'OCEANS AND LAKES'

INTERUNIVERSITY MASTER IN MARINE AND LACUSTRINE SCIENCE AND MANAGEMENT

Quality Control of Sea Level Variation Observations and Tidal Predictions Based on the IOC Sea Level Station Monitoring Facility

Carolina Machado Lima de Camargo

June, 2018

Thesis submitted in partial fulfillment for master degree in Marine and Lacustrine Science and Management

Promotor: Prof. Dr. Karline Soetaert

Supervisor: Francisco Hernandez (VLIZ)

Table of Contents

List of Figures

List of Tables

List of Acronyms

BODC – British Oceanographic Data Centre

BP – Breakpoint Module

GLOSS – Global Sea Level Observing System

go – Global Outlier Filter

IOC – Intergovernmental Oceanographic Commission

IOC-SLMF – IOC Sea Level Monitoring Facility

IOOS - U.S. Integrated Ocean Observing System

MSL – Mean Sea Level

NOAA - National Ocean Service of the United States of America

og – Outlier Gloss Filter

OPPE – Organismo Público Puerto del Estado, Spain

or – Out-of-Range Filter

PSMSL – Permanent Service for Mean Sea Level

QA – Quality Assurance

QC – Quality Control

RMSE – Root Mean Square Error

RSL – Relative Sea Level

RT – Real Time data

RTQC – Real Time Quality Control

SSE – Sum of Squares of Errors

SSR – Sum of Squares of Regression

SST – Total Sum of Squares

UHSLC – University of Hawaii Sea level Centre

VLIZ – Flanders Marine Institute

Abstract

The IOC Sea Level Monitoring Facility (IOC-SLMF) connects about 800 tide gauge stations worldwide. The Facility provides real time (RT) visualization and access to the sea level data of such stations. However, a minimal Quality Control (QC) is applied to the data. QC is important to assure credibility of the data being used and stored. Nonetheless, applying QC to RT data can be a challenging activity, for example the distinction between real outliers and signal fluctuations caused by e.g. tsunamis and storm surges may not be very clear. A good approach for a simple QC of sea level measurements in RT is to compare the observations with a tidal prediction. The purpose of the present work was to obtain tidal predictions based on the data from the 12 selected tide gauge stations, and to use the prediction as a rough QC of the RT observations at the IOC-SLMF. In order to obtain the tidal pattern correctly, a QC procedure was applied in the archived data of the IOC-SLMF. The QC was composed of 5 modules: Correction to mean sea level (MSL) as reference; Stability Check; Outlier Detection; Speed of Change Check; Spike Detection. As the QC method developed here had the purpose of tidal studies, it should not be applied to real-time data. Tsunami and storm surge signals were removed during the QC mainly by the Speed of Change and Outlier Detection modules. On average, the QC flagged 15% of the data, thus detecting and removing the noise of the time series. Regarding tidal predictions, there was no significant difference in using data in minute intervals, as provided by most of the stations, or using data after passing by an hourly filter. Furthermore, because the oldest time series considered had 12 years of data, it is only possible to solve 37 of the harmonic components with high accuracy. The tidal forecast was able to predict the sea level variation for the first months of 2018 with a small source of error, making possible to distinguish unpredicted events, such as tsunamis and storm surges, from the tidal curve.

1. Introduction

Sea level height is an extensively used oceanographic parameter (Merrifield et al., 2009). Besides the growing interest in sea level height due to sea level rise and climate change, the observations are also important for the understanding of ocean dynamics and tides (Rickards & Kilonsky, 1997; Van Onselen, 2000). For practical purposes, the continuous monitoring of sea level is applied in coastal defence projects, such as tsunami and storm surge warning and coastal erosion protection (Holgate et al., 2008). After the extensive damages caused by the Indian Ocean Tsunami in 2004, the need of a global sea level monitoring network was accentuated, leading to the creation of the IOC Sea Level Monitoring Facility (IOC-SLMF) (Merriefield et al., 2005).

The IOC-SLMF connects more than 800 tide gauge stations worldwide, which provide high frequency sea level height in real or near-real time (RT) to the Facility. Tide gauges are one of the oldest and more traditional techniques used to measure local sea level height (Van Onselen, 2000). However, with the advance of technology new measuring methods were developed. Nowadays, in addition to the basic float gauge, pressure, acoustic and radar systems are new technologies used for measuring sea level (IOC, 2006). Apart from the *in situ* measurements of tide gauges, sea level height can also be obtained via satellite altimetry. While tide gauges provide the relative sea level (RSL) in a local scale, satellite altimeters supply an absolute sea level data with global spatial coverage (Vinogradov & Ponte, 2011).

The aim of the IOC-SLMF is to provide data access and visualization with the minimum latency possible, especially if the data is part of a tsunami monitoring system (GLOSS, 2011). To avoid any delays in proving data, minimal quality control (QC) is applied at the Facility. The IOC-SLMF network is hosted by the Flanders Marine Institute (VLIZ), which is responsible for gathering, storing and transferring the RT data. Later on, the information collected if transferred to the University of Hawaii Sea Level Center (UHSLC) and to the British Oceanographic Data Center (BODC), where a comprehensive QC is applied to return research quality data (Woodworth et al, 2017). The Permanent Service for Mean Sea Level (PSMSL) stores the long-term quality controlled mean sea level records.

To ensure credibility and value of any type of data, quality assurance (QA) and QC procedures are necessary (IOOS, 2016). While QA include practices are related to the instrument itself, such as sensor calibration and *in-situ* verification of the measurements, QC involves checks on the collected data (IOOS, 2016). However, performing RTQC can be a challenging activity. First, the computational time of QC needs to be minimal, so that no significant delays result from the process. Another challenge of RTQC is to not remove tsunami signals by detecting out of range data (GLOSS, 2011). A simple alternative to RTQC is to use tidal predictions as a reference to the expected sea level height.

The observed sea level can be decomposed in three main components: Mean sea level (MSL), described as the average of hourly values of sea level measured for at least a year; Tides, characterized by its periodicity, the tidal component is the result of the gravitational attraction and rotation of the system earth-moon and earth-sun; And the meteorological residuals, or surge residual, which are the irregular non-tidal variations of the sea level caused by climatic fluctuations (IOC, 1985). Wind waves also influence the height of the sea surface; however this effect is usually filtered out in the measurements (Weisse et al., 2011). The difference between the measured sea level and the tidal prediction is mainly due to meteorological effects. Thus, comparing the expected curve (i.e., tidal prediction) with real observations can give an idea of the expected residuals, and roughly detect if the deviations are real signals or instrumentation failures.

The purpose of the present work was to obtain tidal predictions with the data from the tide gauge stations, and use the prediction as a rough quality control of the real-time observations at the IOC-SLMF. In order to obtain the tidal pattern correctly, a QC procedure was applied to the archived data of 12 stations from the Facility. The QC developed here comprised 5 modules: Correction to MSL as reference; Stability Check; Outlier Detection; Speed of Change Check; Spike Detection. Once the QC had the purpose of tidal studies, the checks described here should not be applied to RTQC. The results of the present work will be implemented in the IOC-SLMF.

1.1. Objectives

1. Develop QC procedure for the tide gauge data of the IOC Sea Level Monitoring Facility.
2. Obtain a reliable tidal prediction, using the tide gauge data as input of the model.

2. Material & Methods

Data from more than 800 stations is available at the IOC-SLMF. For the purpose of the present work, only 12 stations were selected from the active stations (Figure 1). The selection tried to represent the different levels of data quality from the network, i.e., ranging from very good data series with small gaps and few outliers, to very bad data series with long gaps and a lot of outliers. It also included some stations that have registered tsunamis and/or storm surges according to the news feed on the IOC-SLM Facility website, for example the station “mata”, located in Matarani, Perú.

To solve the main harmonics of a tidal prediction, at least one year of data is necessary (Pugh, 1987). Therefore only stations that had a minimum of year of data until December of 2017 were considered. Furthermore, data obtained by different sensors were selected, to ensure that the quality of the treatment was not determined by the sensor itself. Table 1 gives a brief explanation of why the stations were chosen. The Station Summary in Annex 1

gives some of the relevant information of the selected stations, such as location and creation date.

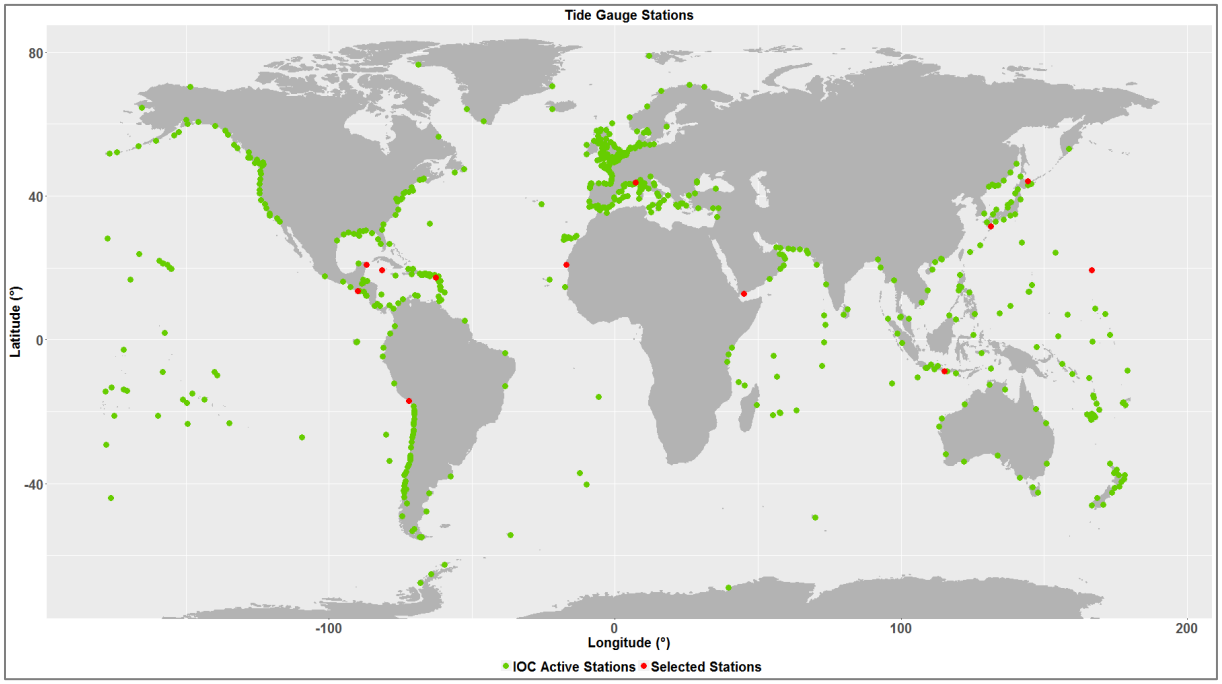


Figure 1. Map of the IOC-SLMF stations. In red are the selected stations for this work, and in green all the active stations of the IOC-SLMF.

Table 1.List of the selected stations, with a brief explanation of why they were chosen.

Station Code	Reason for Selection
“abas”	GLOSS Station with systematic error every year and with short gaps.
“abur”	GLOSS Station with few gaps and few outliers
“acaj”	GLOSS Station with 3 different sensors and 10 years of data.
“aden”	One of the oldest GLOSS stations, with 11 years of data, but with a lot of gaps.
“bass”	Station with a lot of outliers.
“beno”	One of the oldest GLOSS stations, with 12 years of data.
“geor”	Register of a small tsunami wave after earthquake on Jan 10, 2018.
“mata”	Registered waves after earthquake on Jan 14, 2018.
“nice2”	Station with short gaps
“noua”	Station with very few gaps
“pumo2”	Registered small waves after earthquake on Jan 01, 2018.
“wake2”	Tsunami registered after earthquake on Jan 22, 2017.

Access to the database was provided by VLIZ, and the data of the selected stations was downloaded from the IOC-SLM Facility via R Studio Server®. For each station, three data sets were made: one with data since the creation date of the station until 31/12/2017; one with only one year of data (from 01/01/2017 until 31/12/2017); and one with data of 2018 (from 01/01/2018 until 30/04/2018). This allowed for the methods to be tested first with only one year of data, and then applied and tested for the full dataset, optimizing computational time and cost. The dataset of 2018 was used to evaluate the tidal prediction.

Some stations deployed more than one sensor for measuring sea level. For example, station "acaj" in Acajutla, El Salvador has three different sensors: Bubbler, Radar and Pressure. In such case, one file for each sensor was created and treated separately. This resulted in a total of 23 datasets, even though only 12 stations were selected. Information about the different types of sensors can be found in the Literature Review of this present work, submitted apart, and in Manual II, IV and V for Sea Level Measurement and Interpretation (IOC, 1985, 1994, 2016).

Data treatment was performed in the open source software R (R core team, 2018), and divided in 2 main modules, QC and Tidal Prediction, as illustrated in the flowchart (Figure 2) bellow. The following sections explain in details each module.

