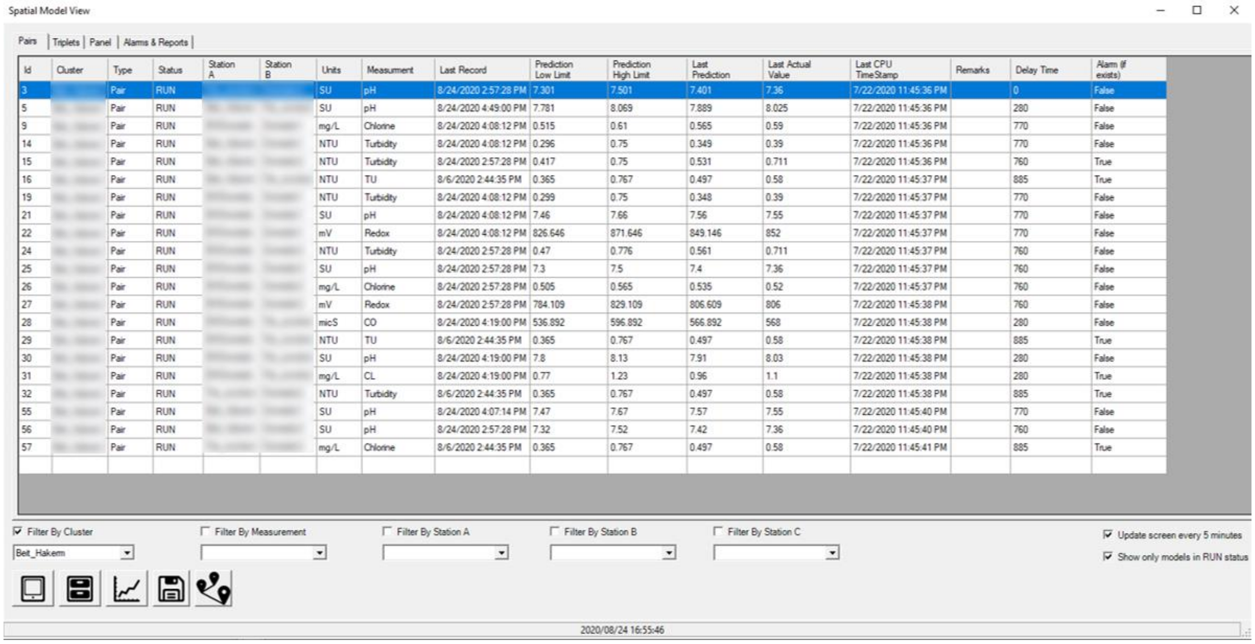
Figure 137: EDS thin client Alarms Center.

Clicking on the "Spatial model" item under the "File" menu will show the "Spatial Model View" window, discussed in the next subsection.

4.11.1 The Spatial Model View in the EDS Thin Client

This section covers the Spatial Model view in the EDS thin client. This view is accessible through the "Spatial Model" option, under the "File" menu in the EDS thin client. Figure 138 shows the Spatial Model View, with it's tab, selection drop down boxes and toolbar. In essence the Spatial Model View is a simpler version of the Spatial Model manager presented in Section 4.1. The main change between the manager and viewer is that models and definitions cannot be changed from the viewer.

The pairs tab is shown in Figure 138, with the pair descriptive field and state indicators shown in Figure 76 in Section 4.1



Id	Cluster	Type	Status	Station A	Station B	Units	Measurment	Last Record	Prediction Low Limit	Prediction High Limit	Last Prediction	Last Actual Value	Last CPU TimeStamp	Remarks	Delay Time	Alarm (if exists)
3	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 2:57:28 PM	7.301	7.501	7.401	7.36	7/22/2020 11:45:36 PM		0	False
5	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 4:49:00 PM	7.781	8.069	7.889	8.025	7/22/2020 11:45:36 PM		280	False
9	[redacted]	Pair	RUN	[redacted]	[redacted]	mg/L	Chlorine	8/24/2020 4:08:12 PM	0.515	0.61	0.565	0.59	7/22/2020 11:45:36 PM		770	False
14	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	Turbidity	8/24/2020 4:08:12 PM	0.296	0.75	0.349	0.39	7/22/2020 11:45:36 PM		770	False
15	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	Turbidity	8/24/2020 2:57:28 PM	0.417	0.75	0.531	0.711	7/22/2020 11:45:36 PM		760	True
16	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	TU	8/6/2020 2:44:35 PM	0.365	0.767	0.497	0.58	7/22/2020 11:45:37 PM		885	True
19	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	Turbidity	8/24/2020 4:08:12 PM	0.299	0.75	0.348	0.39	7/22/2020 11:45:37 PM		770	False
21	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 4:08:12 PM	7.46	7.66	7.56	7.55	7/22/2020 11:45:37 PM		770	False
22	[redacted]	Pair	RUN	[redacted]	[redacted]	mV	Redox	8/24/2020 4:08:12 PM	826.646	871.646	849.146	852	7/22/2020 11:45:37 PM		770	False
24	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	Turbidity	8/24/2020 2:57:28 PM	0.47	0.776	0.561	0.711	7/22/2020 11:45:37 PM		760	False
25	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 2:57:28 PM	7.3	7.5	7.4	7.36	7/22/2020 11:45:37 PM		760	False
26	[redacted]	Pair	RUN	[redacted]	[redacted]	mg/L	Chlorine	8/24/2020 2:57:28 PM	0.505	0.565	0.535	0.52	7/22/2020 11:45:37 PM		760	False
27	[redacted]	Pair	RUN	[redacted]	[redacted]	mV	Redox	8/24/2020 2:57:28 PM	784.109	829.109	806.609	806	7/22/2020 11:45:38 PM		760	False
28	[redacted]	Pair	RUN	[redacted]	[redacted]	micS	CO	8/24/2020 4:19:00 PM	536.892	596.892	566.892	568	7/22/2020 11:45:38 PM		280	False
29	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	TU	8/6/2020 2:44:35 PM	0.365	0.767	0.497	0.58	7/22/2020 11:45:38 PM		885	True
30	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 4:19:00 PM	7.8	8.13	7.91	8.03	7/22/2020 11:45:38 PM		280	False
31	[redacted]	Pair	RUN	[redacted]	[redacted]	mg/L	CL	8/24/2020 4:19:00 PM	0.77	1.23	0.96	1.1	7/22/2020 11:45:38 PM		280	True
32	[redacted]	Pair	RUN	[redacted]	[redacted]	NTU	Turbidity	8/6/2020 2:44:35 PM	0.365	0.767	0.497	0.58	7/22/2020 11:45:38 PM		885	True
55	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 4:07:14 PM	7.47	7.67	7.57	7.55	7/22/2020 11:45:40 PM		770	False
56	[redacted]	Pair	RUN	[redacted]	[redacted]	SU	pH	8/24/2020 2:57:28 PM	7.32	7.52	7.42	7.36	7/22/2020 11:45:40 PM		760	False
57	[redacted]	Pair	RUN	[redacted]	[redacted]	mg/L	Chlorine	8/6/2020 2:44:35 PM	0.365	0.767	0.497	0.58	7/22/2020 11:45:41 PM		885	True

☑ Filter By Cluster: Beit_Hakem
☐ Filter By Measurement
☐ Filter By Station A
☐ Filter By Station B
☐ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

2020/08/24 16:55:46

Figure 138: Pairs tab in Spatial Model Viewer in the

The Spatial Model Viewer toolbar contains the following options:

- See model definitions
- Refresh view
- See charts
- Export table to file
- See selected SM object on map

Focus on object in map: By clicking on an SM object and then clicking on the "see select SM object on map icon" (), the map in the EDS thin client will focus on the selected object. See Figure 139 for an example of a map focused on an object (A pair object with ID 45).

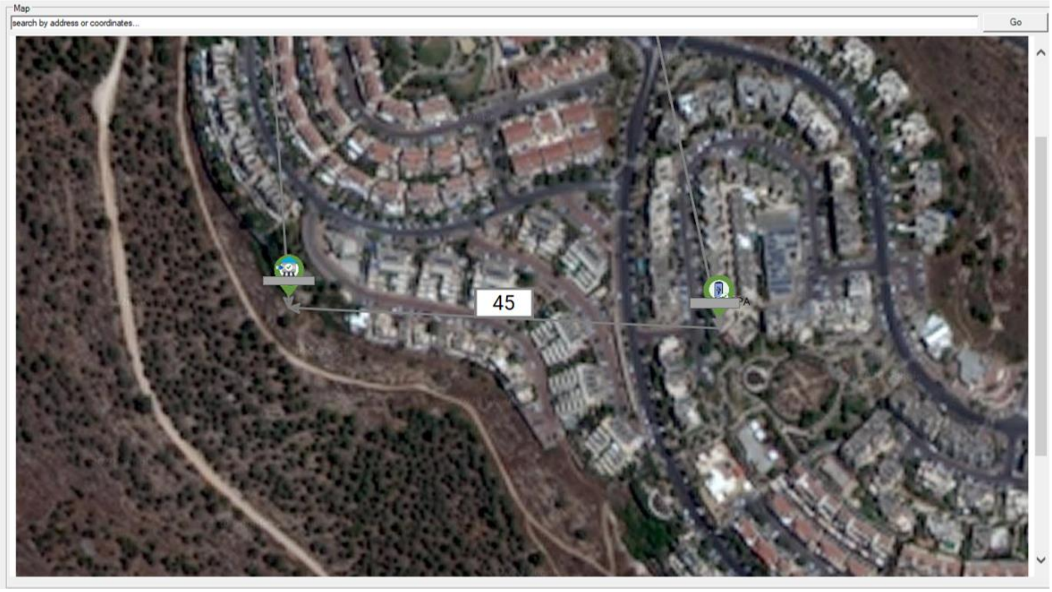
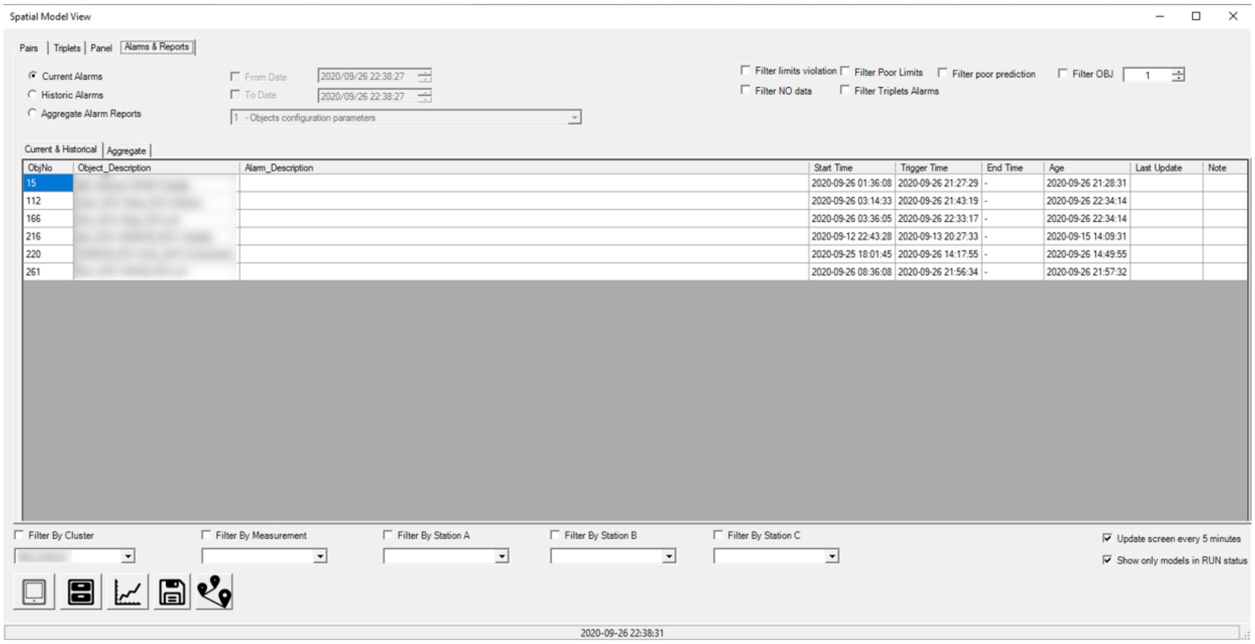


Figure 139: An example for a map focused on an object after clicking on the "map focus" icon.

Reports and alarms: Current alarms, historical alarms and reports can be viewed through the Alarms and reports panel, shown in Figure 140. This panel is similar to the Alarms and Reports panel shown on Figure 81 in Section 4.1.