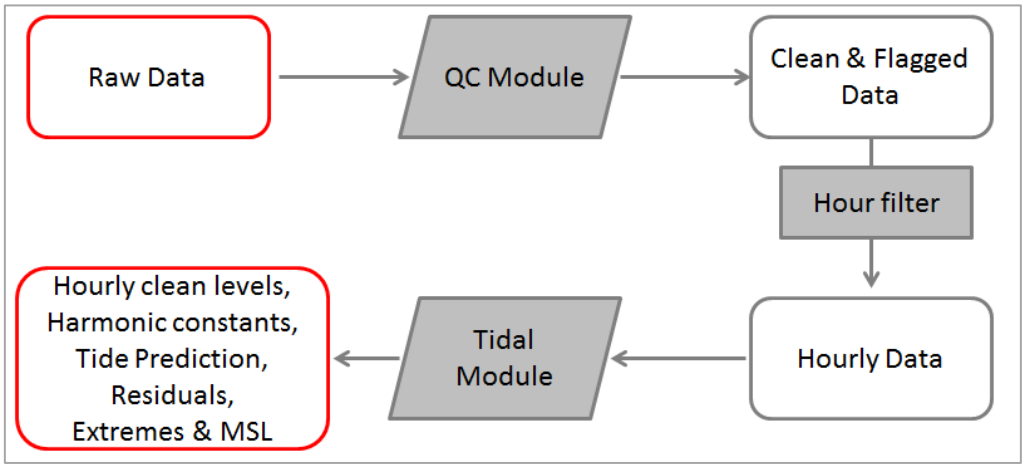


Figure 2.Flowchart of the data treatment. The rounded rectangles refer to datasets, parallelograms to data treatment modules and the rectangle to the hourly filter applied. Highlighted in red is the input and output datasets.

2.1. Quality Control

QC procedures were gathered from the available literature, focusing mainly on tests described in the manuals from GLOSS (2011), BODC (2007) and IOOS (2016). The QC module, illustrated in Figure 3, was divided in 5 sub-modules: Breakpoint detection (MSL Correction); Stability check; Outlier detection; Speed of change check; and Spike detection.

Using R (R core team, 2018), a function was created for each sub-module and filter. A main function for the QC module was created, combining the 5 sub-modules. After each sub-module a flag was created, with 0 for the suspected and removed values, and 1 for values that were kept. Through the entire QC, the raw data set was always kept unmodified. This is important for reevaluation and future use of the data. Annex 2 explains in details each function. Some QC filters used seasonal tolerance values. Considering the predominant effect of the Moon on sea level height and that the tidal spectra can be interpreted according to lunar cycles (Munk & Cartwright, 1966), such tolerance values were calculated according to calendar months and according to lunar months and phases.

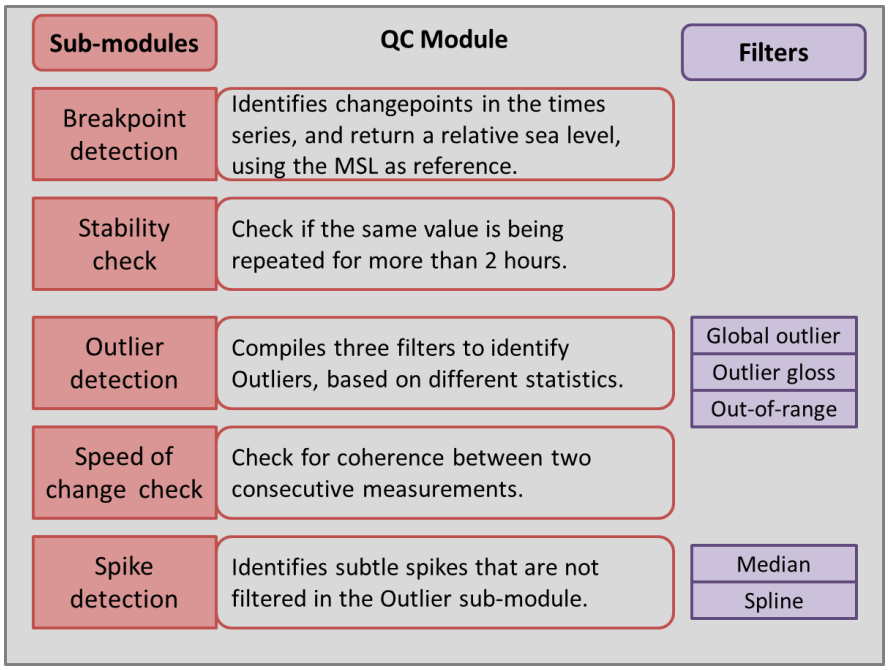


Figure 3.Diagram resuming the Quality Control Module.

2.1.1. Breakpoint sub-module (Correction to MSL as reference)

This sub-module aims to detect change within the time series. A time series is considered to have a change point if, after a determined time *t*, the series can be divided in two subsets (until *t*, after *t*), with different statistical properties, such as mean and variance (Killick & Eckley, 2014). The probability of a time series to have one or more change points increases as the length of the time series increases (Killick & Eckley, 2014).

This step is important for detecting changes in the reference point of the measurements, i.e. datum shifts, and only possible for treating relative sea level observations, such as the ones from tide gauges. For example, consider a fictional station *A* that has sea level data from 2010 to 2016, and which suffered a change in the reference point in 2014 that was not documented in the metadata of the station. As a result, from 2010 until 2014 the measurements ranged from 2 to 6 meters, and from 0 to 4 meters afterwards. If the dataset is treated as only one, then it would return an amplitude of 8 meters, when the actual amplitude is 4 meters. Also the mean, median, standard deviation and 90% percentile would be wrong, and those are important parameters used on the following QC steps. Thus, it is important that the Breakpoint sub-module is applied prior to the other modules.

The identification of change points is made using the R package *changepoints* (Killick et al, 2016). This package incorporates different algorithms and methodologies available to detect multiple change points within a time series, such as the binary segmentation, the segment neighborhood and the PELT algorithm. It also allows to look for the change in relation to the mean, variance or both. More information on the package and the applied methodology can be found in Killick & Eckley (2014).

After change points are identified, the time series is divided according to it and a "mean sea level correction" is applied to each subset. Following the previous example of station *A*, a change point would be identified in 2014. Then the dataset would be subdivided into 2 subsets: 2010-2014, and 2014-2016. Next, the mean is calculated and subtracted from each subset. As a result, both subsets now range from -2 to +2 meters, and are centered on 0. After this "MSL correction" the subsets are joined back, returning a unique time series, with the same length as the initial one. The outcome of the sub-module is a corrected time series, with the MSL as a reference point.

2.1.2. Stability check sub-module

Considering a sinusoidal curve and assuming that sampling frequency is high enough to represent the curve without aliasing (Han et al., 2004), a maximum of two consecutive points can have the same value (BODC, 2007). For a sea level curve, the allowed number of consecutive equal values will be affected by the sampling interval, resolution of the tide gauge and local tidal range. According to the QC manual of BODC (2007), in practice the same value can be repeated for up to 2 hours, any longer periods implies an instrument mal-function and should be flagged.

This sub-module applies a simple test that looks for constant values in a window longer than 2 hours. If found, then the entire period for which the value was constant is flagged, and not considered in the following steps.

2.1.3. Outlier detection sub-module

This sub-module aims to detect and flag values that are beyond established limits for a time series. It combines three different filters that check for outliers, each one based on different statistical properties to the dataset. If an observation fails the test proposed by the filter, then it is flagged and not considered in the next sub-module.

a. Global Outlier filter: A simple and fast filter, based on the global mean (μ) and standard deviation (σ) of the dataset. Being h_t the sea level observation at time t:

$$|h_t - \mu| > 4\sigma$$

b. Outlier Gloss filter: filter to identify outliers according to the IOC Sea Level Monitoring Facility (http://www.ioc-sealevelmonitoring.org/service.php). According to the treatment, an outlier is a value that after subtracted from the median exceeds a tolerance value. The tolerance is defined as 3 times the 90% Percentile minus the median.
   At the IOC-SLMF there is an option to calculate the median in a 12 hours, 24 hours, 7 days and 30 days window. Here, the median was only calculated based on a 30 days window. With h_t as the sea level observation at time t, and $\tilde{x}_m$ and Q_{90} the median and percentile 90 of the month, respectively:

$$|h_t - \tilde{x}_m| > 3|Q_{90} - \tilde{x}_m|$$

The mean and standard deviation were calculated considering calendar months and the lunar cycle (moon passes by the 4 phases).

c. Out-of-range filter: This filter is based on the QC treatment applied by the Organismo Público Puerto del Estado (OPPE) (GLOSS, 2011). It checks if a value is beyond the seasonal limits of a station, defined as 3-sigma from the monthly mean. OPPE defines the seasonal limit is as 2-sigma from the climatological mean for a specific area and month, which is informed at the metadata of each station.
   Here the seasonal limit was modified because no climatological values were available. Being h_t the sea level observation at time t, and μ_m and σ_m the mean and standard deviation of the month, respectively:
$$|h_t - \mu_m| > 3\sigma_m$$
   An extra option is to calculate the mean and standard deviation considering a lunar cycle (moon passes by the 4 phases).

2.1.4. Speed of change check sub-module

This sub-module aims to check for coherence between two consecutive measurements. Considering two consecutive measurements h_1 and h_2, the difference between h1 and h2 cannot surpass a tolerance value. This test is defined in both GLOSS (2011) and BODC (2007) QC manuals. According to both manuals, the tolerance value is defined as:

$$tol = 2.\pi.A.dt/720$$

Where A is the amplitude of the tide and dt the time interval between h_1 and h_2. In addition, the BODC manual provides a corrected tolerance value, which allows for a 20 % increase of the asymmetry in the tidal curve. Here it was used the same approach as in the BODC Manual (2007).

The sub-module applies the test:

$$|h_1 - h_2| > tol * A$$

Where tol is calculated depending of the frequency of the dataset, and A is the amplitude calculated for a given month (with the option to calculate for a lunar cycle).

2.1.5. Spike detection sub-module

This sub-module aims to detect less obvious spikes, which are not identified in the Outlier sub-module. It was based on the description of the spike detection algorithm from OPPE in the GLOSS Manual (2011). This sub-module has two filters:

a. Median filter: calculates a running median on the required window (usually 3). This returns a smoothed curve, which can be used to distinguish out-of-range noises (Wang, 2014).

b. Spline filter: applies a smoothing spline to the curve, with a moving window of 12 hours. The result from the median filter is used to calculate the spline, which gives a better approximation of the measurements. Next, a test is applied:

$$|h_t - h_{ts}| > 3\sigma$$

Where h_t is the level at time t, h_{ts} is the result of the median and spline filter at time t, and σ is the standard deviation. A value that fails this test is flagged as a spike, and removed from the time series.

2.2. Tidal Prediction

According to the Equilibrium Tide Theory, the tidal level of a given location is determined by the time and latitude (Pugh, 1987). However, the observed tide differs significantly from the Equilibrium Tide, once the tidal level is not only affected by relationship earth-moon-sun, but also by the real depths and boundaries of the oceans. Given a long enough record of tidal observations, it is possible to predict the tidal behavior with much higher precision. By means of spectral analysis (Fourier) the tidal curve can be decomposed in a series of regular sine waves (Consoli et al, 2013). These regular waves, also known as harmonic components or constants, make it possible to predict the water level for a given location at any time (Consoli et al, 2013). By analyzing a data set of observed sea level, i.e. tidal analysis, it is possible to obtain these harmonic constants, which then can be used for tidal predictions (BODC, 2007). With a sufficient long series of water level, the true values of the constants can be approximated with high precision (Pugh, 1987). Therefore, the longer the time series being analyzed, the more reliable will be the prediction. As described in Pugh (1987) and Foreman (1977), the harmonic method for tidal analysis is done by least-squares fitting method.

In the present work, the tidal analysis and prediction was done with the R Packages Oce (Kelley et al, 2018) and TideHarmonics (Stephenson, 2016). At least one year of data with hourly values was used as input for the analysis. To obtain hourly values, a simple filter was applied (see hourly filter function, Annex 2) that retrieves the median within the hour. For future work, a Doodson filter should be applied to obtain the hourly value (GLOSS, 2011).

2.2.1. Evaluating prediction accuracy

To evaluate forecast accuracy it is necessary to see how well new data, that was not used to create the model, is fits the prediction (Hyndman, 2011). To obtain a genuine forecast, the last day of data considered in the tidal analysis was 31/12/2017. Hence, the tidal prediction could be evaluated with the dataset of 2018 (from January until April). The root mean squared error (RMSE) was used to assess the prediction accuracy. With f as the predicted value for the observation h, the RMSE is given by:

$$RMSE = \sqrt{mean((f-h)^2)}$$

RMSE is a scale-dependent measure, i.e., the unit of the error is the same as the dependent variable being predicted (Hyndman, 2011). As a consequence, this measure should only be used to compare the forecast for a single time series, and not among data sets (Hyndman & Athanasopoulos, 2018). Other methods can be used to compare different data sets, such as the Mean absolute percentage error (MAPE) which is scale-independent. However, percentage errors are very sensitive to outliers, and are not well defined if the value h is close to zero (Hyndman & Athanasopoulos, 2018). The accuracy of tidal predictions made by NOAA, the National Ocean Service of the United States of America, are also evaluating by computing the residuals and computing the RMSE (NOAA, 2018). Thus, the accuracy of the tidal predictions computed in the present work was evaluated by means of the RMSE.

2.2.2. Source of Sea Level Variation

Sea level observations can be divided in the mean sea level plus variations (IOC, 1985). In turn, the variations can be decomposed into tides, meteorological residues and systematic errors from the instrument. Considering systematic errors to be minimum, the calculated tidal prediction can be used to explain how much of the sea level measurement is affected by the tides and how much by climatic fluctuations. For that, the variation between the observations and the predictions were analyzed with the *Sum of Squares* method, which describes the total variation (SST) as the sum of the variation of the regression (SSR) with the variation of the errors (SSE) (Crawley, 2011):

$$\boldsymbol{SST} \quad = \quad \boldsymbol{SSR} \quad + \quad \boldsymbol{SSE}$$

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Where y and $\hat{y}$ represent the sea level observation and prediction, respectively, in a time series that starts in $t = 1$ until N; and $\bar{y}$ represents the mean of the observations, which in tis case is the mean sea level for the given data set The observation is compared with the respective prediction at time i.

Considering the SST as 100% of the variation, it is possible to see how much of the variation was due to tides (SSR) and how much were from other sources (SSE). The SSR illustrates how much of the observations can be explained by the model (Crawley, 2011).

3. Results

Here, only the main results are showed and discussed. In addition to some metadata information, Annex 2 shows a summary of the results for each station, such as: the MSL and amplitude after the QC; the percentage of removed points by the QC; the RMSE of the tidal prediction; and which set of harmonics had the smaller prediction error. Furthermore, 4 figures are included in Annex 2: A figure with the data before and after the QC, in grey and

green, respectively; a plot of the tidal prediction against the raw observations, in blue and red, respectively; for the first two weeks of 2018; a pie chart illustrating how much of the removed data was flagged by each sub-module; and a pie chart illustrating the source of sea level variation.

3.1. Quality Control

3.1.1. Sub-modules

In order to evaluate each sub-module of the QC, 17 different combinations of treatments (from here on called algorithms) were created, by turning on and off each module. Table 2 gives a brief description of each combination. Each algorithm was tested using calendar months and lunar phases, resulting in a total of 34 treatments that were applied to each data set of the year 2017. After passing by each algorithm, the number of suspected points flagged and removed by the treatment and the final amplitude and mean sea level was calculated and saved in a separate tables.

Table 2. Algorithm's table. The same algorithms were calculated according to the lunar cycle instead of calendar months. For the calendar month, the algorithms are represented with the letter A in front of the number, and with the letter B for the calculations made according to the lunar cycle. In total, 34 algorithms were calculated. When only one sub-module or filter is mentioned to be off (e.g. algorithm 3), it means that all the other sub-modules or filters were on.

Algorithm	Description
1	All sub-modules and filters on
2	All sub-modules on, Global Outlier filter off in Outlier Sub-module
3	Breakpoint sub-module off
4	Stability sub-module off
5	Outlier sub-module off
6	Speed of change sub-module off
7	Spike sub-module off
8	Only Stability sub-module on
9	Only Outlier sub-module on, with 3 filters on
10	Only Speed of Change sub-module on
11	Only Spike sub-module on, with 2 filters on
12	Only Outlier sub-module on, with Global Outlier filter off
13	Only Outlier sub-module on, with only Outlier_gloss filter on
14	Only Outlier sub-module on, with only Out-of-range filter on
15	Only Outlier sub-module on, with only Global Outlier filter on
16	Only Spike sub-module on, with only Median Filter on
17	Only Spike sub-module on, with only Spline Filter on

Figure 4 shows the percentage of data removed for each station after each treatment described in Table 2. While the bars represent the percentage of data flagged, the colors represent different stations. The lower and upper panel shows the result of different stations just for the purpose of scale: the upper panel goes until 100%, and the lower until 80%. During the year 2017, the radar sensor in station *beno* was not working properly, sending a continuously constant signal to the server, which was not detected as a mal-function by the database. As it is possible to see in Figure 4 (pink bars, upper panel), the QC module was very efficient in detecting this mal-function. By analyzing the graph, it becomes clear that the module responsible for this detection is the Stability Check: 100% of the data is removed whenever the stability filter is on (A1, A2, A3, A5, A6, A7, and A8); however, when this filter is off (algorithm A4 and A9-17), only 0.31% is removed from the data.

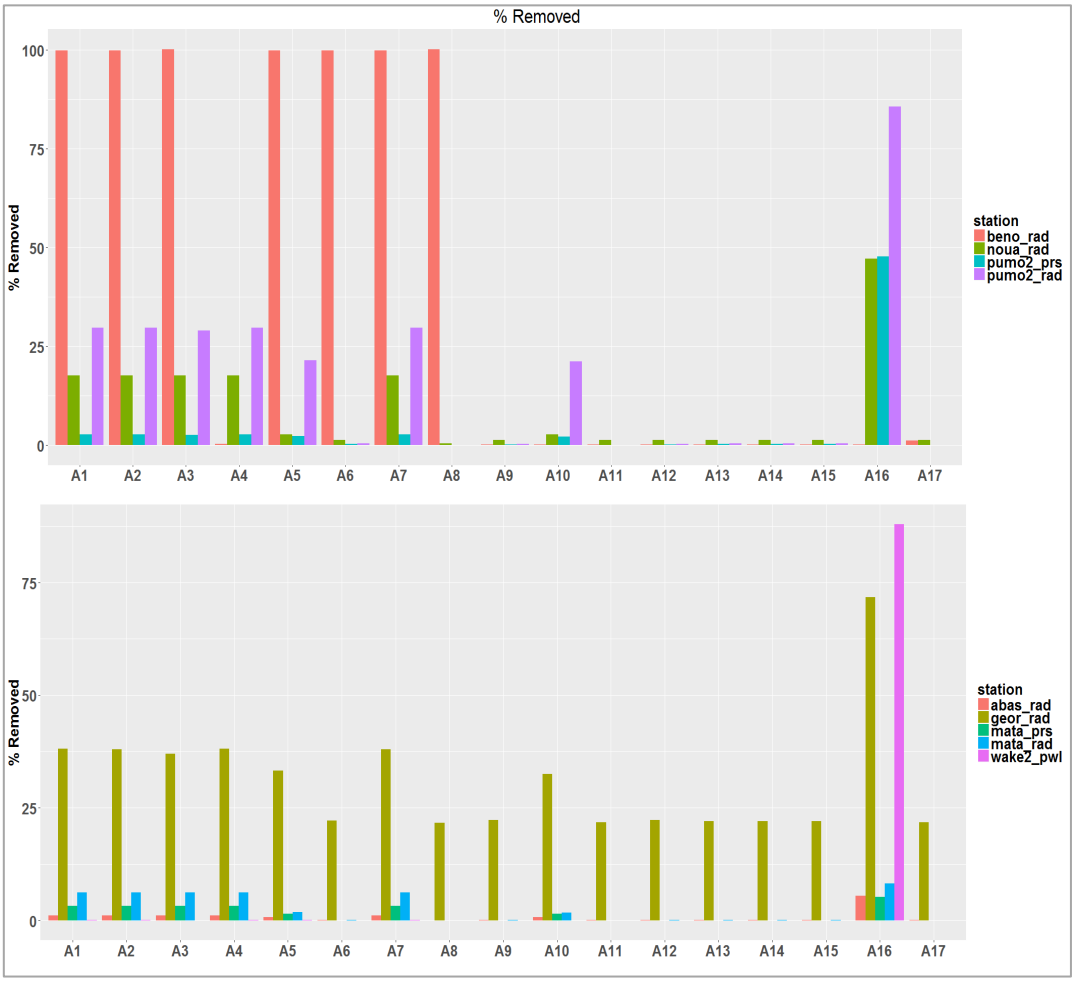


Figure 4. Percentage of the data removed by each algorithm.

The graph in Figure 4 also makes it clear that algorithm A16, which only applied the Spike Module with the Median filter, is clearly faulty: it removes a much higher percentage of data than all the other algorithms for most of the stations. Therefore, it can be assumed that there was a mistake when only the median filter was being used, and that correct data was being flagged and removed. Surprisingly, this bias is not repeated when the Spike Module is using both the median and spline filter. Furthermore, the bars also highlight the efficiency of

the outlier and speed of change checks: in algorithm 5 and 6, when the outlier and speed of change checks are off, the percentage of bad data detected reduces in average to 6 and 8%, respectively.

Figure 4 still illustrates two important results. First, there can be a large difference of quality between different sensors for the same station: the first panel shows the results for station *Pumo* for the pressure and radar sensor (blue and purple bar, respectively), while the same sensors for station *Mata* are shown in the second panel (light green and blue bar, respectively). In both cases, the radar sensor provided much more out of range values than the pressure float. It is important to note though, that this is not enough evidence to state that a pressure float is better than radar. Secondly, the graph also shows the different levels of quality that can be found in the database: while for some stations the percentage of errors detected is around 25-30% (e.g. *Pumo2_rad* in upper panel), others have only 5% or less of bad data (e.g. *abas_rad* in the lower panel). It is also worth to highlight that for some stations, like *Aden*, the quality of the data is very poor. In such cases, even after passing by QC the data is still bad and should be used with caution in further analysis.

While the bar plot clearly demonstrates the efficiency of the outlier and speed of change checks, the other sub-modules are not well represented. However, by considering algorithm 1, with all modules and filters on, as the best treatment, it is possible to estimate the efficiency of the other combinations by comparing them to A1. For that, the amount of data removed by A1 was considered as 100%, and the percentage removed by the other algorithms were calculated in relation to A1. For example, in A4, which had the Stability sub-module off, removed 95% of the same points as A1 for station *acaj_bub*. Thus, it is possible to estimate that the Stability module had an efficiency of 5% in this case. The upper panel in Figure 5 shows the result of this comparison.

Although their contributing percentage is small, an average of 0.2, 0.1 and 0.5% for the stability, spike and breakpoint modules, respectively, the final outcome is better when those modules are on. Figure 5 also shows that the breakpoint module in some cases can lead to a smaller number of points rejected by the QC. This happens especially in stations where tsunamis and storm surges are more frequently. However, increasing the number of rejected points does not necessarily mean that the sub-module is not performing well, since it is possible that it is avoiding false-negative points. In such cases, visual inspection of the data is necessary to make a proper evaluation.

The outlier sub-module was composed by 3 filters: Global Outlier (*go*), Outlier Gloss (*og)* and Out-of-range (*or)*. The efficiency of each of these filters is looked more in detail in Figure 5 (lower panel), which shows the efficiency of each filter by comparing it with A9, when the three filters were on. The red line, which represents the treatment when only the *go* filter was off, is always at zero, meaning that using or not the *go* filter gives the exact same outcome. The green, blue and purple symbols indicate the algorithms when only one of the filters was on. Curiously, the deviation from using all filters and only using one is always the same, no matter which filter is being compared.

Figure 5. Estimation of treatment efficiency, based on the difference between treatment with all modules and treatment with only one. Upper panel: Comparison with different submodules. Lower panel: Comparsion among different filters of Otliwer sub-module.

The second algorithm to be tested was with the *go* filter off, because before introducing the Breakpoint (BP) module, this filter was returning several false-negatives, i.e., it was removing points that were actual good data. But after the introduction of the BP module, this was fixed. Now, the use of this filter does not impact the final outcome, because the other filters of the sub-module are able to detect the same outliers.

Another way of evaluating the various algorithms is by looking at the mean amplitude of the sea level after the QC treatment, as illustrated in Figure 6. The results from *aden_pr2, aden_rad* and *nice2_rad* make it clear that all treatments are quite efficient in removing the extreme values: the raw amplitude of 80 meters decreases to more realistic values after the QC. The figure above also highlights the stations that were sending a stable signal throughout the year: the bubbler sensor in station *Acaj* and the pressure sensor in station *Bass* have amplitude of 0 meters, because the same value was sent all the time.