● Current Alarms
○ Historic Alarms
○ Aggregate Alarm Reports

□ From Date 2020/09/26 22:38:27
□ To Date 2020/09/26 22:38:27
1 - Objects configuration parameters

□ Filter limits violation □ Filter Poor Limits □ Filter poor prediction □ Filter OBJ 1
□ Filter NO data □ Filter Triplets Alarms

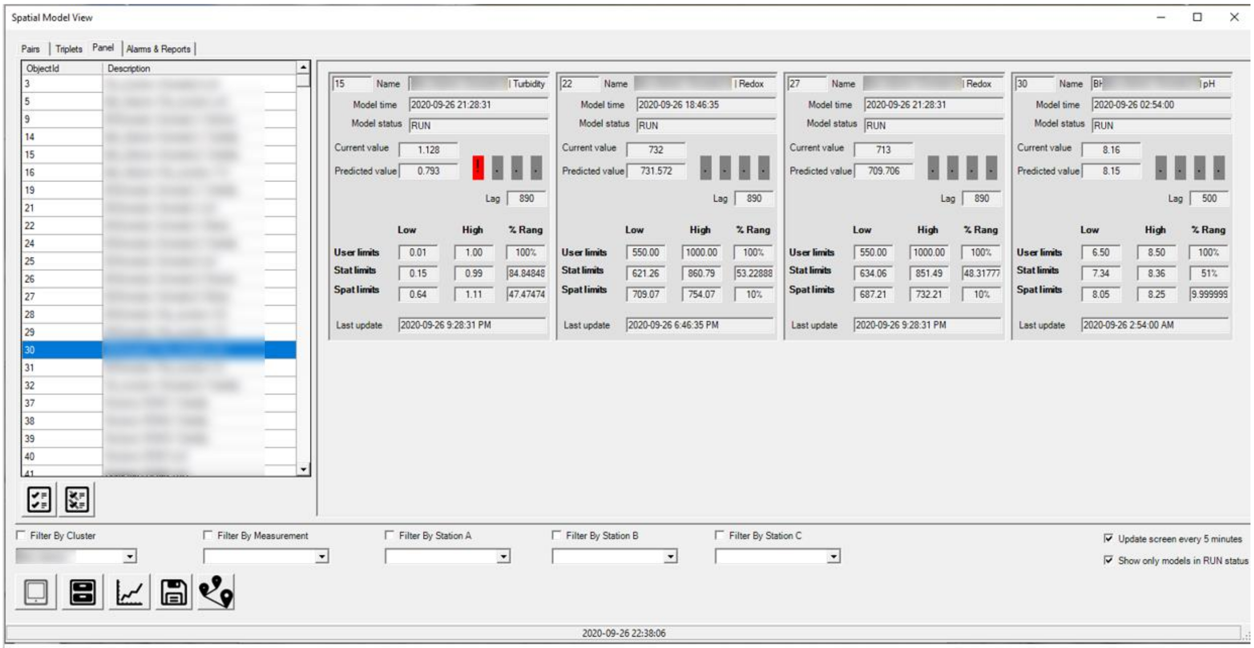
ObjNo	Object_Description	Alarm_Description	Start Time	Trigger Time	End Time	Age	Last Update	Note
15	[redacted]		2020-09-26 01:36:08	2020-09-26 21:27:29	-	2020-09-26 21:28:31		
112	[redacted]		2020-09-26 03:14:33	2020-09-26 21:43:19	-	2020-09-26 22:34:14		
166	[redacted]		2020-09-26 03:36:05	2020-09-26 22:33:17	-	2020-09-26 22:34:14		
216	[redacted]		2020-09-12 22:43:28	2020-09-13 20:27:33	-	2020-09-15 14:09:31		
220	[redacted]		2020-09-25 18:01:45	2020-09-26 14:17:55	-	2020-09-26 14:49:55		
261	[redacted]		2020-09-26 08:36:08	2020-09-26 21:56:34	-	2020-09-26 21:57:32		

□ Filter By Cluster [redacted]
□ Filter By Measurement
□ Filter By Station A
□ Filter By Station B
□ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

2020-09-26 22:38:31

Figure 140:Alarms and reports panel in the EDS thin client

Panel with key indicators: A panel showing the key indicators for each pair is available as well, see Figure 141.


Spatial Model View

Pairs | Triplets | Panel | Alarms & Reports

ObjectId	Description
3	[redacted]
5	[redacted]
9	[redacted]
14	[redacted]
15	[redacted]
16	[redacted]
19	[redacted]
21	[redacted]
22	[redacted]
24	[redacted]
25	[redacted]
26	[redacted]
27	[redacted]
28	[redacted]
29	[redacted]
30	[redacted]
31	[redacted]
32	[redacted]
37	[redacted]
38	[redacted]
39	[redacted]
40	[redacted]
41	

	15	22	27	30
Name	[redacted] | Turbidity	[redacted] | Redox	[redacted] | Redox	BH[redacted] | pH
Model time	2020-09-26 21:28:31	2020-09-26 18:46:35	2020-09-26 21:28:31	2020-09-26 02:54:00
Model status	RUN	RUN	RUN	RUN
Current value	1.128	732	713	8.16
Predicted value	0.793	731.572	709.706	8.15
Lag	890	890	890	500

15	Low	High	% Rang
User limits	0.01	1.00	100%
Stat limits	0.15	0.99	84.84848
Spat limits	0.64	1.11	47.47474

Last update 2020-09-26 9:28:31 PM

22	Low	High	% Rang
User limits	550.00	1000.00	100%
Stat limits	621.26	860.79	53.22888
Spat limits	709.07	754.07	10%

Last update 2020-09-26 6:46:35 PM

27	Low	High	% Rang
User limits	550.00	1000.00	100%
Stat limits	634.06	851.49	48.31777
Spat limits	687.21	732.21	10%

Last update 2020-09-26 9:28:31 PM

30	Low	High	% Rang
User limits	6.50	8.50	100%
Stat limits	7.34	8.36	51%
Spat limits	8.05	8.25	9.999999

Last update 2020-09-26 2:54:00 AM

□ Filter By Cluster □ Filter By Measurement □ Filter By Station A □ Filter By Station B □ Filter By Station C

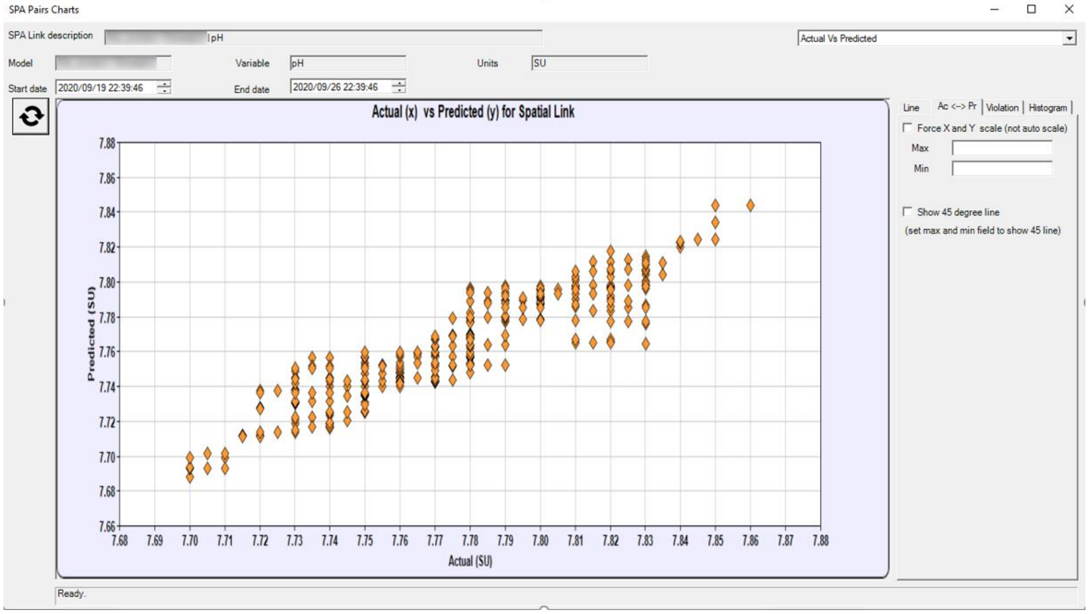
☑ Update screen every 5 minutes

☑ Show only models in RUN status

2020-09-26 22:38:06

Figure 141: A panel showing key indicators for each pair model, in the EDS thin client.

Graphs: Time series graph, error histograms, Source vs Target and Actual vs. Predicted graphs are available for each pair model, by pressing the charts button (📈). Figure 132 shows an example for an Actual vs. Predicted graph for a Pair model in the EDS thin client. Figures 143 and 144 show a histogram for spatial model event lengths, and a time series comparing predictions, measurements, and alarms thresholds, respectively.


SPA Pairs Charts

SPA Link description [redacted] | pH

Model [redacted]

Variable pH

Units SU

Start date 2020/09/19 22:39:46

End date 2020/09/26 22:39:46

Actual Vs Predicted

□ Force X and Y scale (not auto scale)

Max

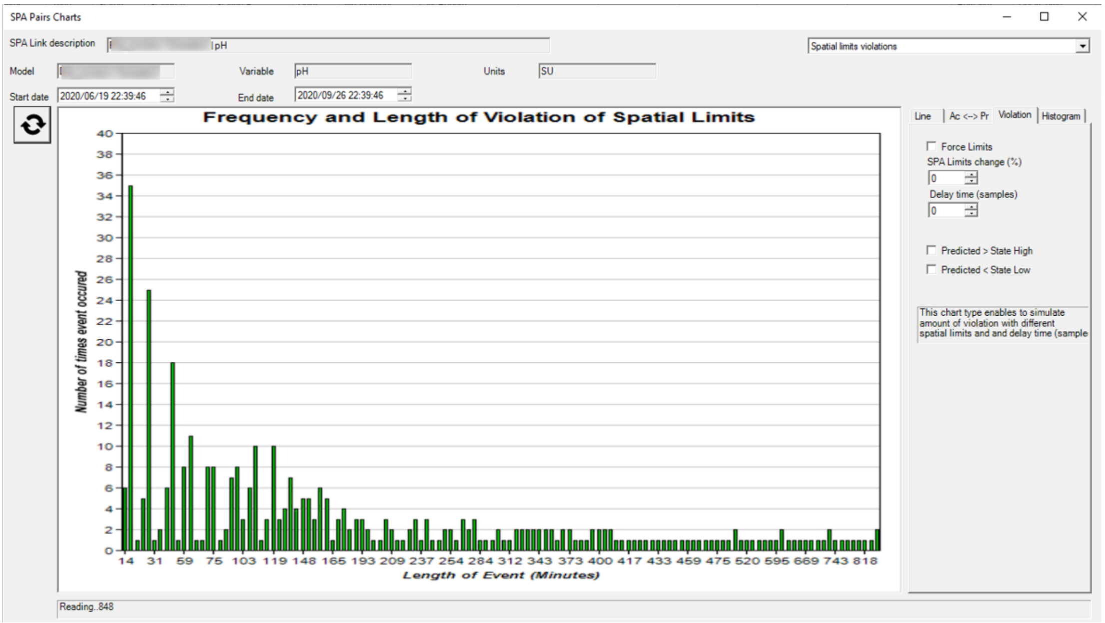
Min

□ Show 45 degree line

(set max and min field to show 45 line)

Ready.

Figure 142: An example for a graph comparing actual and predicted value for a Spatial Model pair model in the EDS thin client.



SPA Pairs Charts

SPA Link description [redacted] pH

Spatial limits violations

Model [redacted]

Variable pH

Units SU

Start date 2020/06/19 22:39:46

End date 2020/09/26 22:39:46

Line | Ac <--> Pr | Violation | Histogram

□ Force Limits

SPA Limits change (%) 0

Delay time (samples) 0

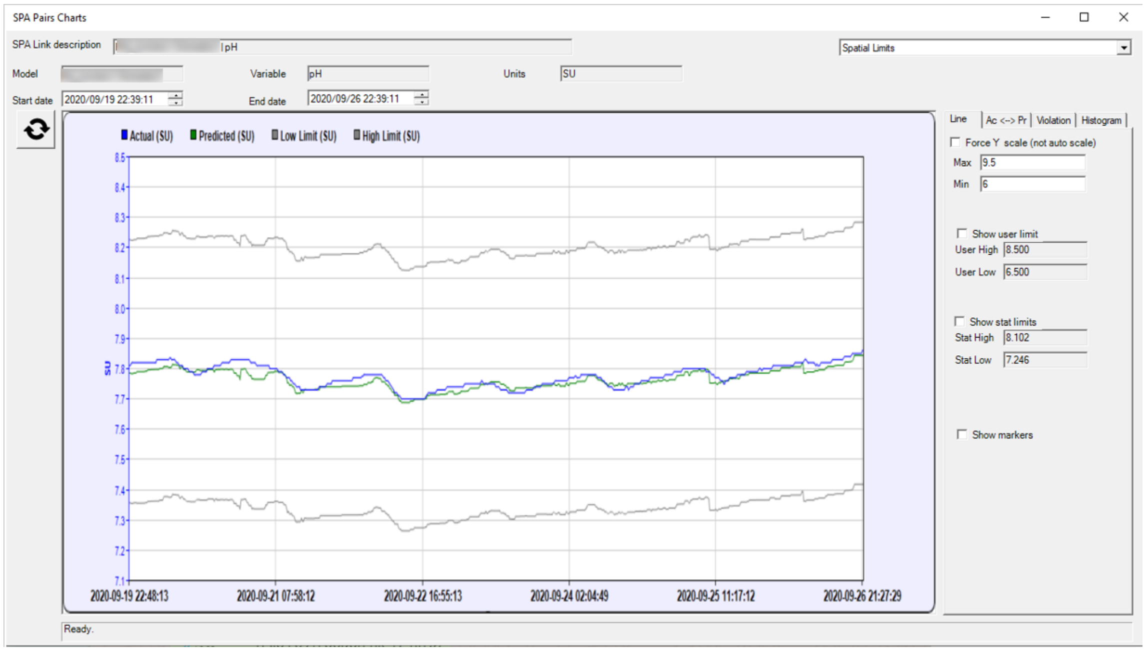
□ Predicted > State High

□ Predicted < State Low

This chart type enables to simulate amount of violation with different spatial limits and and delay time (sample

Reading..848

Figure 143: An example for a histogram of spatial model event lengths, for a Spatial Model pair model in the EDS thin client.



SPA Pairs Charts

SPA Link description [redacted] pH

Spatial Limits

Model [redacted]

Variable pH

Units SU

Start date 2020/09/19 22:39:11

End date 2020/09/26 22:39:11

Line | Ac <--> Pr | Violation | Histogram

□ Force Y scale (not auto scale)

Max 9.5

Min 6

□ Show user limit

User High 8.500

User Low 6.500

□ Show stat limits

Stat High 8.102

Stat Low 7.246

□ Show markers

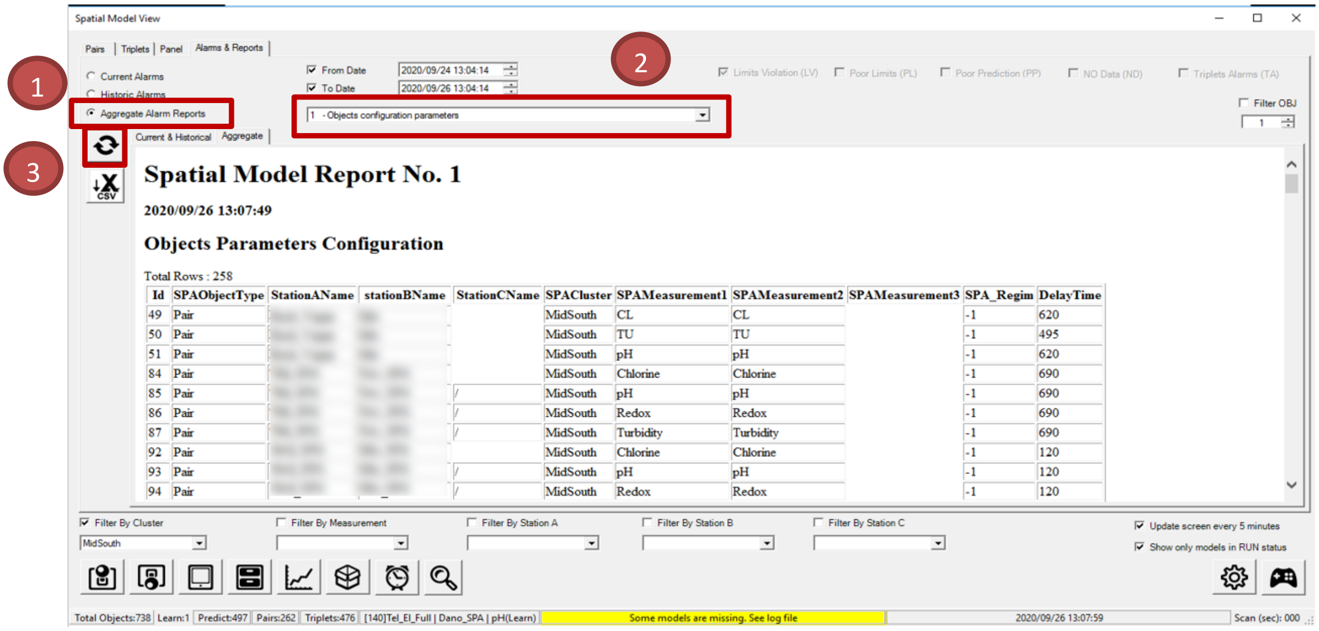
Ready.