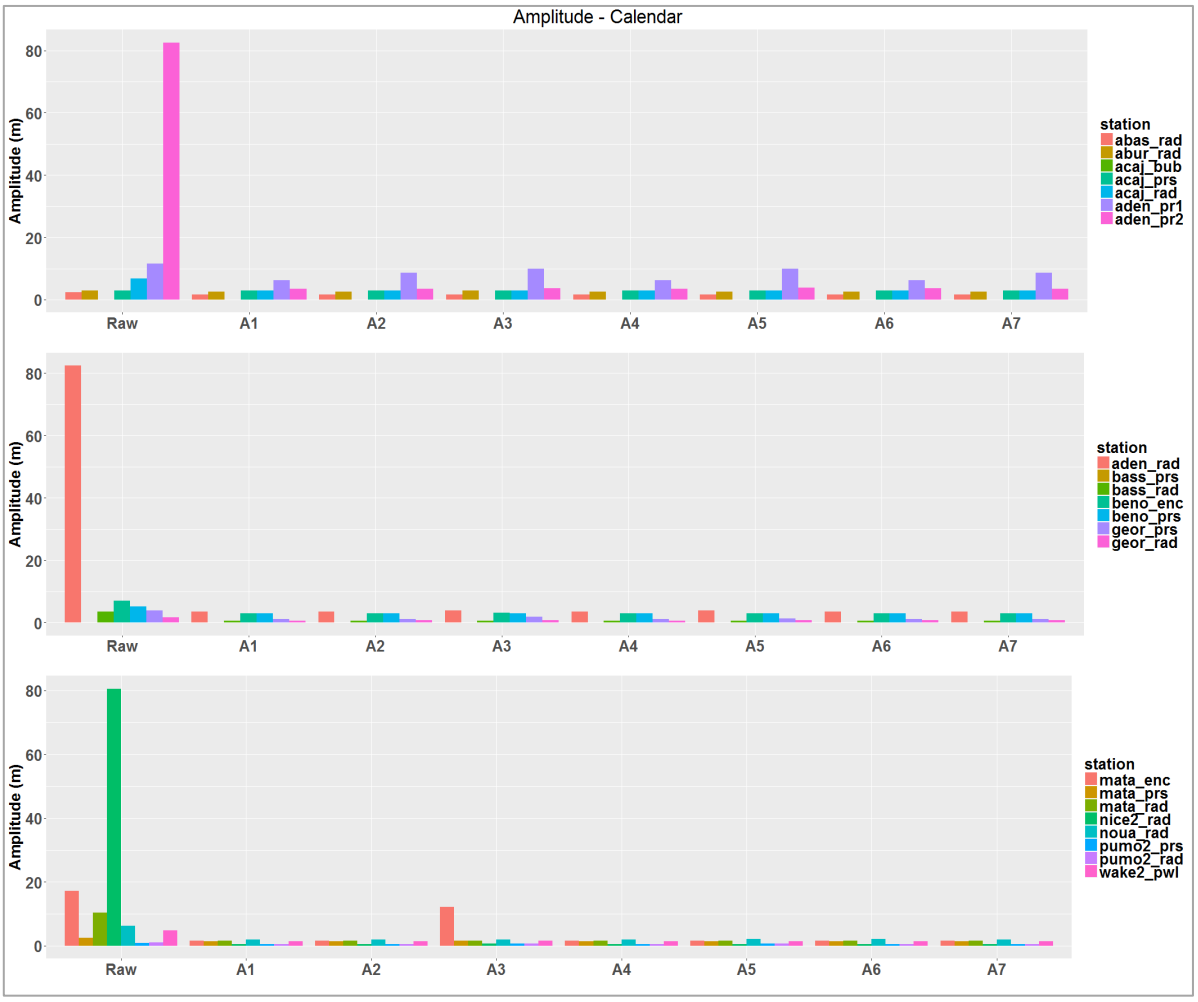


Figure 6. Amplitude of the data sets. On the left there's the raw data (before QC), and treatments from A1-A7. Colors indicate the stations.

By looking at the amplitude variation of the sensor *enc* from station *Mata* (salmon bar in lower panel of Figure 6), the importance of the breakpoint module becomes clear. While all algorithms were returning an amplitude of a couple of meters for that data, when the BP module was off (A3) the amplitude was almost 20 meters. As shown in the Figure 7 (left panel), this sensor had a strong signal fluctuation in 2017. For the first months of 2017, the sea level curve was centered at 16 meters, with amplitude of 1.3 meters; but after July of 2017, the curve was centered around 6 meters, with amplitude of 1.4 meters. If the annual amplitude is calculated without the BP correction, it returns a false value of almost 20 meters. However, if the BP module is applied, and the two “blocks” of data are offset to their average, then the curve evolves always around zero, and the fake amplitude artifact is removed. This problem can be more drastic when a longer time series is considered (Figure 7, right panel).

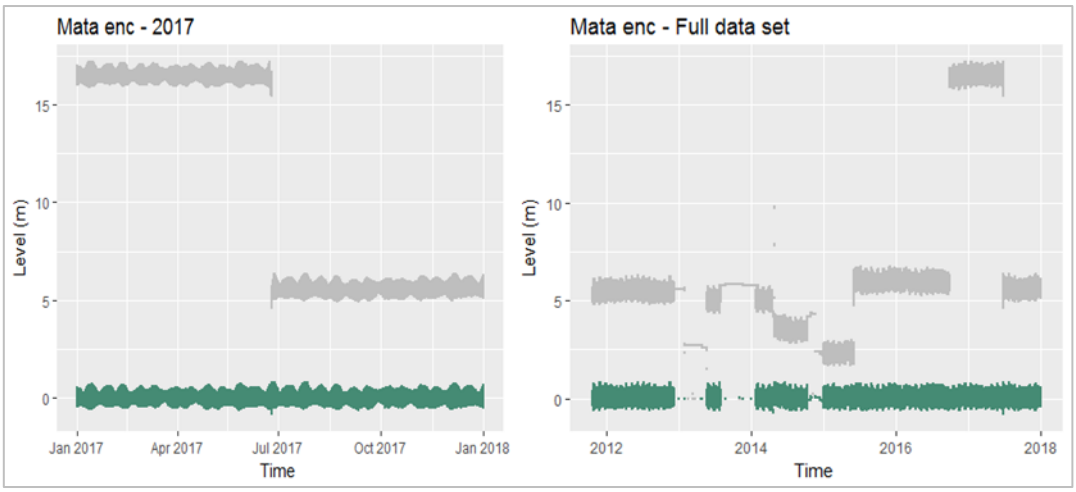


Figure 7. Example of Breakpoint module for station Mata. Left panel: only 2017 data set. Right panel: full data set

The best treatment was considered the algorithm 1, which applied all the modules and filters. Once the *go* filter is redundant, and treatment 1 and 2 give the same result, algorithm 2 can also be considered the best approach. Figure 8 below shows how much each sub-module contributed to detecting bad data in algorithm 1. While the left panel shows the absolute number of removed data by the treatment, the right panel shows the percentage flagged by each sub-module. It becomes clear that the importance of each sub-module is dependent of the data set itself.

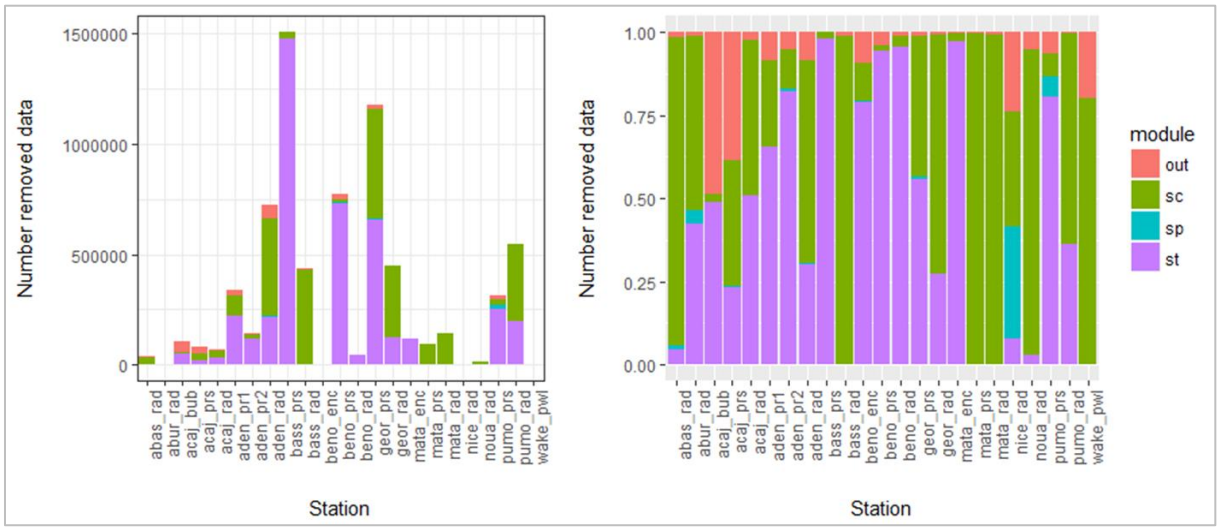


Figure 8. Contributing portion of each sub-module in A1, where all sub-modules were active.

While in some cases the stability check (purple bars, Figure 8) has a big role in detecting wrong data, in other cases this is mainly the result of the speed of change check (green bars, Figure 8). The contributing percentages of these two checks are, in average, very similar, with 44.42 % for the stability check and 45.22% for the speed of change. Despite the smaller contribution of the Outlier Detection module, which on average detected only 8% of the removed data, the importance of this filter should not be undermined. Not using the Outlier module would have a big impact in the outcome of the QC. The Spike module, which aimed

to find smaller variations that were not detected by the Outlier module, played a very small role in the QC, detecting only 2% of the removed data.

The data set of 2017 for sensor *rad* in the station *Geor*, located George Town, Cayman Islands, is a good example of the part played by each module, and it is illustrated in the Figure 9. While the left upper panel (green) shows the result of treatment 1, the outcome of applying only the Stability Check, Speed of Change Check and Outlier Check are shown in the right upper panel (purple), left lower panel (yellow) and right lower panel (red), respectively. This example shows how it is important to apply all the tests. The flat line from mid-April to Jul is only detected by the Stability Check. However, the flat signal that continues through the year is only flagged by the Speed of Change. In the year of 2017, the region of this station was hit by several storms, which impacted the data being recorded by the tide gauge. The results of these storms are the extremes peaks that can be seen in the figure. The Outlier module was responsible for removing the most extreme signals. The figure also shows the importance of the breakpoint module (applied only in A1, left upper panel).

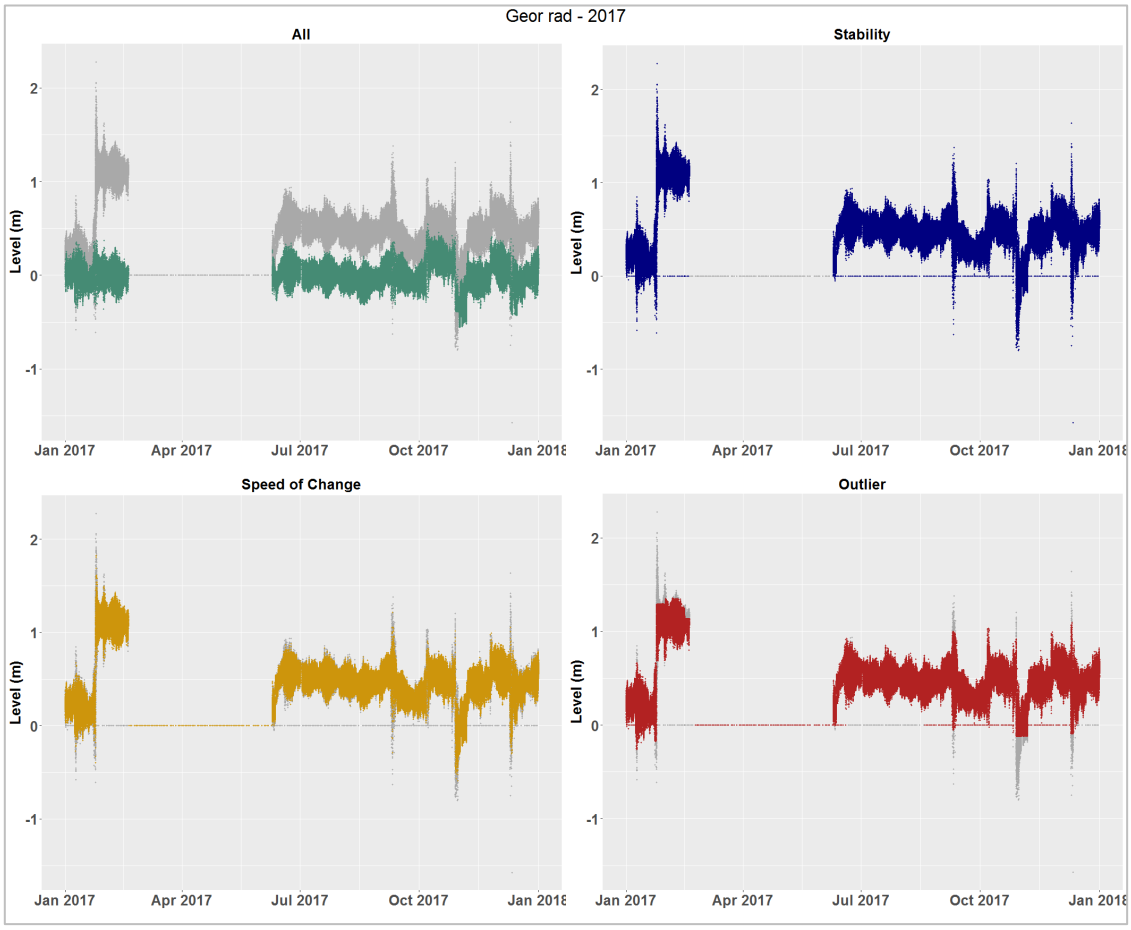


Figure 9. Data of *Geor* with examples of the role of each sub-module. In all panels, raw data is represented in grey. Left upper panel, green represent the result of A1. Right upper panel: purple represent the result of only the Stability Module. Left lower panel: yellow represent the results of only the Stability module. Right lower panel: red presents the results of only the Outlier Module on.

3.1.2. Calendar months X Lunar phase

As explained in the methods section, for some sub-modules, monthly values, such as mean and amplitude, need to be calculated. QC procedures found in literature (GLOSS, 2011; BODC, 2007) usually use climatological values, i.e., values calculated based on a long-term time-series, in the filters. However, the present work focused on using only the data itself to compute these values. As a consequence, the seasonality pattern of such measures could be under-represented. An attempt of reducing some of these biases was to insert the "lunar month/ phase" artifact to the calculations. It was expected that by calculating the amplitude, per example, according to each lunar cycle, the differences between spring and neap tide would be better represented, as well as the differences between the tides in summer and winter solstice and in spring and fall equinox.

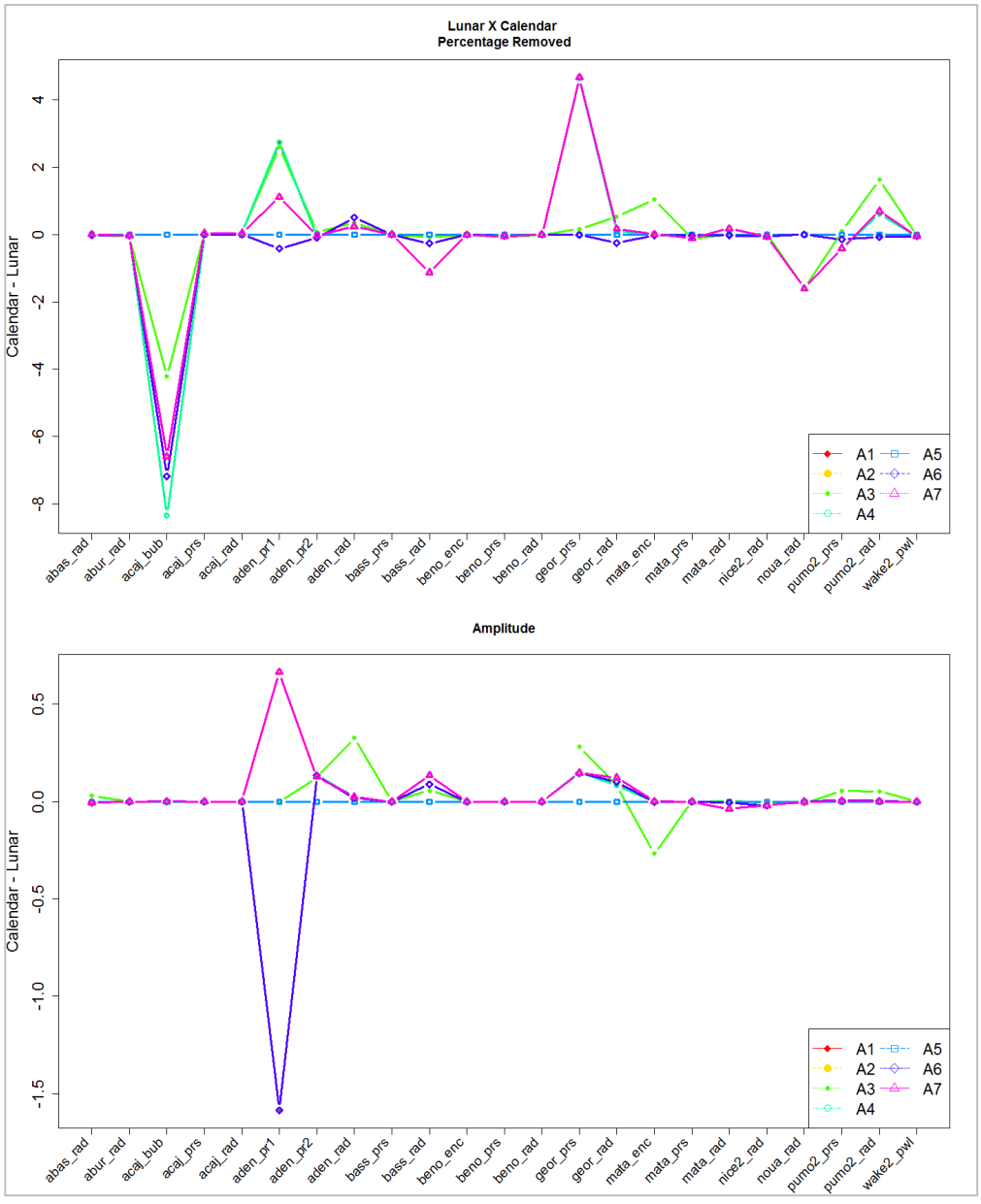


Figure 10. Comparison between QC using Calendar months of Lunar phases. Upper panel: Percentage Removed at each QC treatment. Lower panel: Amplitude after each QC treatment.

Figure 10 shows the differences between applying the QC in relation to the lunar phases and in relation to calendar months. The discrepancy is higher when looking at the percentage of data removed (upper panel) instead of looking at the amplitude (lower panel). The divergence between the algorithms will also depend on the station itself: If the station has a wide tidal range, and the differences between spring and neap tide are large, then using lunar phases instead of calendar months will have a more significant impact in the result of the QC. The main source of variation is the Outlier module. As shown in the Figure 10, when such module is off (algorithm 5) there is no discrepancy between the lunar or calendar months.

Surprisingly, the disparity is stronger in algorithm 3 and 6, which had only the Stability Check and Speed of Change off, respectively, than in algorithm 1, which had all modules and filters on. Only by visual inspection is possible to see and evaluate the difference between the treatments. Figure 11 shows an example of this comparison. The left panel has the result of QC according to calendar months, and the right panel the QC according to lunar phases. Although the difference is subtle, the QC according to the lunar phase was able to remove more of the "flying" points.

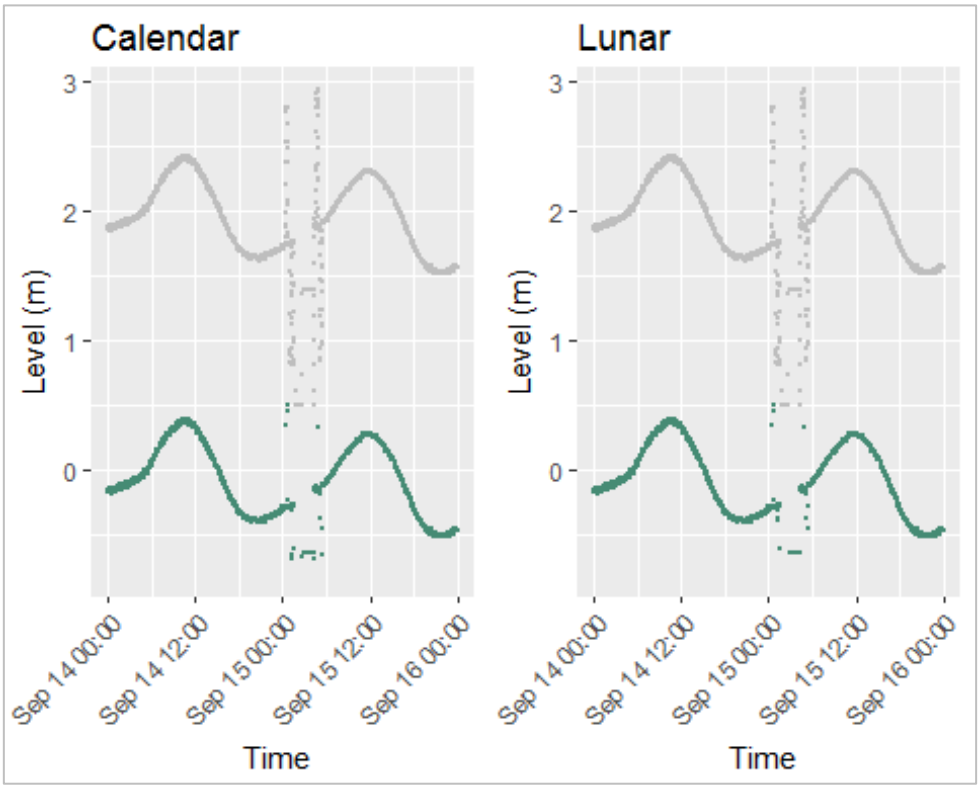


Figure 11. Example of difference between QC using calendar months (left panel) and lunar phases (right panel). In both panels grey represents the raw data and green the clean data.

3.2. Tidal Prediction

Usually, tidal analyses are performed with hourly values of sea level data (GLOSS, 2011). Here, we tested doing the analysis with high frequency data (minute interval) and with hourly values. Figure 12 shows the tidal prediction for station *Mata* for the month of

January, 2018, using minute data and hourly data, in the upper and lower panel, respectively. Both predictions had a good fit with the observed data (red dots), and a small RMSE. However, it becomes clear that the minute-based prediction has stronger residuals, what is expected, since phenomas such as tsunamis, seiches and storms also have a higher frequency (GLOSS, 2011). On the other hand, these high frequency signals are not reflected in the tidal pattern, and therefore there is no improvement in the prediction by using minute data. Yet, the computation time and cost increases significantly with the length of the data set being used in the analysis. Therefore, the prediction was only carried with hourly values.

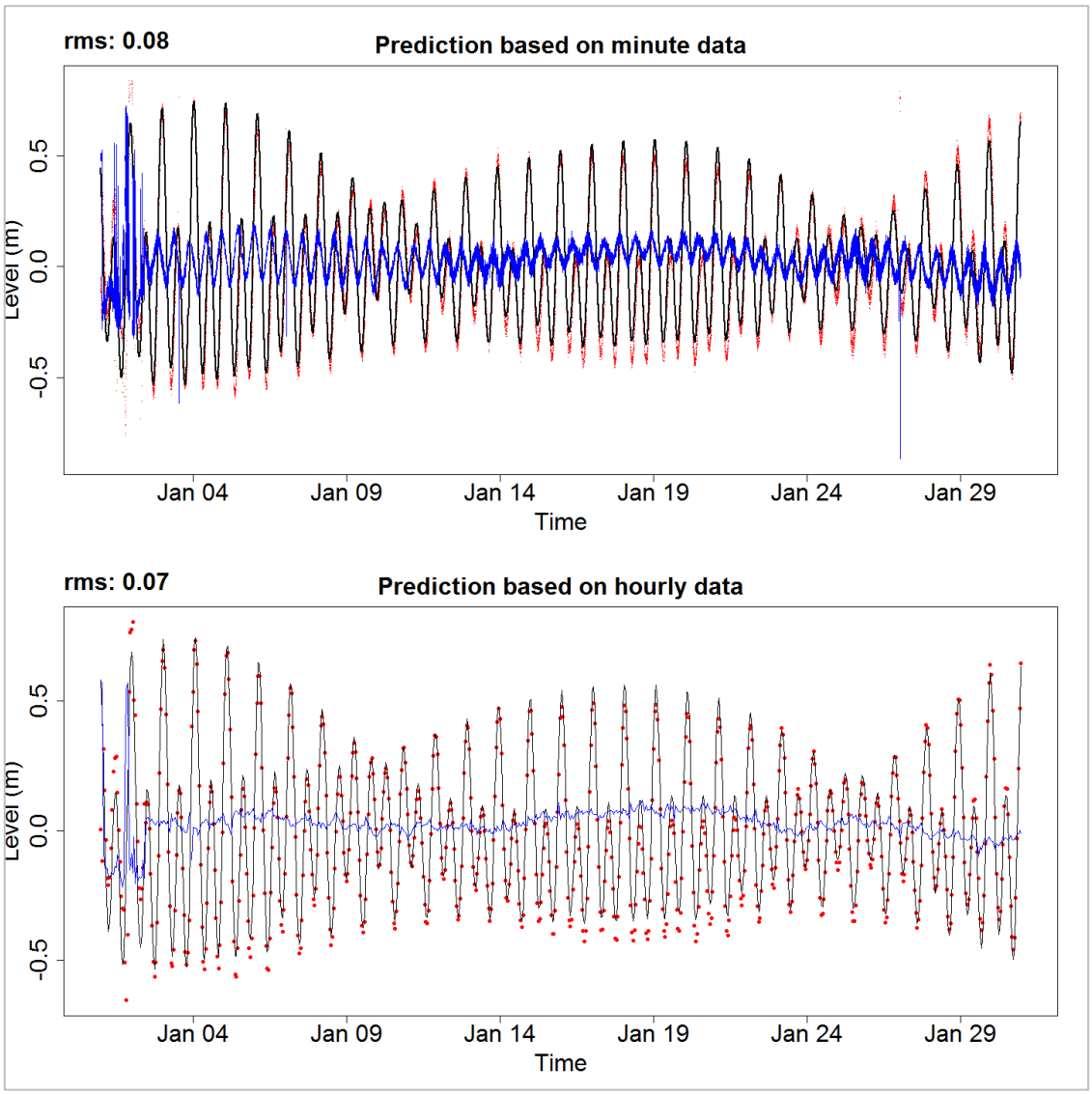


Figure 12. Tidal prediction for station *Mata*. Upper panel the prediction made with minute values, and lower panel the prediction made with hourly values. In both panels: black line is the tidal prediction, red is the sea level observations, and blue is the residual.

The tidal prediction was tested for each station with a set of 7, 37 and 60 harmonics with the package TideHarmonics, and with a set of 7, 37 and 69 harmonics with the package Oce, for the full data set and with only 1 year of data. The sets of harmonics used were the standards provided by the packages. In total 276 tidal predictions were made. A list with the names and brief description of the 37 main harmonics can be found in Annex 3. Figure 13 shows the

RMSE of the predictions with the different sets of harmonics and for each R Package, made with full data set of each station. Both packages had good and similar performances; however, for 93% of the cases, the prediction made with TideHarmonics had a smaller error. While the TideHarmonics package had a mean RMSE of 0.1300, 0.1304 and 0.1306 for the sets of 7, 37 and 60 harmonics, respectively; the Oce package had a mean RMSE of 0.1352, 0.1360 and 0.1360 for the sets of 7, 37 and 69 harmonics. Thus, further analysis was carried out only with the TideHarmonics package.