Figure 144: An example a time series graph comparing predictions, measurements, and alarms thresholds, for a Spatial Model pair model in the EDS thin client.

4.12 Generating reports

This section discusses how aggregate reports can be generated for historical alarms.

Figure 145 shows the "Alarms & Reports" tab in the Spatial Model Viewer. In order to view aggregate level reports, select aggregate alarm reports **(Step I in Figure 145)**, select the report type **(Step II in Figure 145)**, and click the refresh button **(Step III in Figure 145)**.



☑ From Date 2020/09/24 13:04:14
☑ To Date 2020/09/26 13:04:14

1 - Objects configuration parameters

☑ Limits Violation (LV) □ Poor Limits (PL) □ Poor Prediction (PP) □ NO Data (ND) □ Triplets Alarms (TA)

□ Filter OBJ 1

Spatial Model Report No. 1

2020/09/26 13:07:49

Objects Parameters Configuration

Total Rows : 258

Id	SPAObjectType	StationAName	stationBName	StationCName	SPACluster	SPAMeasurement1	SPAMeasurement2	SPAMeasurement3	SPA_Regim	DelayTime
49	Pair	[redacted]	[redacted]		MidSouth	CL	CL		-1	620
50	Pair	[redacted]	[redacted]		MidSouth	TU	TU		-1	495
51	Pair	[redacted]	[redacted]		MidSouth	pH	pH		-1	620
84	Pair	[redacted]	[redacted]		MidSouth	Chlorine	Chlorine		-1	690
85	Pair	[redacted]	[redacted]	/	MidSouth	pH	pH		-1	690
86	Pair	[redacted]	[redacted]	/	MidSouth	Redox	Redox		-1	690
87	Pair	[redacted]	[redacted]	/	MidSouth	Turbidity	Turbidity		-1	690
92	Pair	[redacted]	[redacted]		MidSouth	Chlorine	Chlorine		-1	120
93	Pair	[redacted]	[redacted]	/	MidSouth	pH	pH		-1	120
94	Pair	[redacted]	[redacted]	/	MidSouth	Redox	Redox		-1	120

☑ Filter By Cluster: MidSouth
□ Filter By Measurement
□ Filter By Station A
□ Filter By Station B
□ Filter By Station C
☑ Update screen every 5 minutes
☑ Show only models in RUN status

Total Objects:738 | Learn:1 | Predict:497 | Pairs:262 | Triplets:476 | [140]Tel_El_Full \| Dano_SPA \| pH(Learn) | Some models are missing. See log file | 2020/09/26 13:07:59 | Scan (sec): 000

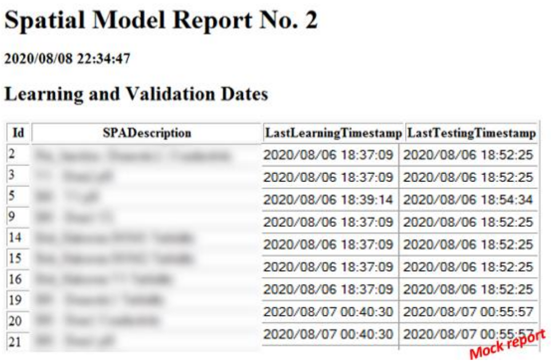
Figure 145: Selection of aggregate alarm reports and report type. Steps for selecting aggregate reports are numbered on the figure.

4.12.1 Report Types

Report No. 1 lists all SM objects, along with their types, stations, measurements, cluster identities, configured network regimes and Delay time for alarms. An example for the report is found in Figure HHH.

For the next reports, an example of the report follows after the report description.

Report No. 2 lists the times and dates each model performed flow time learning and testing for the measurements at the monitored end nodes.


Mock report

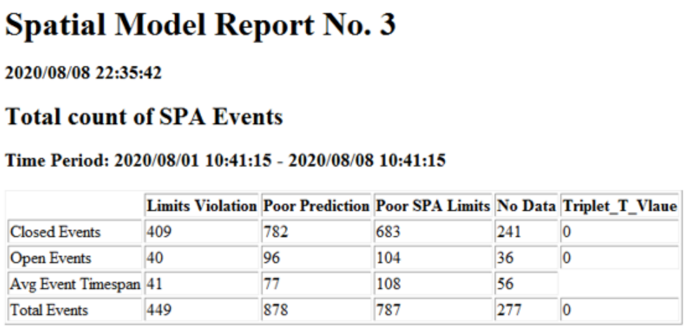
Spatial Model Report No. 2

2020/08/08 22:34:47

Learning and Validation Dates

Id	SPADescription	LastLearningTimestamp	LastTestingTimestamp
2	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
3	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
5	[redacted]	2020/08/06 18:39:14	2020/08/06 18:54:34
9	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
14	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
15	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
16	[redacted]	2020/08/06 18:37:09	2020/08/06 18:52:25
19	[redacted]	2020/08/07 00:40:30	2020/08/07 00:55:57
20	[redacted]	2020/08/07 00:40:30	2020/08/07 00:55:57
21	[redacted]	2020/08/07 00:40:30	2020/08/07 00:55:57

Report No. 3 lists the number of alarms of each type, across the system, with the number of open and closed events, and the average time span of an event in minutes.



Spatial Model Report No. 3

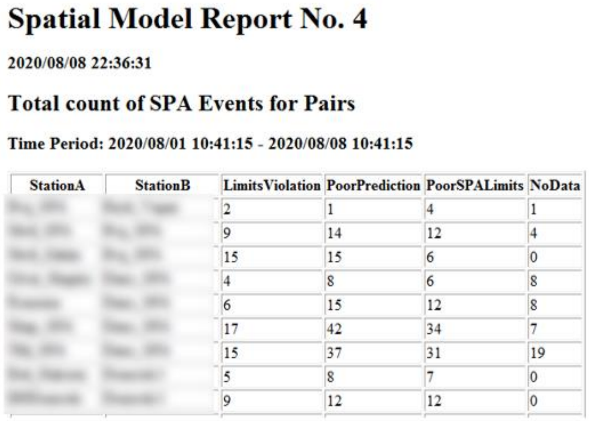
2020/08/08 22:35:42

Total count of SPA Events

Time Period: 2020/08/01 10:41:15 - 2020/08/08 10:41:15

	Limits Violation	Poor Prediction	Poor SPA Limits	No Data	Triplet_T_Vlaue
Closed Events	409	782	683	241	0
Open Events	40	96	104	36	0
Avg Event Timespan	41	77	108	56	
Total Events	449	878	787	277	0

Report No. 4 lists the total number of events of each type, for each pair in the system.



Spatial Model Report No. 4

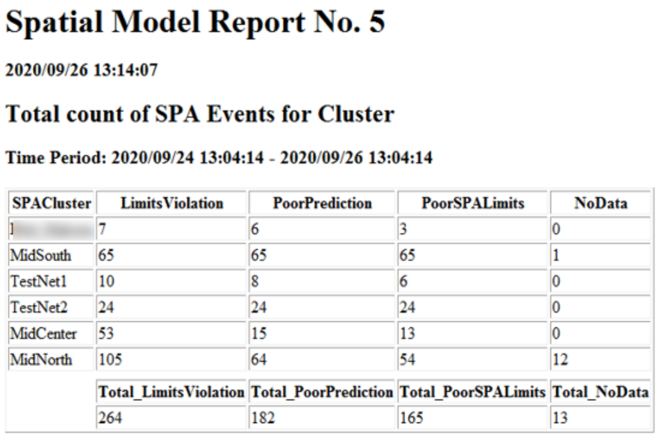
2020/08/08 22:36:31

Total count of SPA Events for Pairs

Time Period: 2020/08/01 10:41:15 - 2020/08/08 10:41:15

StationA	StationB	LimitsViolation	PoorPrediction	PoorSPALimits	NoData
[redacted]	[redacted]	2	1	4	1
[redacted]	[redacted]	9	14	12	4
[redacted]	[redacted]	15	15	6	0
[redacted]	[redacted]	4	8	6	8
[redacted]	[redacted]	6	15	12	8
[redacted]	[redacted]	17	42	34	7
[redacted]	[redacted]	15	37	31	19
[redacted]	[redacted]	5	8	7	0
[redacted]	[redacted]	9	12	12	0

Report No. 5 lists the total number of events, by type, for each cluster of stations in the network.



Spatial Model Report No. 5

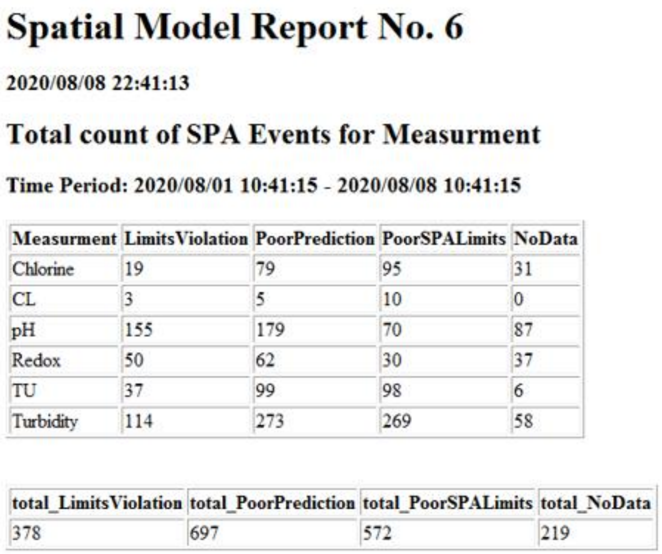
2020/09/26 13:14:07

Total count of SPA Events for Cluster

Time Period: 2020/09/24 13:04:14 - 2020/09/26 13:04:14

SPACluster	LimitsViolation	PoorPrediction	PoorSPALimits	NoData
[redacted]	7	6	3	0
MidSouth	65	65	65	1
TestNet1	10	8	6	0
TestNet2	24	24	24	0
MidCenter	53	15	13	0
MidNorth	105	64	54	12
	Total_LimitsViolation	Total_PoorPrediction	Total_PoorSPALimits	Total_NoData
	264	182	165	13

Report No. 6 shows a cross-tabulation for the number of events by event type and measurement type.



Spatial Model Report No. 6

2020/08/08 22:41:13

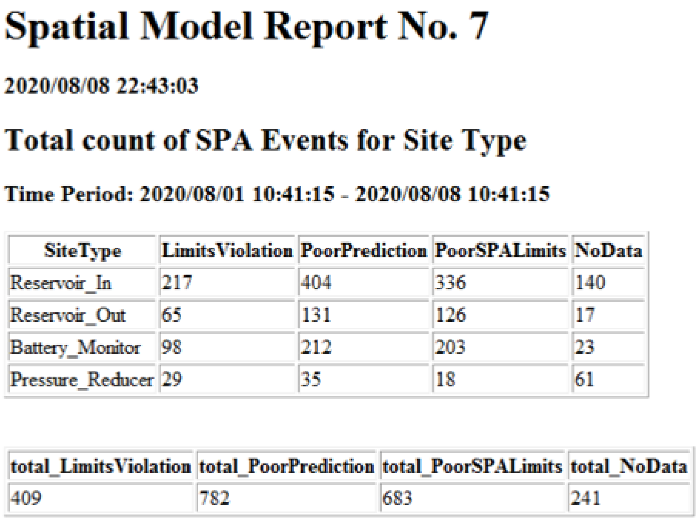
Total count of SPA Events for Measurment

Time Period: 2020/08/01 10:41:15 - 2020/08/08 10:41:15

Measurment	LimitsViolation	PoorPrediction	PoorSPALimits	NoData
Chlorine	19	79	95	31
CL	3	5	10	0
pH	155	179	70	87
Redox	50	62	30	37
TU	37	99	98	6
Turbidity	114	273	269	58

total_LimitsViolation	total_PoorPrediction	total_PoorSPALimits	total_NoData
378	697	572	219

Report No. 7 shows a cross-tabulation of the number of alarms, by alarm type and site type (reservoir in, reservoir out, Battery Monitor, or Pressure reducer).