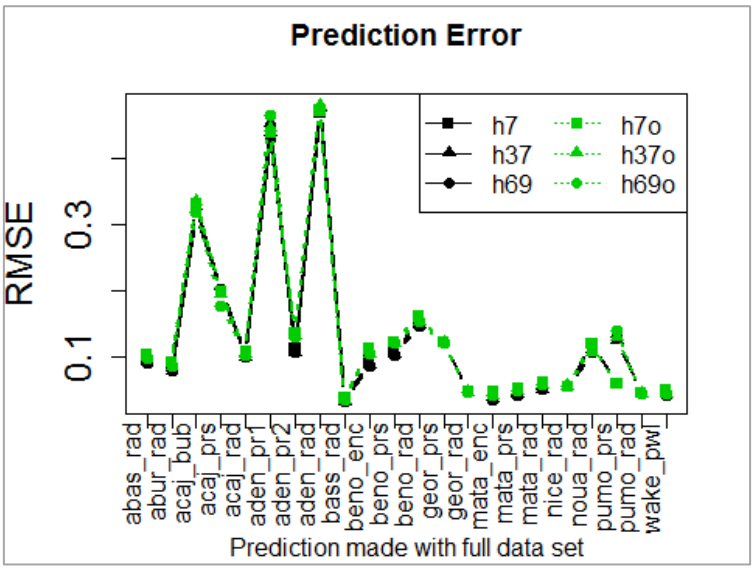


Figure 13. Prediction Error for each station. Green: prediction made with package Oce. black: prediction made with package TideHarmonics. The symbols represent the different sets of harmonics used for prediction.

Comparing the RMSE of the predictions calculated with the different sets of harmonics (Figure 14) it is possible to see that for most cases, the prediction using only the main 7 harmonics gives a higher error. That is expected, as the 7 harmonics (blue bars) cannot reproduce the sea level height with high precision, but only the main pattern. Following the same reasoning, one would expect that the predictions made with a set of 60 harmonics (green bars) would have the smallest residual. However, that is not always the case. With exception of few cases like *Abas*, *Beno_enc* and *Wake_pwl*, the prediction made with 37 harmonics (pink bars) had smaller or equal RMSE as the prediction with 60 constituents. This can be because of the period and quality of the series used in the analysis.

The graph in Figure 14 is ordered by the age of the station. The youngest station was *Nuoa*, with only 2 years of data considered, and the oldest was *Beno*, with 12 years of data. The stations that performed better with the 60 constituents had at least 5 years of data. Besides the length of the time series, another factor influencing the tidal analysis is the quality of the data itself. For example, even though the analysis of station *Aden* was made with 11 years of data, there was no improvement in the prediction when 60 constituents were used instead of 37. That is probably the effect of the poor quality of the data from *Aden*, even after passing by QC.

To see how much of the tide gauge observation could be explained by tides, a *Sum of Squares* analysis was performed. The result of such analysis is shown in Figure 15.

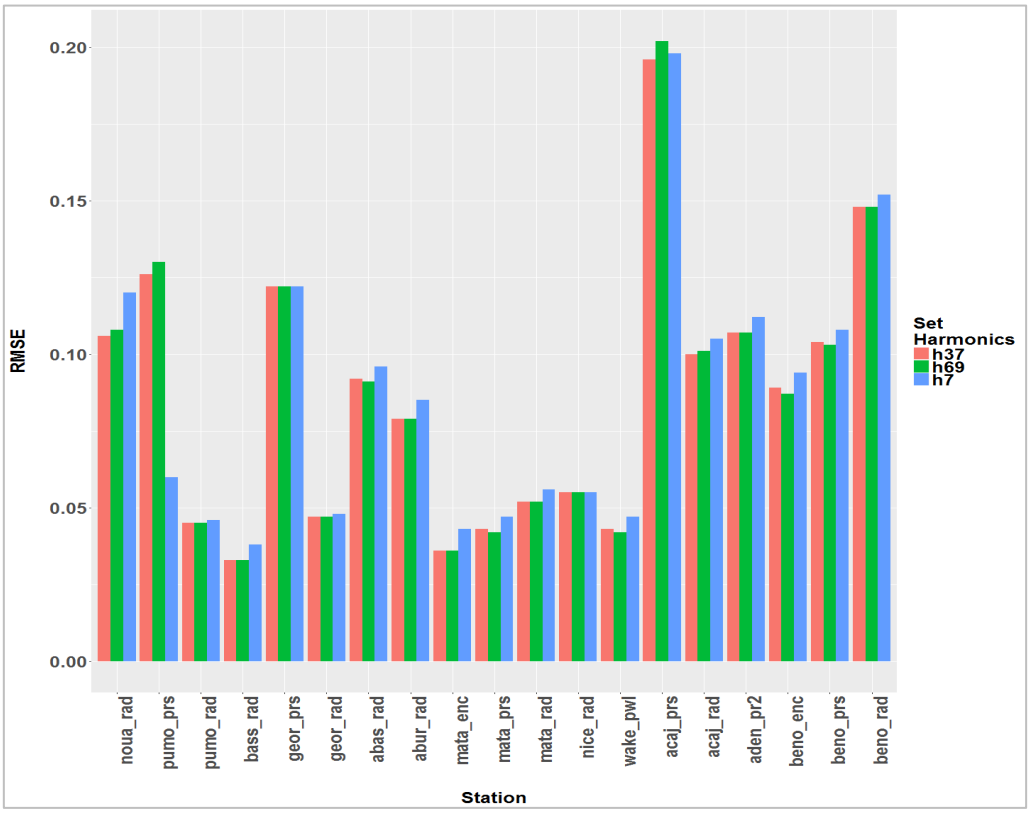


Figure 14. Histogram of prediction error for each station. x-axis is ordered by the age of station: from the left, station *Noua* with only 2 years of data to the right station *Beno* with 12 years of data. Stations Acaj_bub, aden_pr1 and aden_pr2m with RMSE around 0.3, 0.4 and 0.45, respectively, were omitted from the graph for scale purposes.

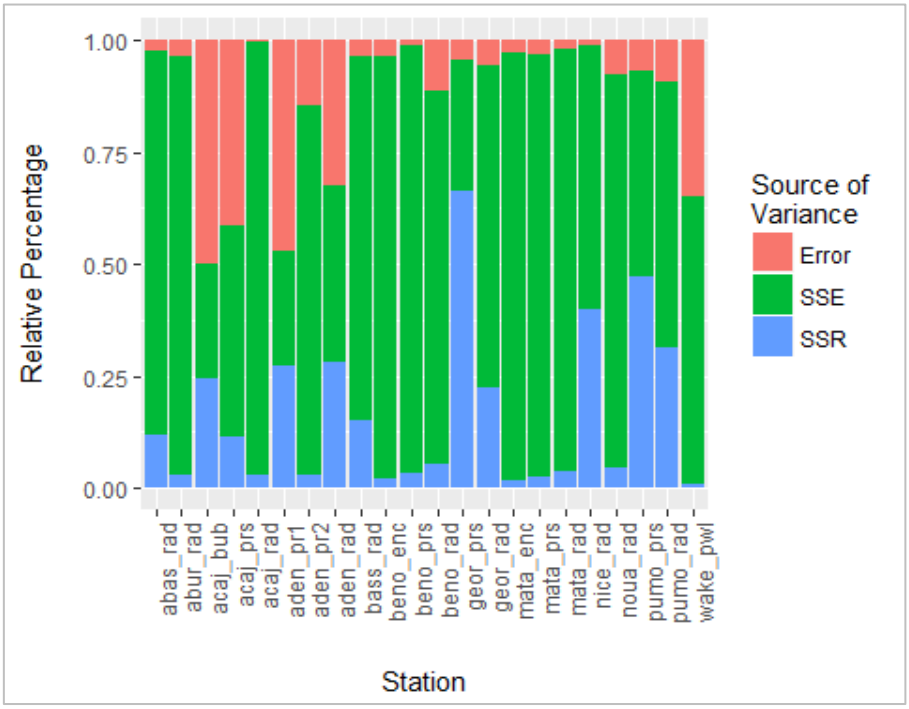


Figure 15. Breakdown of the variation between the sea level observations and the predictions.

As described in section 2.2.2, the total source of variation can be decomposed in the variation explained by the model (SSR) and variation due to errors between the predicted and observed value (SSE). Here, SSR is the variation explained by tides, and SSE is the variation due to climatic factors. However, when adding SSR with SSE not all variation could be explained. This means that there are other outside factors influencing the tide gauge observations, probably measurements and/or data processing introduced errors. Thus, the equation given in section 2.2.2, can be rewritten as:

$$\boldsymbol{SST} \quad = \quad \boldsymbol{SSR} \; + \; \boldsymbol{SSE} \; + \; \boldsymbol{Error}$$

Where the Error is given by:

$$\boldsymbol{Error} \; = \boldsymbol{SST} - (\boldsymbol{SSR} + \boldsymbol{SSE})$$

As shown in Figure 15, most of the sea level observation can be explained tides. The higher residual variances were in stations from stormy areas, like *Pumo*, *Mata* and *Geor* .The *Error* was more significant in cases with poor quality of data, such as for the sensor *bub* and *prs* in station *Acaj* and sensor *pr1* in station Aden, giving a probable origin of the error.

Examples of the tidal prediction for each station can be found in the second plot of Station Summary in Annex 1.

4. Discussion

4.1. Quality Control

The IOC-SLMF provides sea level data from tide gauges all around the word (GLOSS, 2011). While the Facility focusses on maintaining a continuous, fast and smooth communication between the stations, the database and the scientific society, it is also important to provide some kind of QC to ensure the credibility and reliability of the data (Unesco, 1993). However, applying real-time quality control is a challenging activity, which goes beyond the aim of the Facility. A simple approach to give some reference to the real time observation is by having a tidal prediction for each station. However, to calculate the tidal prediction based on the tide gauge data, it is necessary to first remove as much error as possible from the sea level series.

According to GLOSS (2011), two types of errors can be present in the sea level data: random errors, such as electronic noise from the instruments and problems with the sensor and in communication; and systematic errors, such as changes in the instrument itself, in the instrument location, or even changes in the environment surrounding the station. These kinds of error can induce strong discontinuities in the sea level data and produce trends in the tidal pattern. Therefore, it is of extreme importance to minimize any source of error before doing the tidal prediction.

The QC developed here was composed of 5 main steps, and based mainly on the procedures described in the QC manuals of GLOSS (2011) and of BODC (2007). In the first step, change points along the time series were identified. From each change point, the data set was divided in subsets, which were then subtracted from its mean value. That resulted in an artificial offset based on the mean. After passing by this step, the series was always evolving around 0. This module was very efficient in correcting *datum shifts* in the time series, which are sudden jumps in the reference level (BODC, 2007). These shifts can be caused either by natural phenomenon, such as earthquake, or by improper maintenance of the station. The second step, a stability check, was very successful in removing errors caused by failure in the instrument or in the communication. The third and fourth steps, the outlier and speed of change checks, both aimed to remove random errors caused by fault in the measurements. These were the steps mainly responsible for removing "errors" resulting from a tsunami or a storm.

The fifth and final step of the QC was the Spike detection module. The QC manual of GLOSS (2011), briefly describes an algorithm for the detection of spikes which is based on the fit of a spline with a moving window. While such algorithm is considered by OPPE as the main component of the QC-module and that is responsible for detecting 95% of the wrong values of a very bad series (GLOSS, 2011), the attempt of reproducing such treatment here was not that successful. The Spike module of the present work flagged, in average, only 2.11% of data points. This tends to suggest that there was something wrong with the Spike Detection Module here. Although this module was not negatively affecting the QC, by removing data that should not be removed, it also was not effective. Future work should focus on a better implementation of this module.

Despite the small differences among the QC done according to calendar months and according to lunar cycle, considering the Moon in calculating tolerance values can be important in regions with stronger tidal range. At some locations, the tidal amplitude can vary significantly between neap and spring tides. For example, Schettini et al. (2011) showed that the tidal amplitude for Fortaleza, Brazil, varies from 1.5 meters during neap to 2.8 meters in spring tide. If only one of those values would be considered along the QC of a tide gauge in such location, then either the treatment could be too restrictive or not restrictive enough, by using the amplitude of neap or spring tide, respectively. Therefore, it is recommended that QC is done considering the lunar cycle.

Overall, the QC treatment was effective in returning a cleaner series for the tidal prediction. In order to validate the QC done here, the clean series for stations *Abas*, *Abur* and *Wake* was compared with clean hourly values provided by UHSLC (Caldwell et al, 2015) and yearly means available at PSMSL (Holgate et al., 2013). However, such comparison proved not very efficient. The main problem was that the observations have different reference levels: for example, while the mean value of the series for stations *Abas* was 1.5 m for UHSLC, it was 7.34 m for PSMSL. Thus, to allow comparison, each data set was subtracted from its mean

value, returning a series around 0. The differences between the values obtained in the present work and those available at UHSLC and PSMSL platforms were tested with ANOVA for the MSL, and t-test for the median and amplitudes, in the case if normality and homogeneity of variances were respected. If one of the prior assumptions were violated, then a Kruskal-Wallis test was performed. The compared values and result of the tests can be seen in Table 3.