Spatial Model Report No. 7

2020/08/08 22:43:03

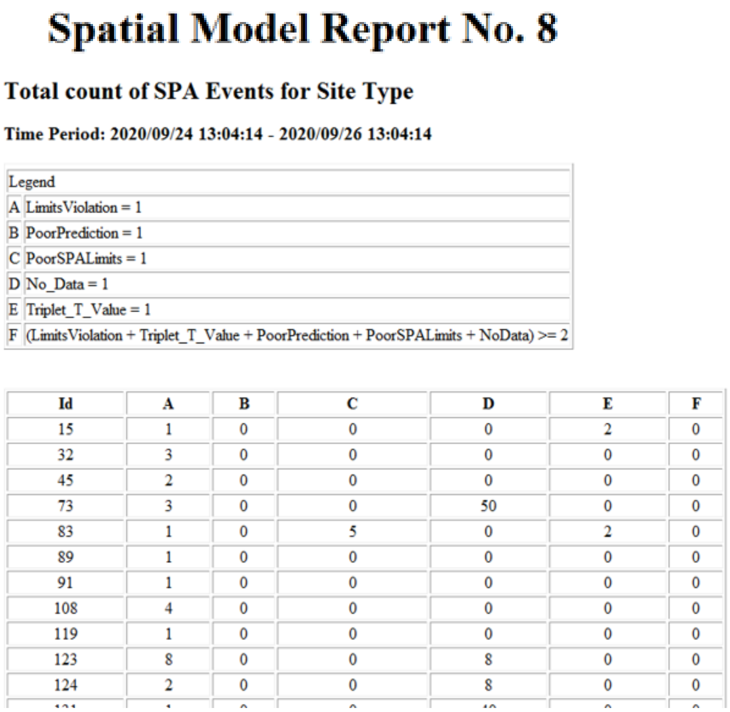
Total count of SPA Events for Site Type

Time Period: 2020/08/01 10:41:15 - 2020/08/08 10:41:15

SiteType	LimitsViolation	PoorPrediction	PoorSPALimits	NoData
Reservoir_In	217	404	336	140
Reservoir_Out	65	131	126	17
Battery_Monitor	98	212	203	23
Pressure_Reducer	29	35	18	61

total_LimitsViolation	total_PoorPrediction	total_PoorSPALimits	total_NoData
409	782	683	241

<u>Report No. 8</u> lists the total number of events by event type (see legend below), and Site ID.



Spatial Model Report No. 8

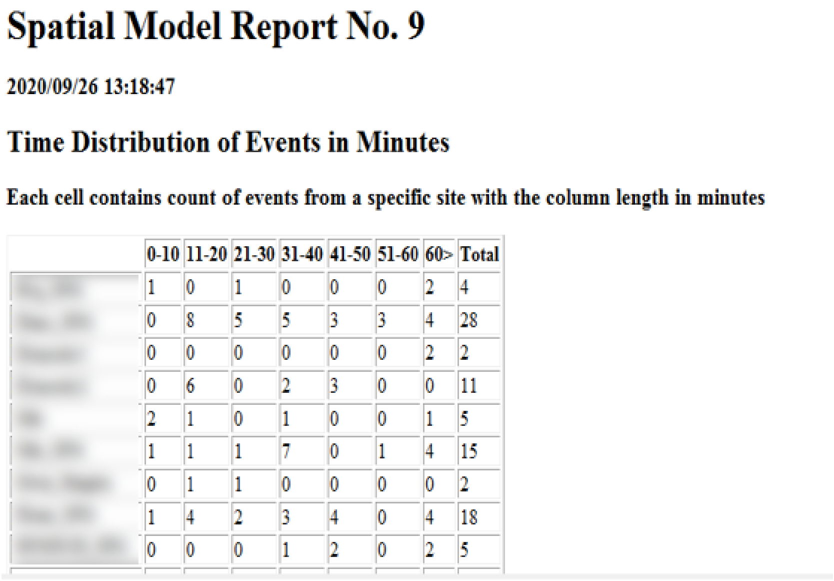
Total count of SPA Events for Site Type

Time Period: 2020/09/24 13:04:14 - 2020/09/26 13:04:14

Legend	
A	LimitsViolation = 1
B	PoorPrediction = 1
C	PoorSPALimits = 1
D	No_Data = 1
E	Triplet_T_Value = 1
F	(LimitsViolation + Triplet_T_Value + PoorPrediction + PoorSPALimits + NoData) >= 2

Id	A	B	C	D	E	F
15	1	0	0	0	2	0
32	3	0	0	0	0	0
45	2	0	0	0	0	0
73	3	0	0	50	0	0
83	1	0	5	0	2	0
89	1	0	0	0	0	0
91	1	0	0	0	0	0
108	4	0	0	0	0	0
119	1	0	0	0	0	0
123	8	0	0	8	0	0
124	2	0	0	8	0	0

<u>Report No. 9</u> the distribution of events time, by 10 minute bins (last bin is for all events time larger than 1 hour), for each network site.



Spatial Model Report No. 9

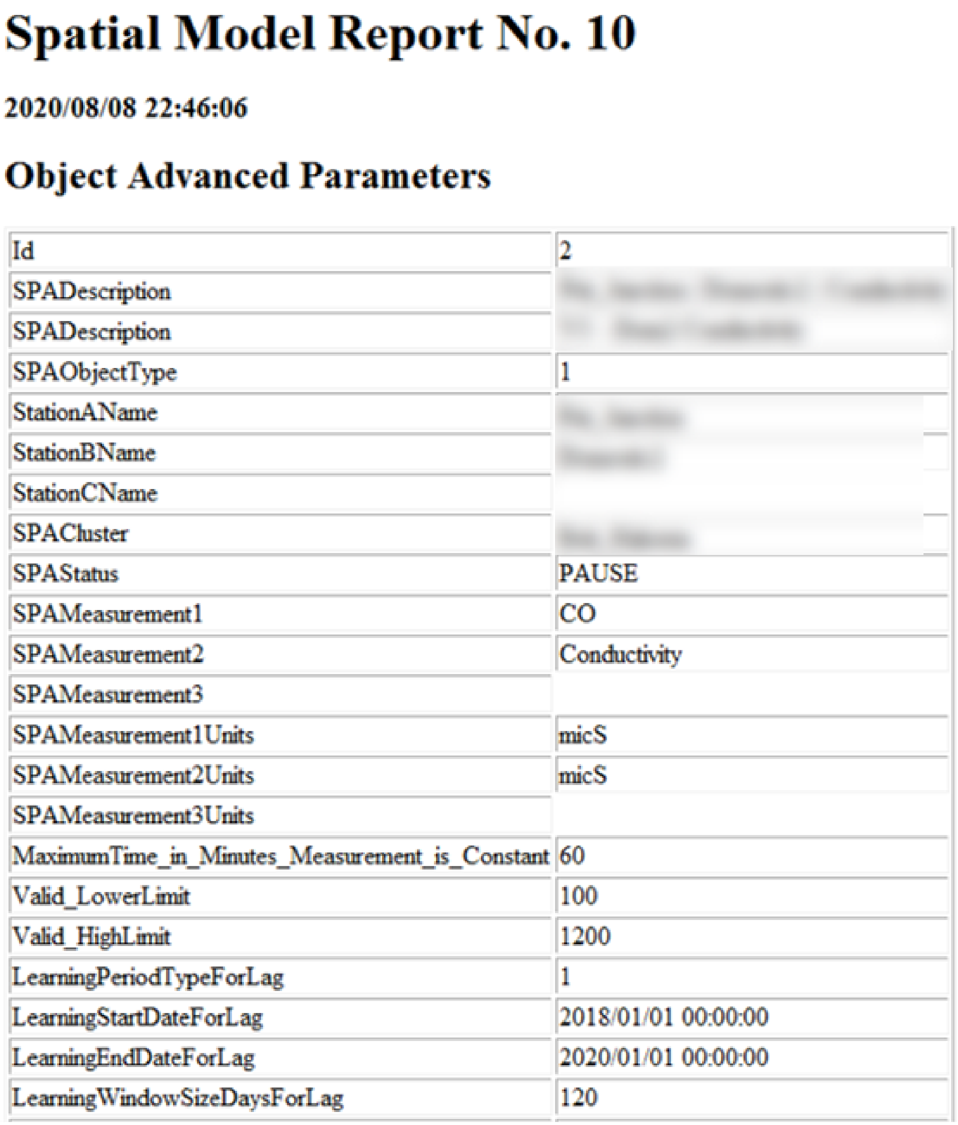
2020/09/26 13:18:47

Time Distribution of Events in Minutes

Each cell contains count of events from a specific site with the column length in minutes

	0-10	11-20	21-30	31-40	41-50	51-60	60>	Total
[redacted]	1	0	1	0	0	0	2	4
[redacted]	0	8	5	5	3	3	4	28
[redacted]	0	0	0	0	0	0	2	2
[redacted]	0	6	0	2	3	0	0	11
[redacted]	2	1	0	1	0	0	1	5
[redacted]	1	1	1	7	0	1	4	15
[redacted]	0	1	1	0	0	0	0	2
[redacted]	1	4	2	3	4	0	4	18
[redacted]	0	0	0	1	2	0	2	5

Report No. 10 lists all object parameters, for all SM objects. An example for a single object is found below.



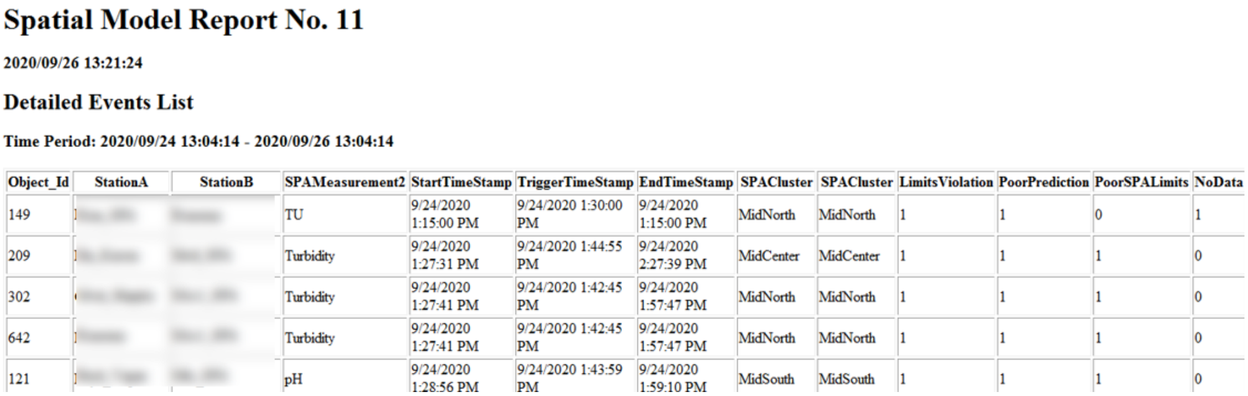
Spatial Model Report No. 10

2020/08/08 22:46:06

Object Advanced Parameters

Id	2
SPADescription	[redacted]
SPADescription	[redacted]
SPAObjectType	1
StationAName	[redacted]
StationBName	[redacted]
StationCName	[redacted]
SPACluster	[redacted]
SPAStatus	PAUSE
SPAMeasurement1	CO
SPAMeasurement2	Conductivity
SPAMeasurement3	
SPAMeasurement1Units	micS
SPAMeasurement2Units	micS
SPAMeasurement3Units	
MaximumTime_in_Minutes_Measurement_is_Constant	60
Valid_LowerLimit	100
Valid_HighLimit	1200
LearningPeriodTypeForLag	1
LearningStartDateForLag	2018/01/01 00:00:00
LearningEndDateForLag	2020/01/01 00:00:00
LearningWindowSizeDaysForLag	120

Report No. 11 lists all historical events, by station, events time, measurement, and event type.



Spatial Model Report No. 11

2020/09/26 13:21:24

Detailed Events List

Time Period: 2020/09/24 13:04:14 - 2020/09/26 13:04:14

Object_Id	StationA	StationB	SPAMeasurement2	StartTimeStamp	TriggerTimeStamp	EndTimeStamp	SPACluster	SPACluster	LimitsViolation	PoorPrediction	PoorSPALimits	NoData
149	[redacted]	[redacted]	TU	9/24/2020 1:15:00 PM	9/24/2020 1:30:00 PM	9/24/2020 1:15:00 PM	MidNorth	MidNorth	1	1	0	1
209	[redacted]	[redacted]	Turbidity	9/24/2020 1:27:31 PM	9/24/2020 1:44:55 PM	9/24/2020 2:27:39 PM	MidCenter	MidCenter	1	1	1	0
302	[redacted]	[redacted]	Turbidity	9/24/2020 1:27:41 PM	9/24/2020 1:42:45 PM	9/24/2020 1:57:47 PM	MidNorth	MidNorth	1	1	1	0
642	[redacted]	[redacted]	Turbidity	9/24/2020 1:27:41 PM	9/24/2020 1:42:45 PM	9/24/2020 1:57:47 PM	MidNorth	MidNorth	1	1	1	0
121	[redacted]	[redacted]	pH	9/24/2020 1:28:56 PM	9/24/2020 1:43:59 PM	9/24/2020 1:59:10 PM	MidSouth	MidSouth	1	1	1	0

4.13 Quiz for Unit C

1. What types of models are available in the Spatial Model?

2. What types of alarms are can be raised by the spatial model?

3. What are the possible network regimes the SM may be in?

4. When is a "poor statistical limits" alarm raised?

5. When is a "poor prediction" alarm raised?

6. What is the role of the delay time in alarm generation, and how does it affect the alarms raised?

7. What type of charts are available in the Spatial Model?

8. What types of reports are available in the Spatial Model?

9. What parameters affect flow time estimation?

10. How can the credibility of flow time estimates be assessed?

11. Why would the user select to copy flow time estimates from one pair model to another?

12. Briefly explain what is the prediction model used to predict the target node measurement value. Explain what are the explanatory and predicted variables, and how flow time is taken into consideration.

13. What parameters are used for configuring the prediction values?

Appendix A – Additional graphs for flow time estimation and cross validation

This appendix shows Score function graphs for additional models.

A.1 additional graph for flow time estimates by regimes

Figures 146-153 show score functions for different objects

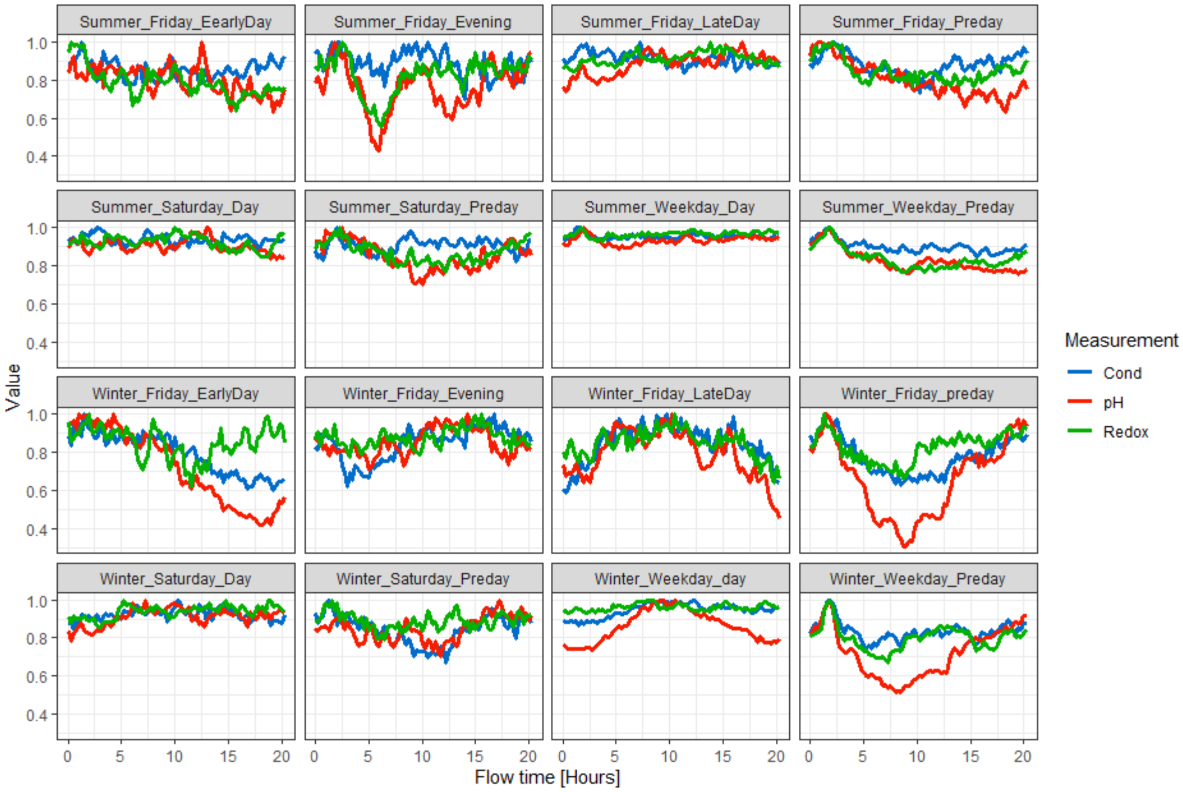


Figure 146: Score function for objects O164,O166,O167, three pair objects for the same source and target stations, using different measurements. Score functions for measurements in the same regime were normalized to have a maximum of 1 in each subplot so that the estimated flow-time (X coordinate of the maximum) could be compared for estimates obtained from different measurements.

O69:

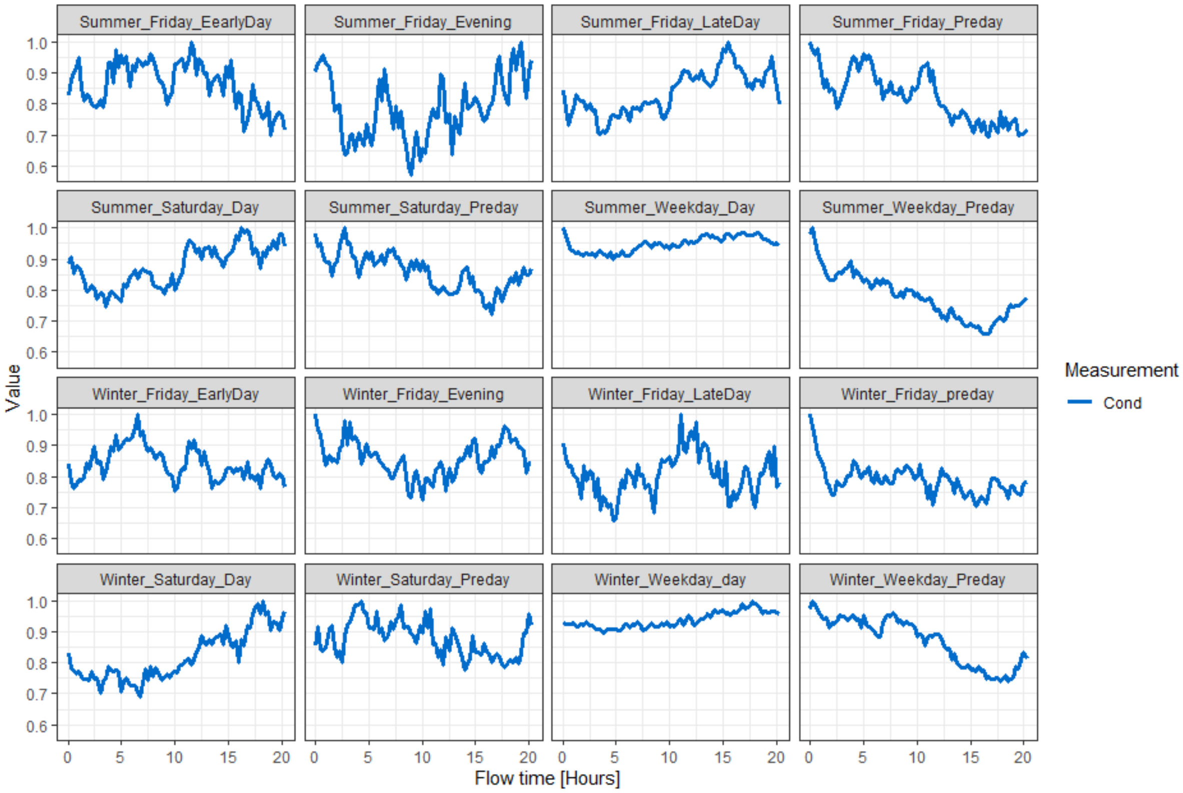


Figure 147: Score function for objects O69, for 16 regimes. Score functions computed based on Conductivity measurements.

O76:

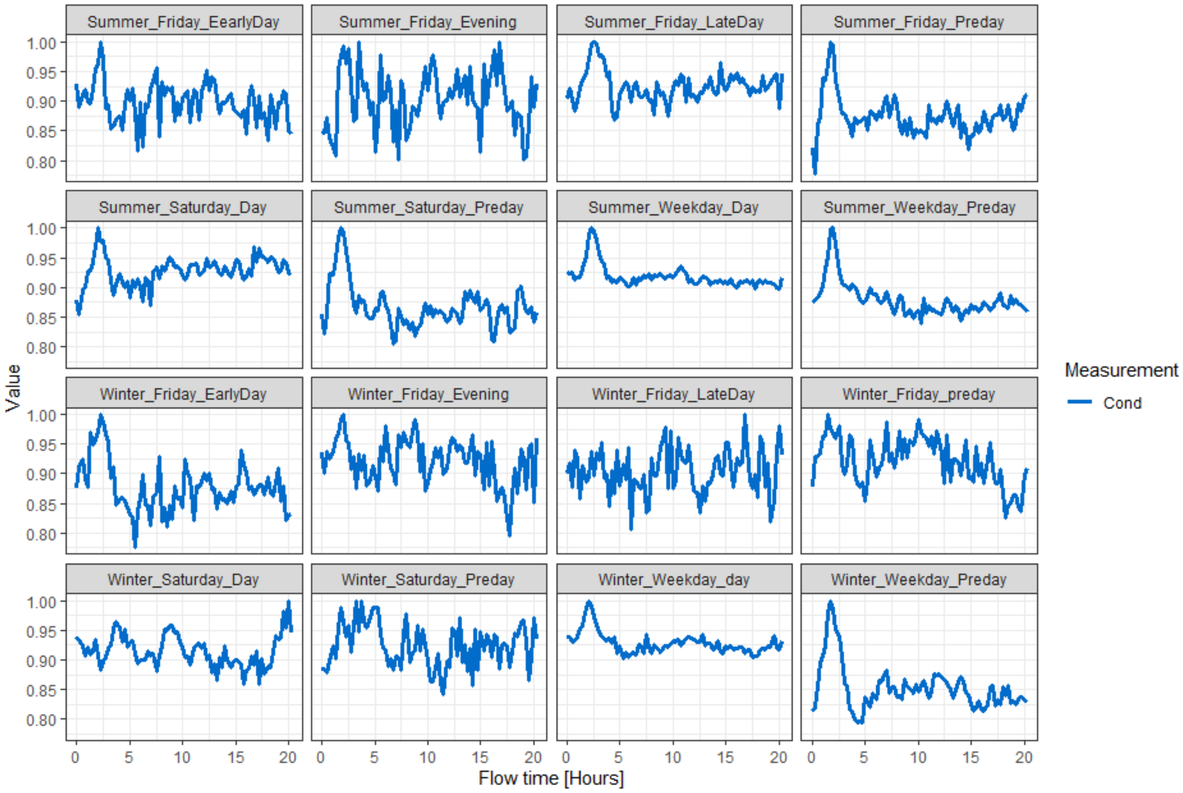


Figure 148: Score function for objects O76, for 16 regimes. Score functions computed based on Conductivity measurements.

O107:

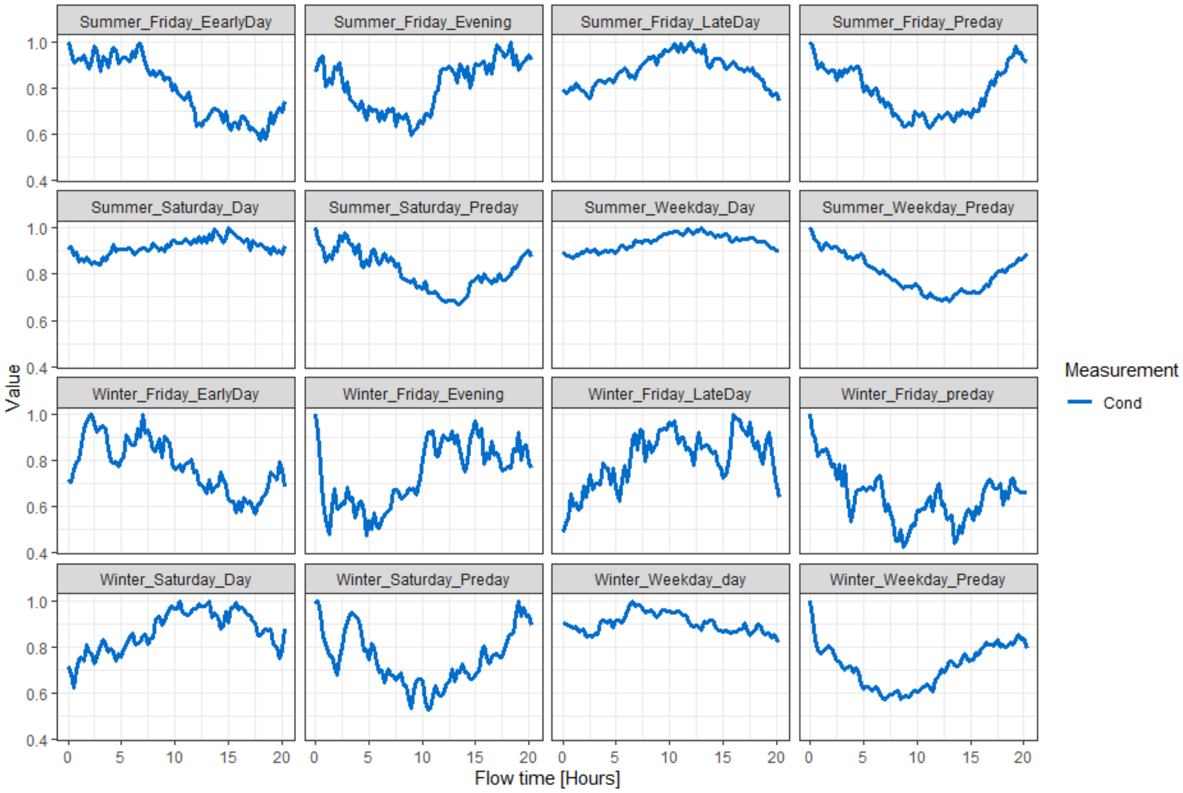


Figure 149: Score function for objects O107, for 16 regimes. Score functions computed based on Conductivity measurements.

O116

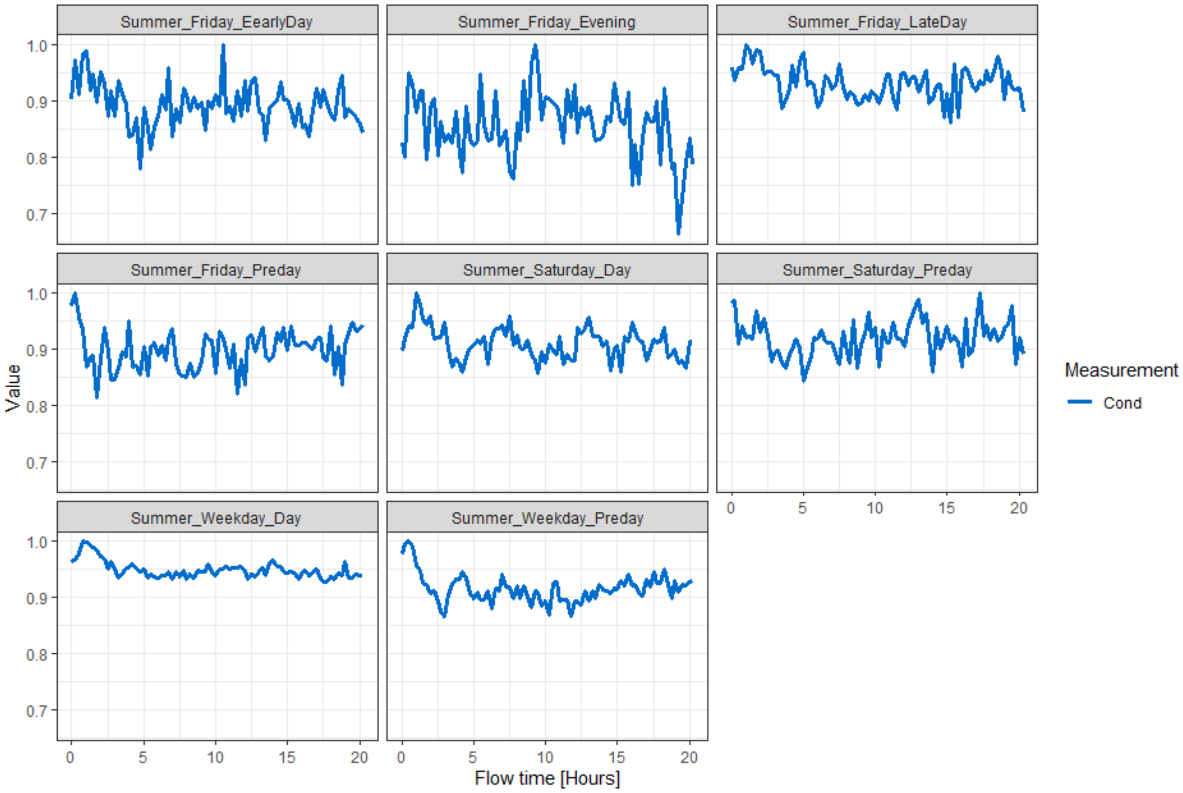


Figure 150: Score function for objects O116, for 16 regimes. Score functions computed based on Conductivity measurements.