Table 3. Yearly mean, median and amplitude calculated from hourly values of the present work (IOC), and of UHSLC. For PSMSL, only yearly means are available. Normality and homogeneity of variances were tested with Shapiro and Levene's test, respectively. P-values for the Anova, t-test and Kruskal-Wallis (K-W) test are also shown for each comparison set (mean, median an amplitude). Highlight values indicate when the null-hypothesis could not be rejected.

	MSL			MEDIAN		AMP		
Year	IOC	PSMSL	UHSLC	IOC	UHSLC	IOC	UHSLC	Station
2012	0.000317	-0.0253	0.0162	-0.0173	-0.0037	1.51	1.59	Abas
2013	-0.000146	0.00866	0.533	-0.013	0.0363	1.54	1.76	
2014	-0.000627	0.01266	0.0571	-0.0163	0.0363	1.51	1.67	
2015	-0.000177	-0.0003	0.044	-0.0185	0.0263	1.56	1.74	
2016	0.000435	0.01166	0.0559	-0.0126	0.0463	1.45	1.55	
2017	0.00168	-0.00733	0.0374	-0.00855	0.0263	1.5	1.62	
Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
H. Var	Yes			Yes		No		
Anova	2.10^{-16}			2.10^{-16}		0.003		
t-test	2.10^{-11}			6.10^{-12}		0.007		
K-W	0.45			0.27		0.44		
2012	0.000576	0.023	0.126	0.0381	0.159	2.43	2.78	Abur
2013	0.000884	-0.031	0.0519	0.0377	0.0789	2.42	2.53	
2014	0.000896	-0.008	0.0748	0.0237	0.0989	2.52	2.89	
2015	0.000673	-0.017	0.0659	0.0276	0.0889	2.35	2.71	
2016	0.000616	0.033	0.115	0.0277	0.139	2.35	2.6	
2017	0.00548		0.0742	0.033	0.0989	2.3	2.82	
Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
H. Var	Yes			Yes		Yes		
Anova	5.10^{-6}			0.0001		0.0004		
t-test	0.0009			0.0008		0.001		
K-W	0.45			0.35		0.44		
2012	0.000548	0.0244	0.0369	0.00092	0.0297	1.09	1.11	Wake
2013	-0.00298	0.0174	0.0956	-0.00982	0.0917	1.2	1.32	
2014	-0.00123	-0.0316	0.0468	-0.00393	0.0437	1.2	1.32	
2015	-0.00057	-0.0226	0.0553	-0.00526	0.0487	1.2	1.24	
2016	0.000016	0.0124	0.091	-0.00246	0.0867	1.22	1.26	
2017	0.00217		0.0984	-0.00101	0.0957	1.32	1.5	
Normal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
H. Var	No			No		Yes		
Anova	2.10^{-5}			0.0001		0.18		
t-test	0.0014			0.0017		0.18		
K-W	0.45			0.44		0.44		

With exception of the Amplitude of station Wake, whenever a more restrictive test was used, i.e. Anova and t-test, the tests indicated that the series being compared are different. On the other hand, when the robust Kruskal-Wallis test was used, the test indicated that the data sets are similar. The fact that the data sets are considered statistically different does not mean that the QC applied here is ineffective. While the tests from the present work were more robust and aimed to be time-efficient, UHSLC applies a more complete QC which returns research level data. Furthermore, the differences amongst the values obtained here and the ones used in the comparison may derive from the method applied to obtain the mean values.

The present study was done with high frequency sea level observations, which had a sampling rate of 1 or 5 minutes. To obtain hourly values, a simple median filter was applied. In turn, the hourly values were simply averaged over time to obtain the daily, monthly and yearly means. While for the data available at the UHSLC portal, hourly, daily and monthly values were obtained by: hourly spot reading, a 119-point convolution filter applied to the hourly data, and a simple average of all daily values, respectively (Caldwell et al, 2015). The averaging method will influence not only the accuracy of the data, but also which information is lost during the filter (Van Onselen, 2000). For example, a low-pass filter, like Doodson X_0, eliminates the diurnal and semi-diurnal tidal constituents, which is useful for studies of mean sea-level and of sub-tidal meteorological effects (Pugh, 1987; Van Onselen, 2000). On the other hand, the Doodson X_F filter is commonly recommended to obtain hourly values for tidal analysis (GLOSS, 2011).

There is still room to improve the QC procedure described here. The quality of the data is directly dependent to the sensor technology, thus developing a QC methodology that can be used for every sensor can be very challenging (IOOS, 2016). The ambition of looking for a QC that can be applied to all the stations, regarding the type of sensor, became a constrain in the treatment. The QC could be more efficient if aimed for only one station or one type of sensor. Future work should look for ways of implementing the different characteristics of the stations in the QC, apart from improving the Spike detection module.

The QC of the present study can be improved by incorporating the results of the tidal analysis. After obtaining the harmonic constituents for a station, it is possible to classify the regional tide by means of the tidal form. The tidal form number F is given by the ratio of amplitudes of the main semidiurnal and diurnal components (Zhu et al., 2015):

$$F = [A_{O_1} + A_{K_1}]/[A_{M_2} + A_{S_2}]$$

where the numerator is the sum of amplitudes of the main diurnal harmonics, being O_1 the lunar and K_1 the solar constituent; and the denominator is the sum of amplitudes of main semidiurnal harmonics, being M_2 the lunar and S_2 the solar constituent.

The dominance of the tide is classified by tidal form number F: For F < 0.25, the tidal form is considered pure semidiurnal; for 0.25 < F < 1.5, the tidal regime is classified as mixed with semidiurnal dominance; if 1.5 < F < 3, the tidal form is mixed with diurnal dominance; and for F > 3, the tidal regime is pure diurnal (Zhu et al., 2015; Frota et al., 2016). However, the classifying threshold can vary among the literature. For example, Stephenson (2016) used 0.5 as the cut-off value between semidiurnal and diurnal or mixed semidiurnal. The tidal form can be used in QC to determine more appropriate tolerance values, for example the allowable difference between two consecutive samples of sea level height will be different for a semidiurnal and a diurnal tidal regime (BODC, 2007). The values used here are assuming a semidiurnal period, thus implementing this modification can result in a better QC procedure.

4.2. Tidal Prediction

The selection of which harmonics constituents should be determined in the analyses can be very challenging, even “thought of as black art” (Pugh, 1987). The more constituents are included in the analysis, the more realistic the prediction should be. However, if the time series is not long enough, then adding more constituents will not result in a more accurate tidal prediction (Munk & Cartwright, 1966). In general, the longer the time series being analysed, the more constituents can be independently determined. If analyzing 19 years of data, more than 300 independent constituents can be determined, while for a series of one year, usually sets of 30 or 60 constituents are solved (Pugh, 1987). The most basic and less restrictive rule to decide which component should be in the analysis is related to the Nyquist criterion: a constituent can only be solved if its period is less than twice the sampling interval (Pugh, 1987). Considering hourly values, the shortest constituent that can be solved will have the period of two hours. Here, the tidal prediction was done with sets of 7, 37 and 60 harmonics. The results showed that prediction error (RMSE) between the different sets were very small, varying in average 0.0002. For most cases, it is possible to see the improvement of the prediction when passing from 7 to 37 constituents. On the other hand, the RMSE for 37 and 60 is usually very similar, indicating that the last 23 constituents are not being well solved. When choosing the set of constituents to be used in the tidal predictions at the IOC-SLMF, it is recommended to compute the predictions with the set of 37 harmonic constituents.

Just as the decision of the harmonic constituents, choosing the time span to use in the harmonic analysis is a thoughtful choice. The longer the time series being analyzed, a higher numbers of harmonic constituents that can be solved independently and thus, the higher the precision of the tidal prediction (Pugh, 1987). On the other hand, several studies perform tidal analysis for each year of data. On the present study, when the analysis made with the full data set was available, 3 to 12 years depending on the station, the tidal pattern was better represented than by using only one year of data. However, by computing tidal analysis for each year of data available, it is possible to see how the harmonics change with

time. GLOSS (2011) suggests that the tidal analysis should be done for each year, and that a vector mean analysis should be applied to choose the adequate harmonic value for the tide prediction. In addition, by calculating harmonic constituents for each year, it is possible to see how the tidal form varies through the study period (Frota et al., 2016). Thus, the use of annual time series can be advantageous depending on the purpose of the tidal prediction itself.

A prediction can only be really evaluated when compared with new data, which were not used in the making of the prediction (Hyndman, 2011). Here, the tidal analysis was done with data until 31/12/2017, and evaluated with data of the first four months of 2018. The prediction had an average error of 0.13 meters, which is a relatively small error considering that the mean amplitude of the data was 1.53 meters. Most of the variation from the sea level observations (SST) could be explained by the residuals of the prediction (SSR). The variance that could not be explained by the prediction (SSE) was associated with the non-astronomical sea level signal. This interpretation should be further corroborated by analyzing the sub-tidal frequencies of the sea level signal, which is mainly influenced by meteorological events, local morphological characteristics and river discharge (Miranda et al., 2002; Frota et al., 2016). Such frequencies can be separated from the tidal frequency by applying a low-pass filter to the sea level data set (Pugh, 1987), and should also be analyzed with local wind and pressure data (Frota et al., 2016). Still, for almost all the stations the total variation of the sea level height could not be explained by only summing the tidal variance and the meteorological residue. This third source of variance was assigned to noises induced by the instrumentation and/or methods used for data collecting and processing.

The main purpose of the present work was to obtain a tidal prediction for each station, which can be then be used as a reference of what is being visualized in real time. One of the challenges of RT evaluation of the data is to differentiate real errors from signals variations caused by tsunamis or storms. The graph in Figure 16 exemplifies a situation of a small tsunami in George Town (Cayman Islands), generated by an earthquake in north of Honduras. The figure shows in red the raw data, with no QC, as it is seen in real time. While the upper panel illustrates the hourly filtered values, the lower panel shows the high frequency data, in minute interval. The black line represents the tidal prediction.

When the event reaches the station, the distance between the observed and predicted values increases significantly. However, it is possible to see that the tidal pattern is kept in the observations, only with higher amplitude than the expected. This is not expected to happen in the case of a real error. It is interesting to notice that the waves generated by the earthquake do not cause positive fluctuations; in contrary, the elevation curve is dislocated downwards. However, the amplitude of the curve is accentuated by the event. This demonstration validates the use of the tidal prediction computed here as a rough QC for the tide gauge observations.

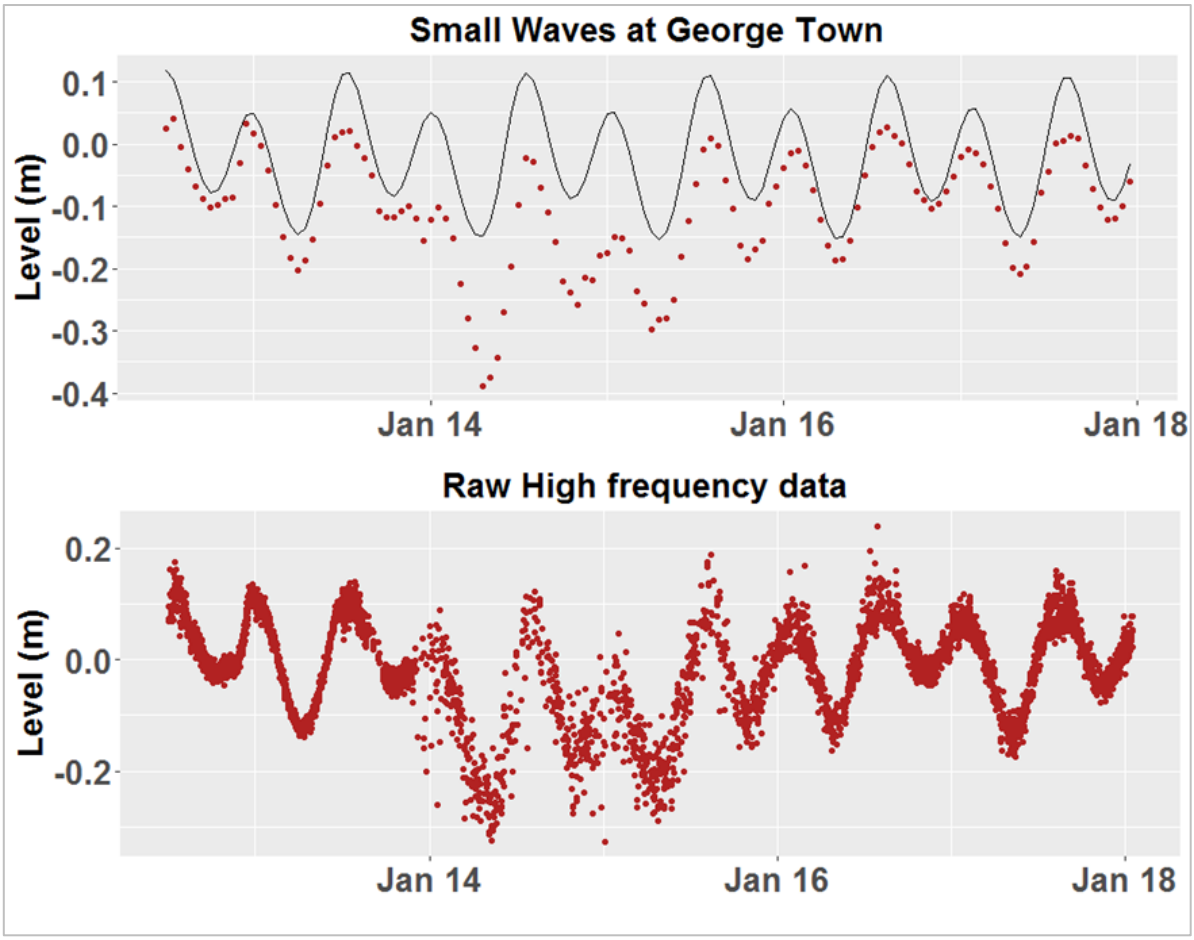


Figure 16. Example of an event at station *Geor*, in January of 2018. In red: raw data. Black: tidal prediction. Lower panel shows the high frequency data, in minute interval; and upper panel the hourly filtered data.