O240:

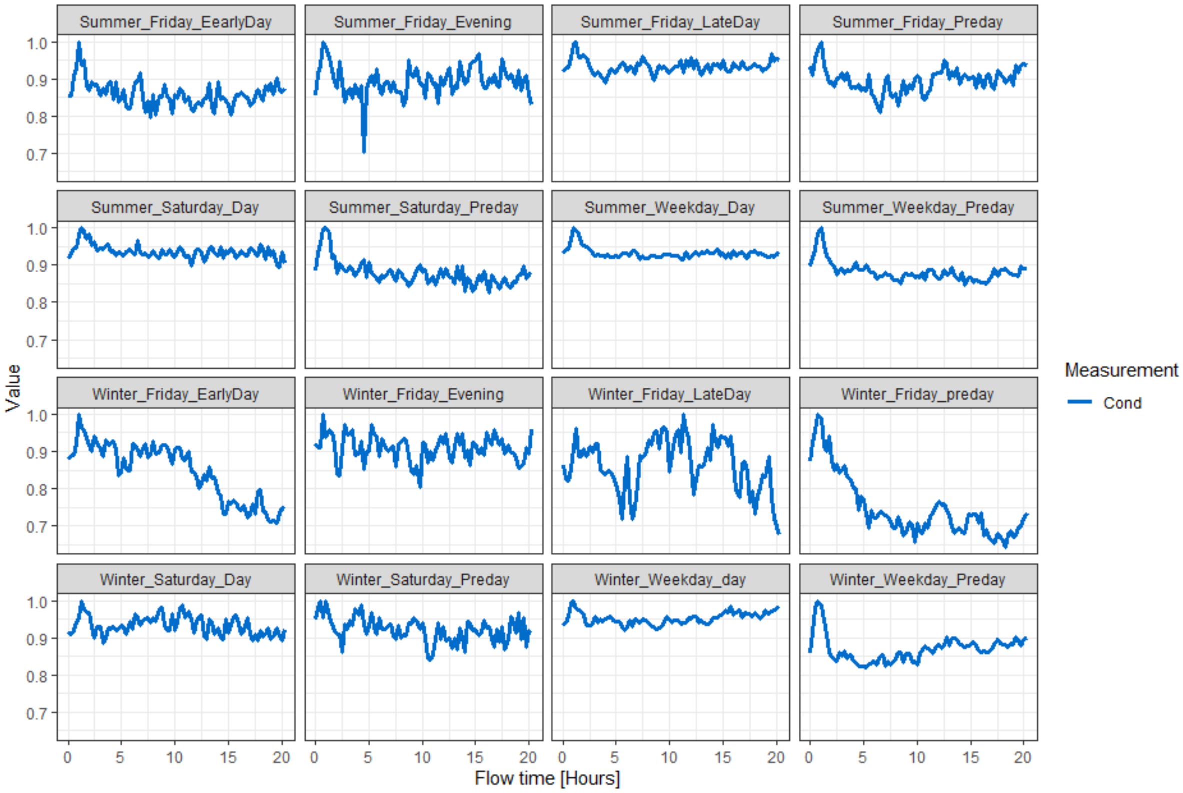


Figure 151: Score function for objects O240, for 16 regimes. Score functions computed based on Conductivity measurements.

O164:

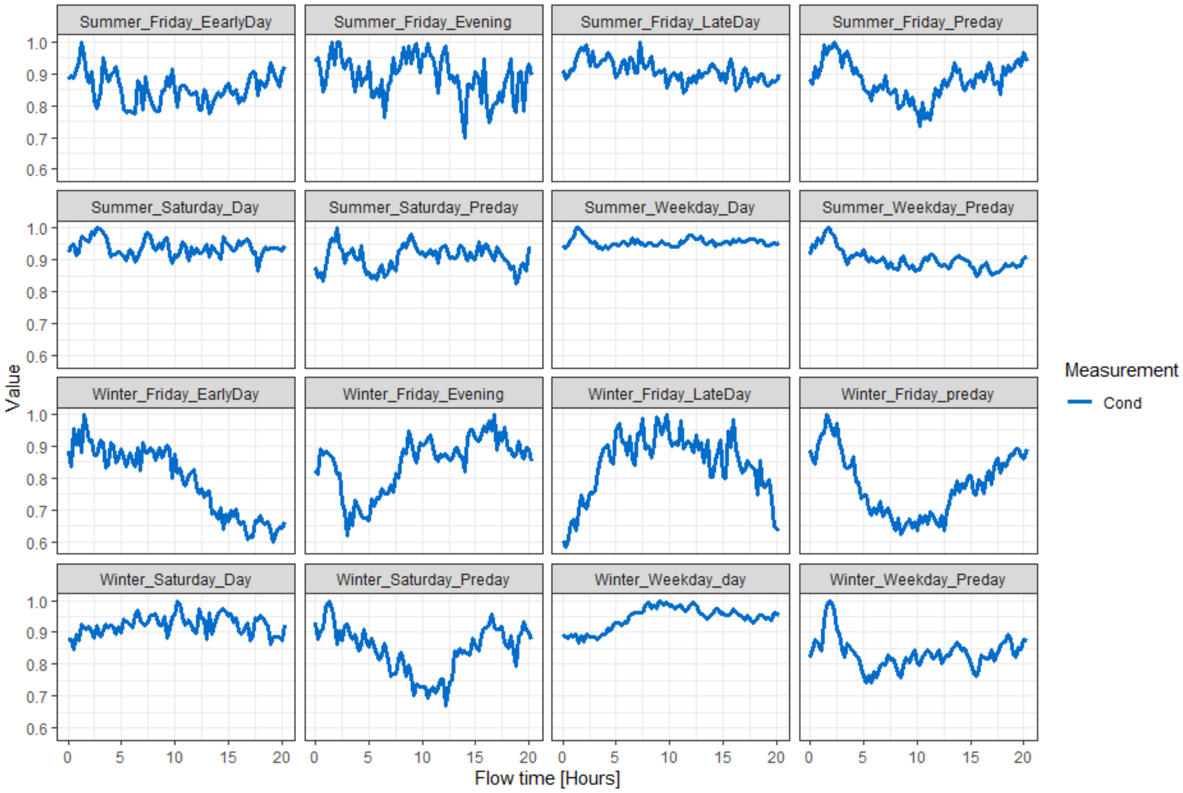


Figure 152: Score function for objects O164, for 16 regimes. Score functions computed based on Conductivity measurements.

O186

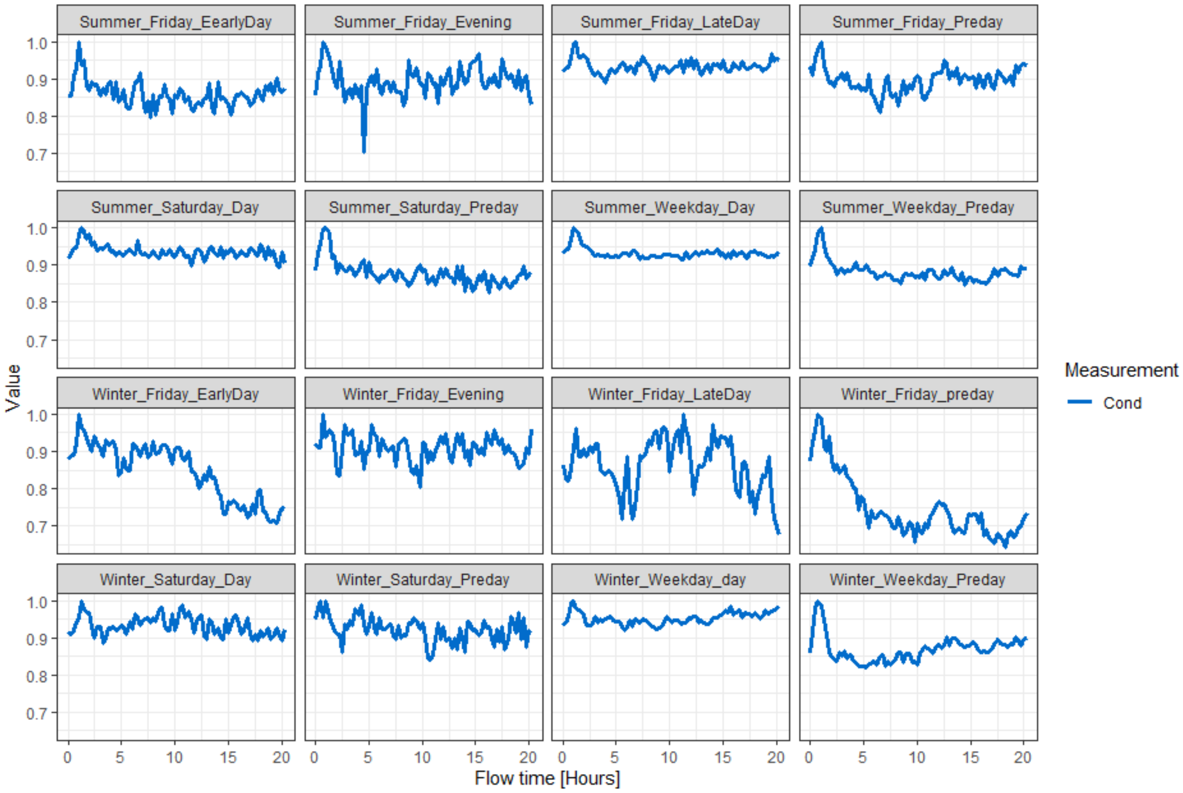


Figure 153: Score function for objects O186, for 16 regimes. Score functions computed based on Conductivity measurements.

A.2 Additional graph for error estimates for estimated flow times

Figures 154-158 show cross-validation estimates of flow times for different objects. The cross validation procedure is presented in Section 3.4.
Note that for some Saturday-preday and Friday Evening Regimes, flow times are large and cross validation estimates do not agree.

O116 -

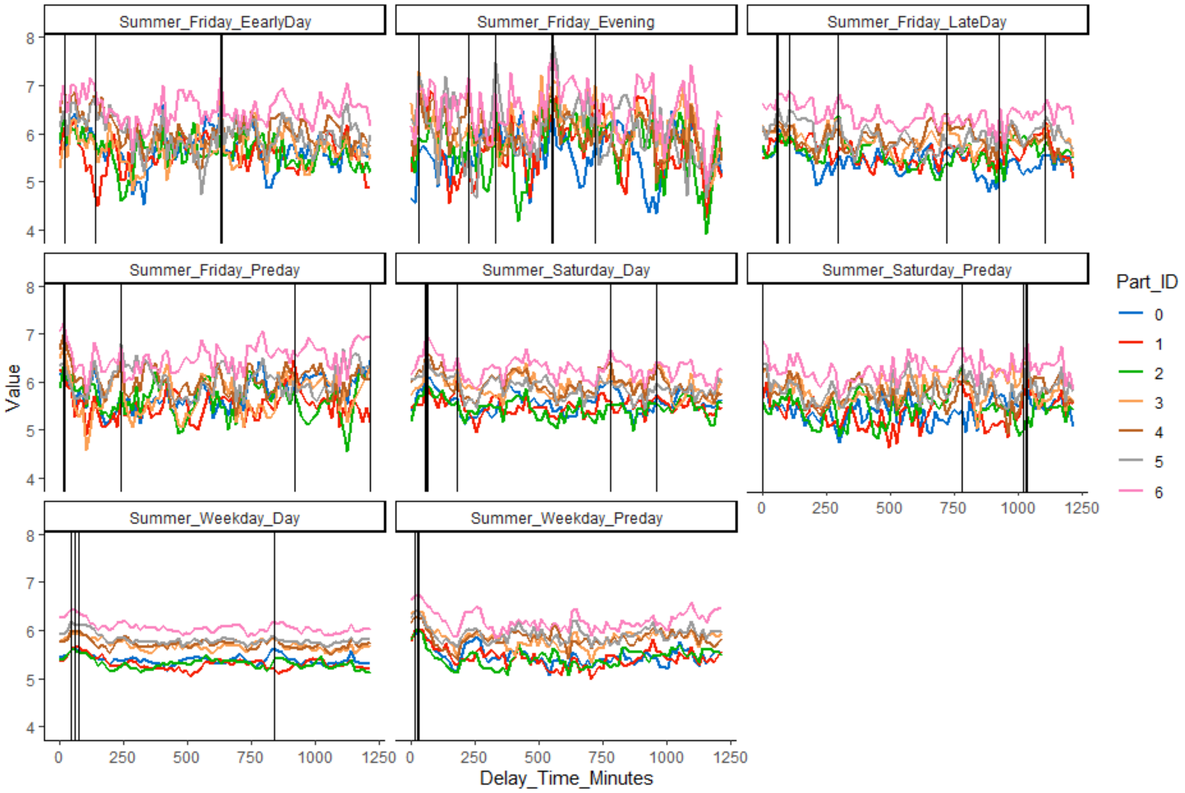


Figure 154: Score functions for cross-validations, for object O116.

Object -O76

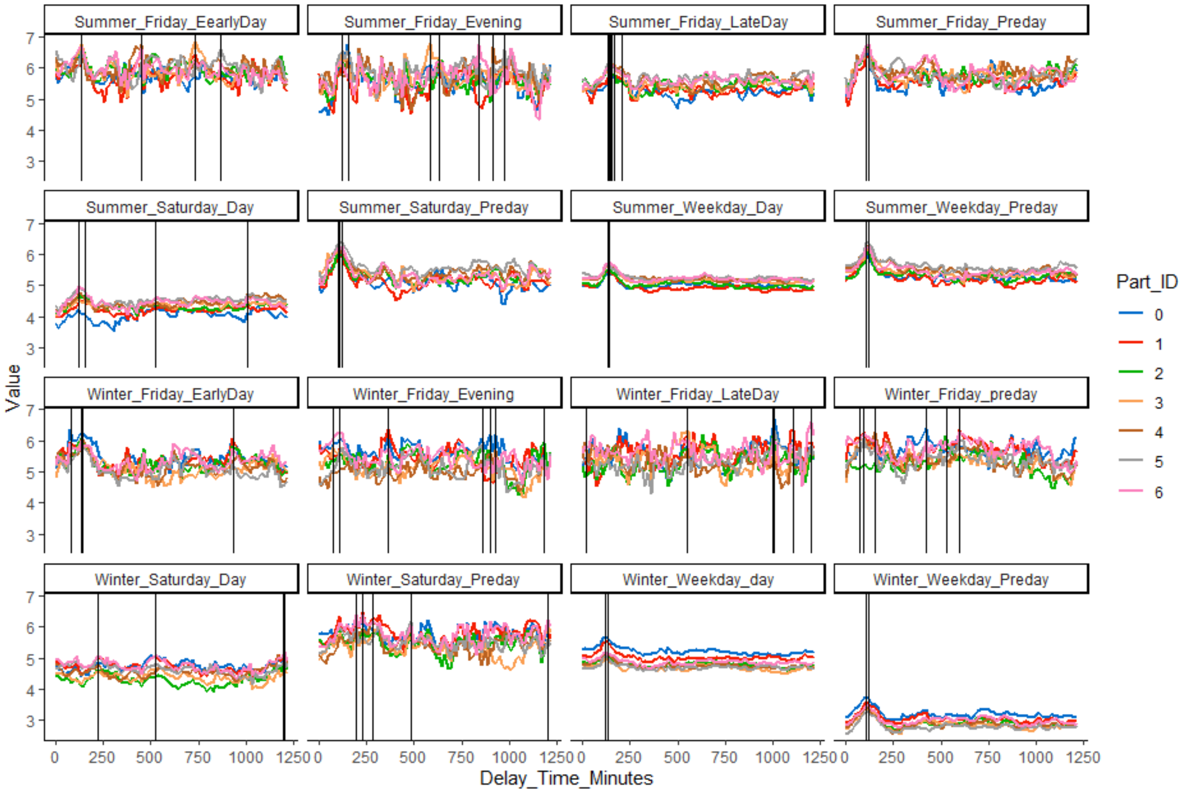


Figure 155: Score functions for cross validations, for object O76.

O240 -

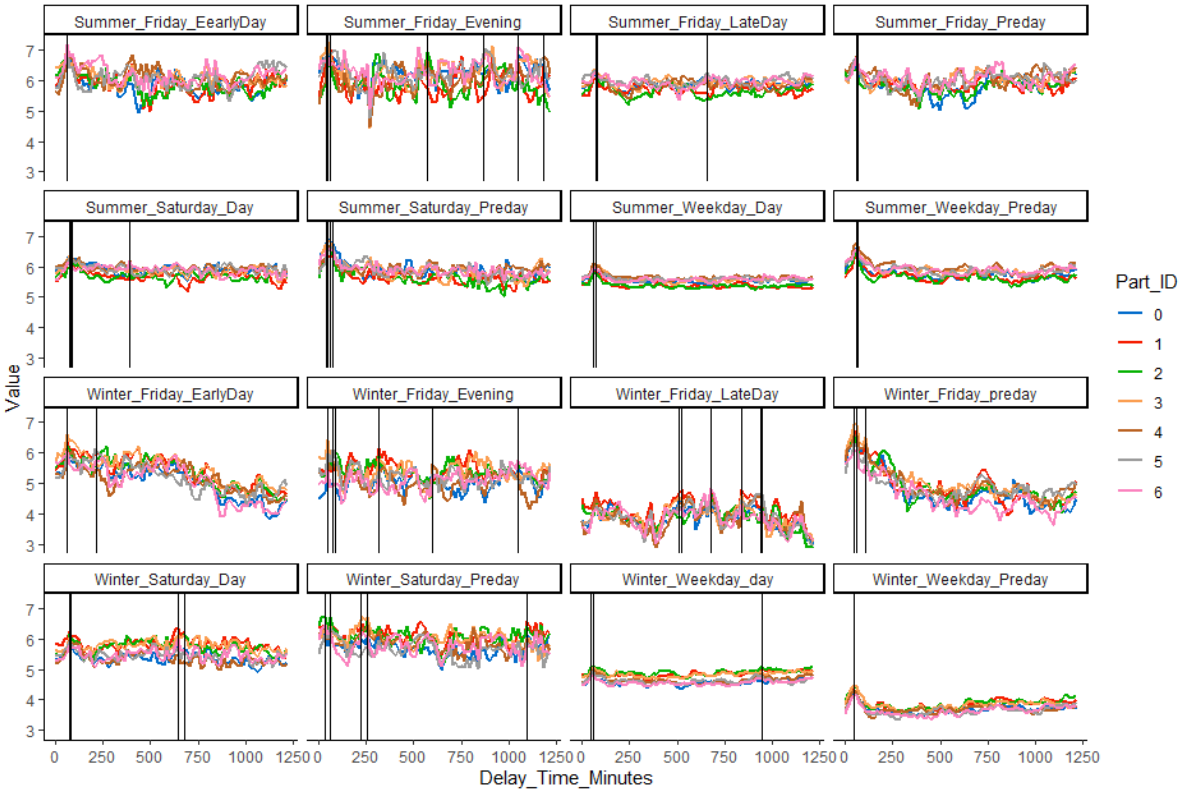


Figure 156: Score functions for cross validations, for object O240.

O164 -

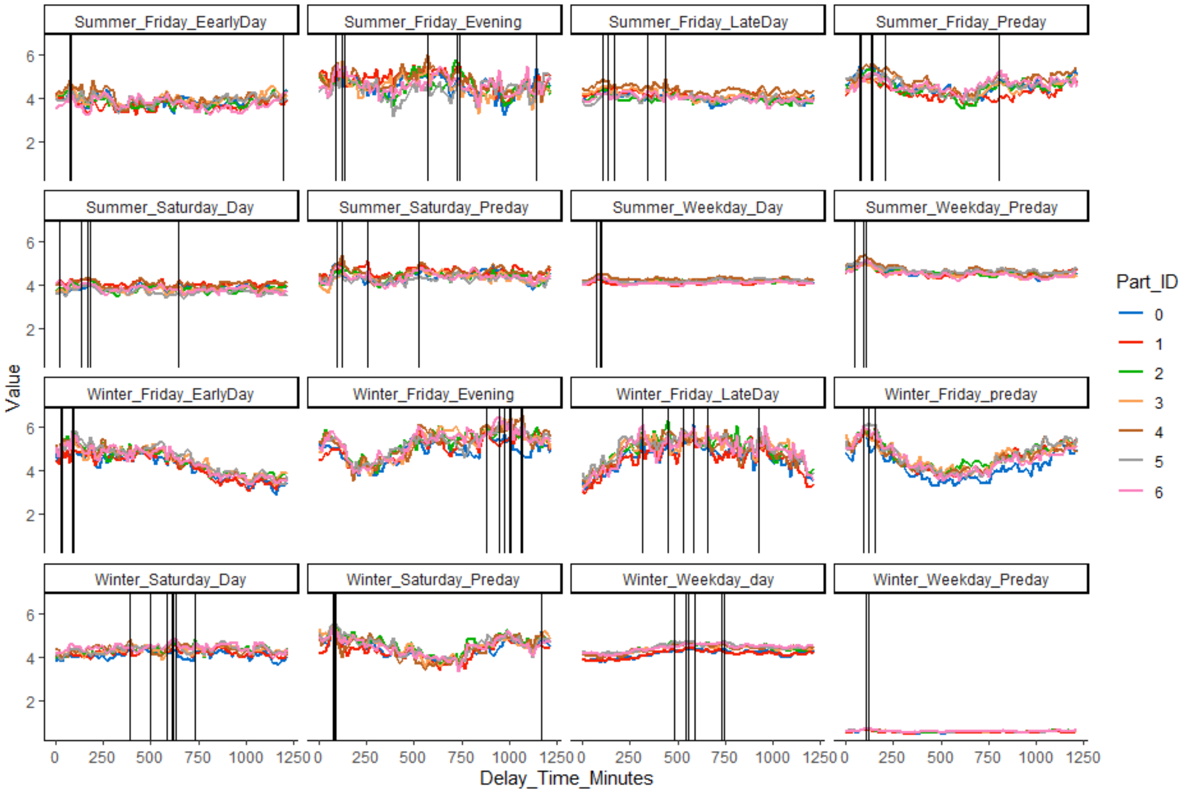


Figure 157: Score functions for cross validations, for object O164.

O186 -

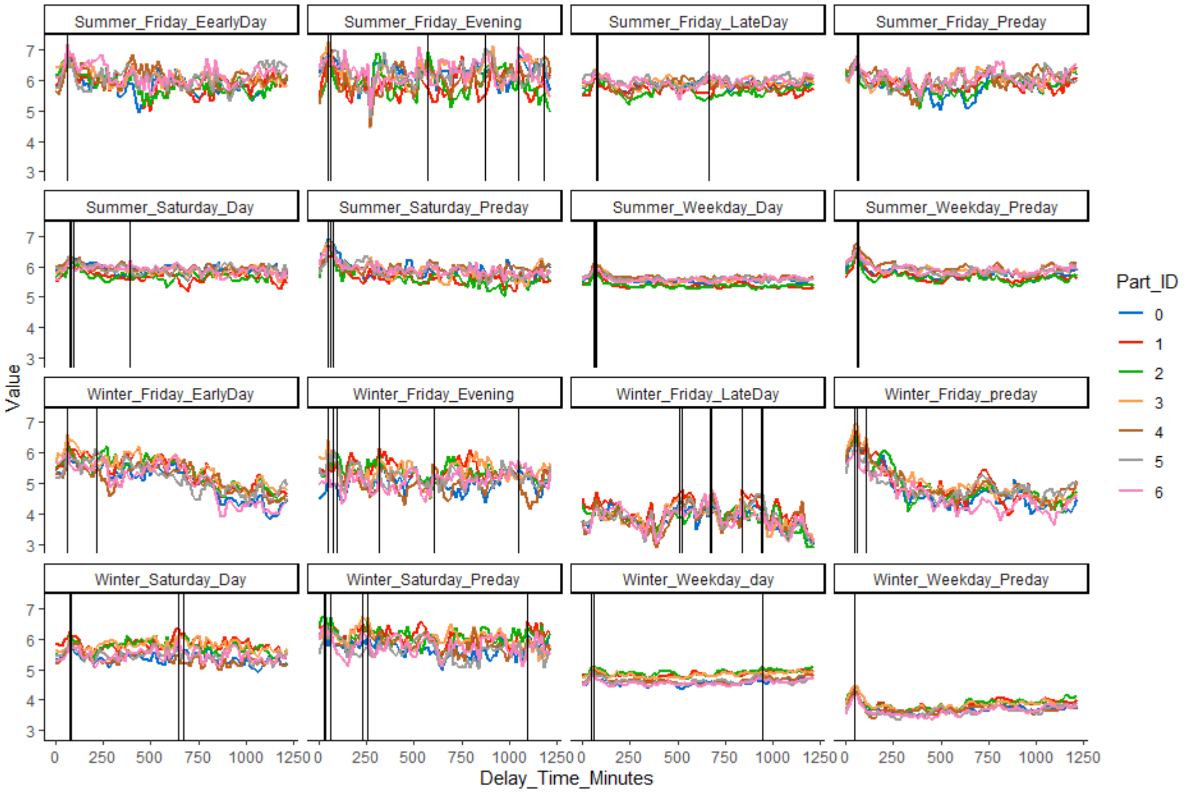


Figure 158: Score functions for cross validations, for object O186.

Appendix B – Additional graphs for the triplets-gamma hypothesis

This appendix shows additional examples for the distribution of T statistics, and their fit to gamma distributions.

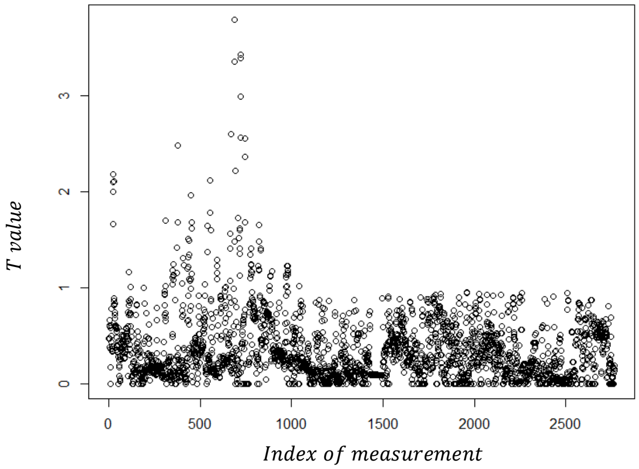


Figure 159: A sample of data points from Triplet 470, one hour intervals

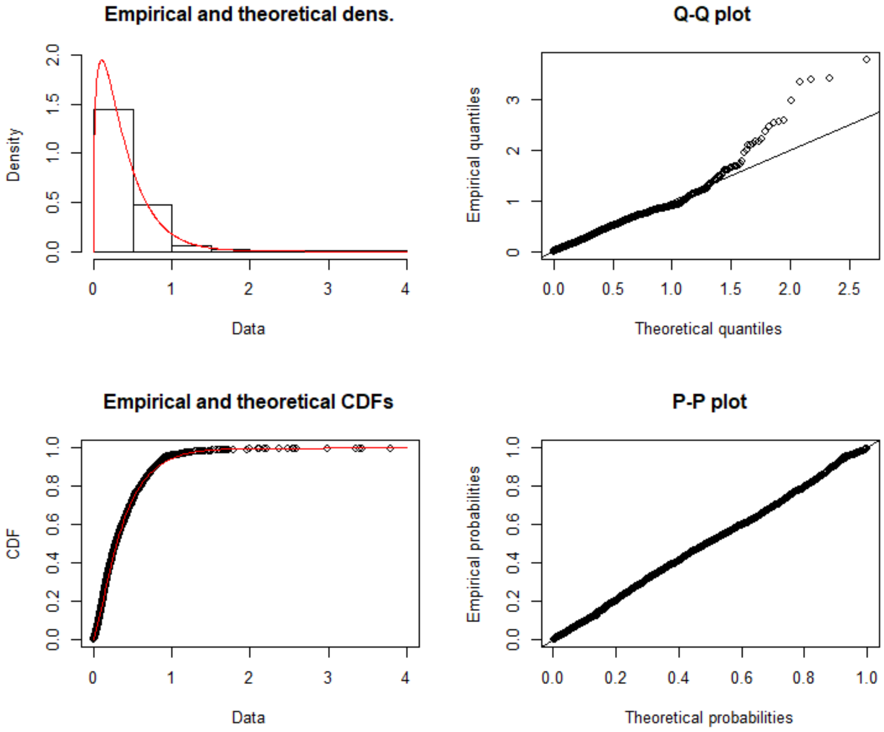


Figure 160: Comparison between the data shown in Figure 159, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.

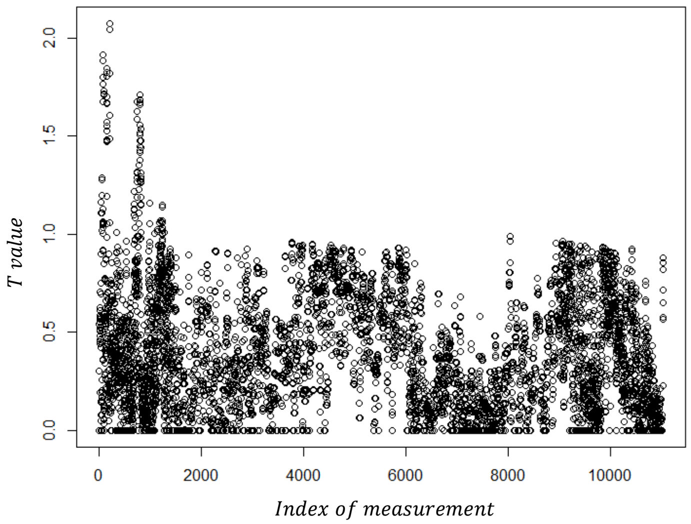


Figure 161: A sample of data points from Triplet 476, one hour intervals

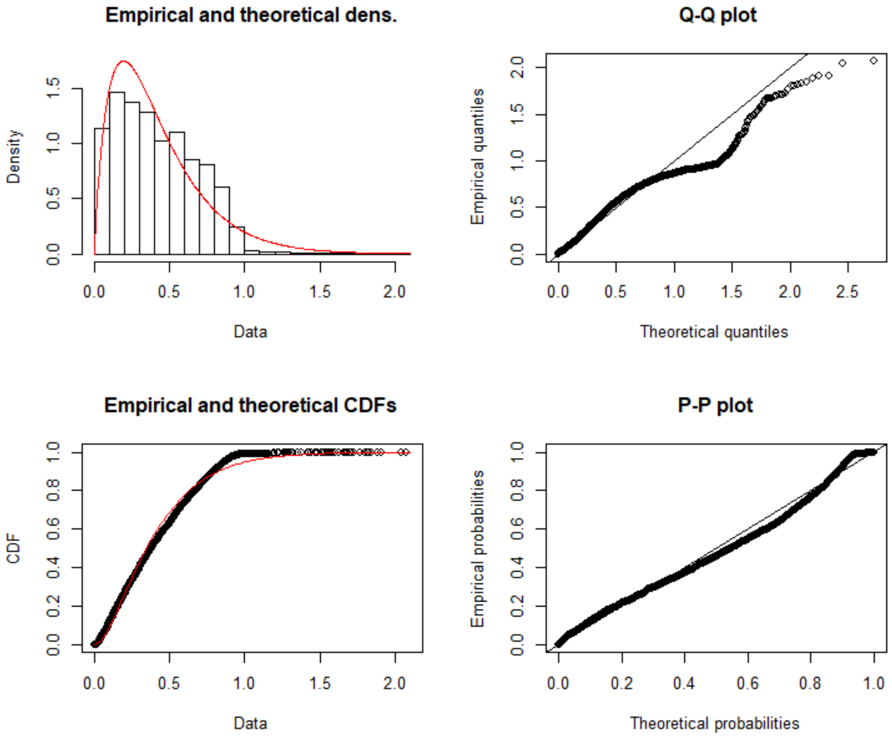


Figure 162: Comparison between the data shown in Figure 161, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.

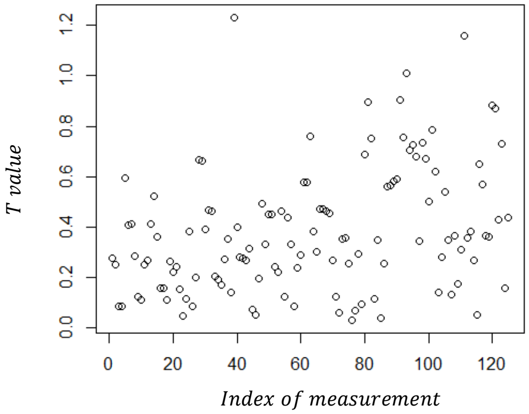


Figure 163: A sample of data points from Triplet 308, half hour intervals

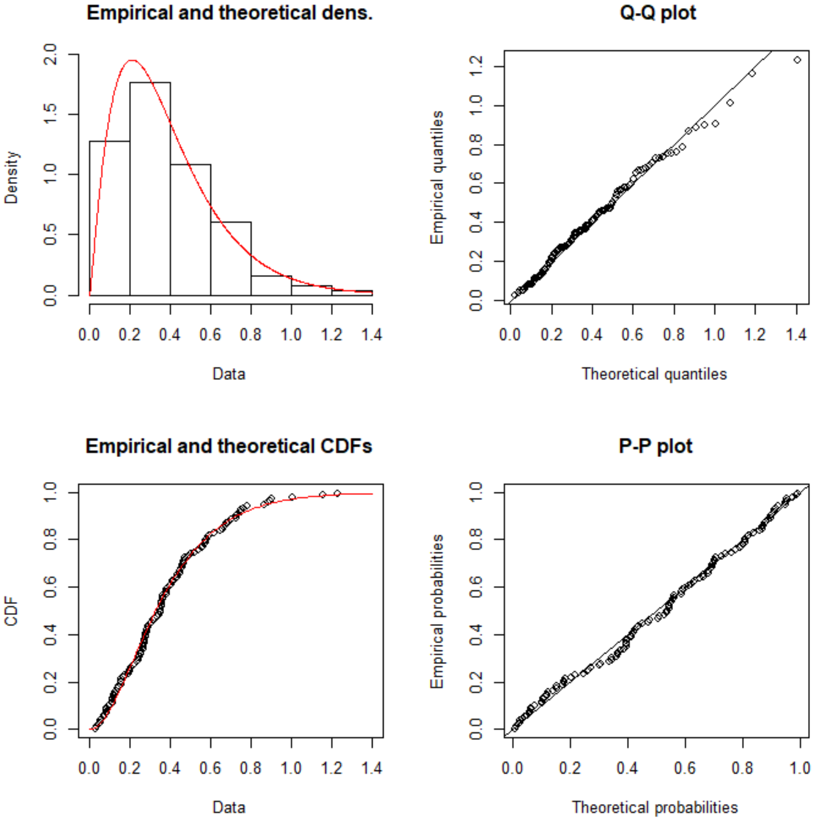


Figure 164: Comparison between the data shown in Figure 163, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.

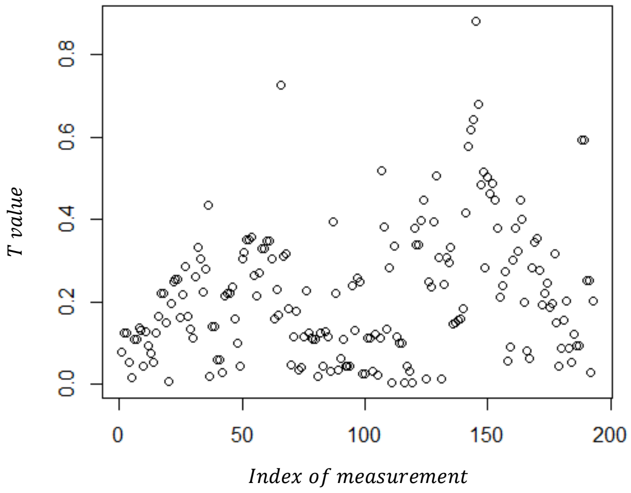


Figure 165: A sample of data points from Triplet 318, half hour intervals

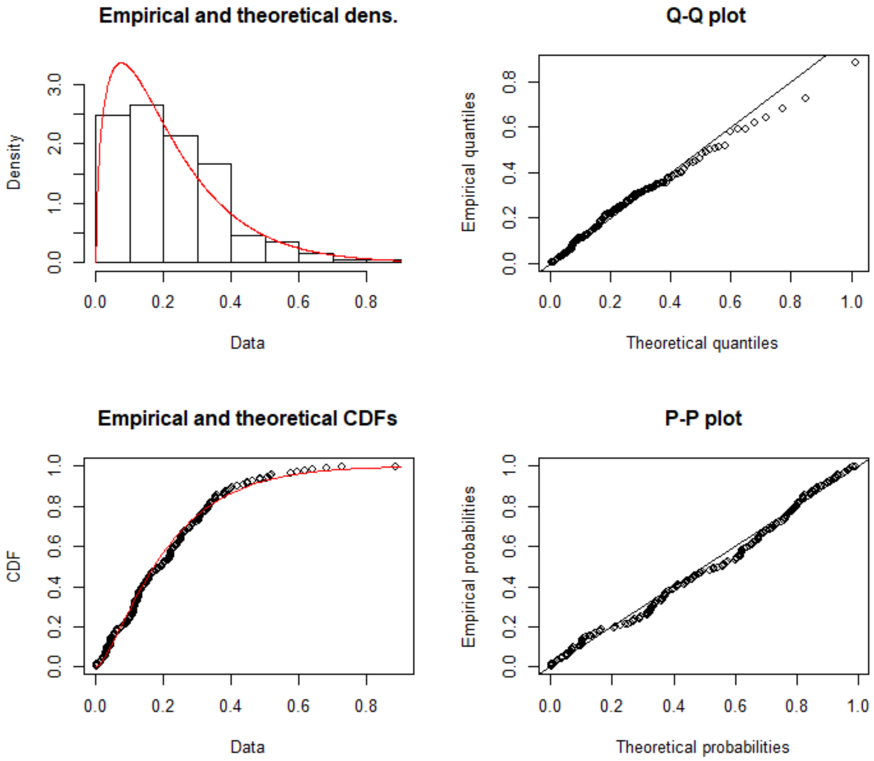


Figure 166: Comparison between the data shown in Figure 165, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.

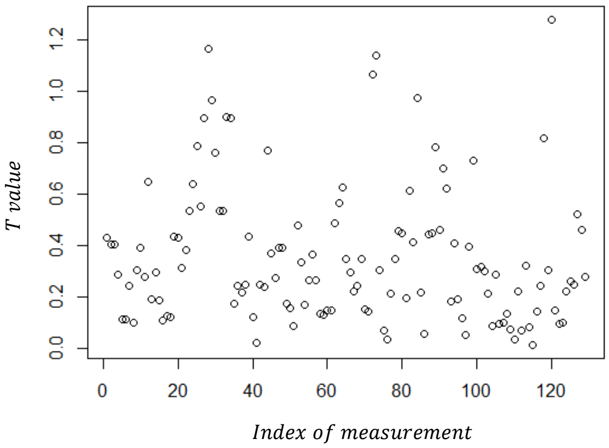


Figure 167: A sample of data points from Triplet 31, one hour intervals

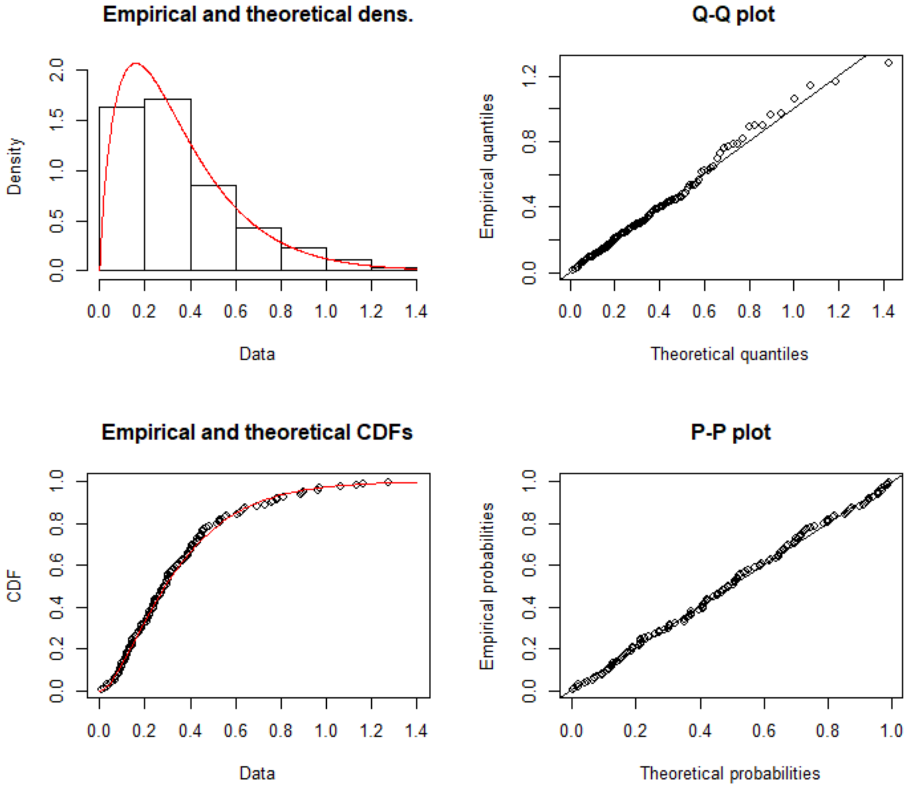


Figure 168: Comparison between the data shown in Figure 167, to the best fitted gamma distribution. Subplots show a histogram comparing the data to the maximum-likelihood fit, Q-Q plot, Empirical and theoretical CDFs and a P-P plot.