5. Conclusion

The QC applied at the present work was successful in returning a clean series, getting rid of random and systematic errors. Despite the limited contribution of the Spike module, algorithm 1, which applied all the modules seems to have the best outcome. The *global* outlier go filter is redundant, and can, therefore, be removed from the QC. On the other hand, if a fast and robust check is needed, the *go* test can be applied. Especially the Breakpoint module was very effective in dealing with sharp discontinuities and datum shifts in the time series. Through all the QC steps, visual interpretation of the results showed to be very important to determine if actual wrong values were being detected by the tests.

The tidal prediction made with the analysis of the tide gauge data proved to be reliable. It is recommended that the tidal analysis at the IOC-SLMF to be done with 37 harmonic constituents. The stations with the stronger deviations between the predicted and observed values were the ones in stormy areas and the ones with bad quality of data. Nevertheless, the predictions can be applied to real-time data visualization in order to obtain a reference level of what would be expected to see.

Future Work

Future work should focus on improving the QC described here, especially the Spike Identification Module. In addition, should look for ways of implementing the different characteristics of the stations in the QC, apart from trying to add more checks. However, it is important to not compromise the speed of the QC process. Some checks made during the QC removed the signal fluctuations caused by tsunamis and storms. Future work should try to add a special flag for these occasions. Furthermore, a Doodson filter should be implemented to pass from the high frequency data to hourly values. This could improve significantly the tidal analysis and predictions. In order to apply the QC to the 800 stations available at the IOC-SLMF, it is still necessary to make some modifications on the QC procedure, making it faster and more applicable to a large database.

Outlook

Despite its technical character, the present work is very important to improve the services of the IOC-SLMF. Apart from the use of the results by VLIZ, this work can become base for other works which can connect the sea level observations and tidal predictions available at the IOC-SLMF with biological databases, such as Life Watch. Another application is to use the database to characterize natural disasters events. From analyzing the residual between the tidal prediction and the sea level observations, is possible to look for characteristics of the records of tsunamis and storm surges, and try to gather mean information of range of frequency, period and amplitude of these events. Furthermore, based on this 10 years database of high frequency sea level measurements, one can try to answer the question “Is the frequency of tsunamis and storm surges increasing?”

6. Acknowledgments

This work was achieved by the supervision of Prof. Dr. Karline Soetaert, and by the help and oversight of Francisco Hernandez during the entire data treatment, and to both I would like to express my gratitude for the guidance, support and patience. Data access was provided by VLIZ and the IOC Sea Level Monitoring Facility. I would like to thank Global Sea Mineral Resources (GSR) and the International Seabed Authority (ISA) for the fellowship which allowed to join this Master programme.

I also want to thank my dearest Oceans & Lakes family for making these two years a very happy and memorable time. Being away from home is not always easy, but your friendship made Brussels feel like home. Thank you for all the help, advices and patience during these two years. A special thanks to Leonie, Pedro and Tati for bringing a happy atmosphere to the thesis work sessions. I would also like to thank the Stack Overflow community for all the help with R. Finally, I want to thank my family for always supporting my decisions, even when they took me to the other side of the Atlantic.

7. References

British Oceanographic Data Centre – BODC (2007). Data Quality Control Procedures, *Version*
British Oceanographic Data Centre – BODC (2007). Data Quality Control Procedures, *Version 3.0.*(September 2007), 75pp.

Caldwell, P. C., M. A. Merrifield, P. R. Thompson (2015). Sea level measured by tide gauges from global oceans — the Joint Archive for Sea Level holdings (NCEI Accession 0019568), Version 5.5, *NOAA National Centers for Environmental Information*, Dataset, doi:10.7289/V5V40S7W.

Crawley, M.J. (2011). Statistics: An Introduction using R. John Wiley & Sons Ltda. DOI:10.1002/9781119941750. Available online at http://www3.imperial.ac.uk/pls/portallive/docs/1/1171920.PDF.

Consoli, S., Recupero, D.R., Zavarella, V. (2013). A survey on tidal analysis and forecasting methods for tsunami detection. Science of Tsunami Hazards. 58 pp.

Global Sea-Level Observing System – GLOSS (2011). Manual on Quality Control of Sea Level Obsevations, Version 1.0, 38pp. *Draft*.

Emburi, S. M. (2004). The Importance of Data Quality Control. NERC Microarray Data Quality Control and Analysis Workshop.

Foreman, M.G.G. (1977). Manual for tidal heights analysis and prediction. Pacific Marine Science Report 77-10

Frota, F.F., Truccolo, E.C., Schettini, C.A.F. (2016). Tidal and sub-tidal sea level variability at the northern shelf of the Brazilian Northeast Region. Annals of the Brazilian Academy of Sciences. 88 (3): 1317-1386. http://dx.doi.org/10.1590/0001-3765201620150162

Han, S-C., Jekeli, C., Shum, C.K. (2004). Time-variable aliasing effects on ocean tides, atmosphere, and continental water mass on monthly mean GRACE gravity field. Journal of Geophysical Research, 109, B04403, doi:10.1029/2003JB002501.

Holgate, S.J., Woodworth, P., Foden, P.R., Pugh, J. (2008). A study of delays in making tide gauge data available to tsunami warning centers. Journal of Atmospheric and oceanic Technology, 25, 475-481. DOI: 10.1175/2007JTECHO544.1

Holgate, S.J., Matthews, A., Woodworth, P.L., Rickards, L.J., Tamisiea, M.E., Bradshaw, E., Foden, P.R., Gordon, K.M., Jevrejeva, S., and Pugh, J. (2013). New data systems and products at the Permanent Service for Mean Sea Level. Journal of Coastal Research, 29(3), 493–504. Coconut Creek (Florida), ISSN 0749-0208.

Hyndman R.J. (2011) Forecasting: An Overview. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-04898-2

Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: Principles and Practice. 2nd Edition. O Texts, Monash University, Australia. Book available online at https://otexts.org/fpp2/

IOC. (1985). Manual on Sea level measurement and interpretation. Volume I – Basic Procedures. Intergovernmental Oceanographic Commission Manuals and Guides No. 14. IOC, Paris, 83pp.

IOC. (2006). Manual on Sea level measurement and interpretation. Volume IV – An update to 2006. Paris, Intergovernmental Oceanographic Commission of UNESCO. 78pp. (IOC Manuals and Guides No. 14, vol IV; JCOMM Technical Report No.31; WMO/TD No. 1339).

Killick, R., Eckley, I. A. (2014). changepoint: An **R** Package for Changepoint Analysis. *Journal of Statistical Software*, Vol 58, issue 3.

Merrifield, M.A., Firing, Y.L., Aarup, T., Agricole, W., Brundrit, G., Chang-Seng, D., Farre, R., Kilonsky, B., Knight, W., Kong, L., Magori, C., Manurung, P., McCreery, C., Mitchell, W., Pillay, S., Schindele, F., Shillington, F., Testut, L., Wijeratne, E.M.S., Caldwell, P., Jardin, J., Nakahara, S., Porter, F.Y., Turesky, N. (2005). Tide gauge observations of the Indian Ocean tsunami, December 26, 2004. Geophysical Research Letters, 32, L09603, doi:10.1029/2005GL022610.

Merrifield, M.A., Aarup, T., Allen, A., Aman, A., Bradshaw, E., Caldwell, P., Fernandes, R.M.S., Hayashibara, H., Hernandez, F., Kilonsky, B., Martin Migues, B., Mitchum, G., Péres Gómez, B., Rickards, L., Rosen, D., Schöne, T., Szabados, M., Testut, L., Woodworth, P., Woppelmann, G., Zavala, J. (2009). The Global Sea Level Observing System (GLOSS). Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Vol. 2. (Venice, Italy).

Miranda, L.B., Castro, B.M., Kjerfve, B. (2002). Princípios de Oceanografia Física de Estuários. São Paulo, Editora da Universidade de São Paulo - EDUSP. 424 p.

Munk, W., Cartwright, D.E. (1966). Tidal spectroscopy and prediction. Phil Trans R Soc London A 259: 533-581.

National Oceanic and Atmospheric Administration, NOAA. (2018). Tide Prediction Error for the United States Stations. Available online at https://tidesandcurrents.noaa.gov/pdf/Tide_Prediction_Error_for_the_United_States_Coastline.pdf

Pugh, D. T. (1987). Tides, Surges and mean sea-level. Book. *John Wiley & Sons*, 472.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rickards, L.., Kilonsky, B. (1997). Developments in sea level data management and exchange. *Ocean Data Symposium*, *Dublin*, *Irland* (October).

Schettini, C.A.F., Maia, L.P., Truccolo, E.C. (2011). Análise da variabilidade do nível da água na costa de Fortaleza, Cerá.Arquivos de Ciëncias do Mar, Fortaleza, 44(1):27-32

Stephenson, A. G. (2016). Harmonic Analysis of Tide Using TideHarmonics. URL: https://CRAN.R-project.org/package=TideHarmonics.

Van Onselen, K.I. (2000). The influence of data quality on the detectability of sea-level height variations. Delft: Netherlands Geodetic Commission. Publication on Geodesy 49. 204 pp.

Vinogradov, S.V., Ponte, R.M. (2011). Low-frequency variability in coastal sea level from tide gauges and altimetry. Journal of Geophysical Research, Vol 116., C07006, doi:10.1029/2011JC007034.

Unesco (1993). Manual of Quality Control Procedures for Validation of Oceanographic Data. Prepared by: CEC: DG-XII, MAST and IOC: IODE, Manual and Guide 26, SC-93/WS-19.

U.S. Integrated Ocean Observing System – IOOS (2016). Manual for the Use of Real-Time Quality Control of Water Level Data Version 2.0: A Guide to Quality Control and Quality Assurance of Water Level Observations. 46pp.

Wang, R. (2014). Low-pass Filtering and Smoothing. Available online at http://fourier.eng.hmc.edu/e161/lectures/smooth_sharpen/smooth_sharpen.html .

Weisse, R., Storch, H., Niemeyer, H.D., Knaack, H. (2011). Changing North Sea storm surge climate: An increasing hazard? Ocean & Coastal Management, 68, pp 58-68, doi:10.1016/j.ocecoaman.2011.09.005

Woodworth, P.L., Hunter, J.R., Marcos, M., Caldwell, P., Menéndez, M., Haigh, I. (2017). Towards a global higher-frequency sea level dataset. Geoscience

Zhu, Jianrong & Wu, Hui & Li, Lu. (2015). Hydrodynamics of the Changjiang Estuary and Adjacent Seas. 19-45. 10.1007/978-3-319-16339-0_2.

7.1. R Packages

Constantine, W., Percival, D. (2017). fractal: A Fractal Time Series Modeling and Analysis Package. URL: https://cran.r-project.org/web/packages/fractal/fractal.pdf .

Kelley, D., Richards, C., Layton, C. (2018). Oce: Analysis of Oceanographic Data. URL: https://cran.r-project.org/web/packages/oce/index.html

Killick, R., Haynes, K., Eckley, I., Fearnhead, P., Lee, J. (2016). changepoint: Methods for Changepoint Detection. URL: https://github.com/rkillick/changepoint/ .

Lazaridis, E. (2015). lunar: Lunar Phase & Distance, Seasons and Other Environmental Factors. URL: http://statistics.lazaridis.eu .

Stephenson, A. G. (2016). Harmonic Analysis of Tide Using TideHarmonics. URL: https://CRAN.R-project.org/package=TideHarmonics.

Trapletti, A., Hornik, K., LeBaron, B. (2018). tseries: Time Series Analysis and computational finance. URL: https://cran.r-project.org/web/packages/tseries/tseries.pdf .

Wickham, H., Chang, W. (2016). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. URL: http://ggplot2.tidyverse.org, https://github.com/tidyverse/ggplot2 .

Wickham, H., Francois, R., Henry, L., Muller, K. (2017). dplyr: A Grammar of Data Manipulation. URL:http://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr .

Wuertz, D., Setz, T., Chalabi, Y., Maechler, M., Byers, J.W. (2018) timeDate: Rmetrics - Chronological and Calendar Objects. URL: https://cran.r-project.org/web/packages/timeDate/index.html, https://www.rmetrics.org.

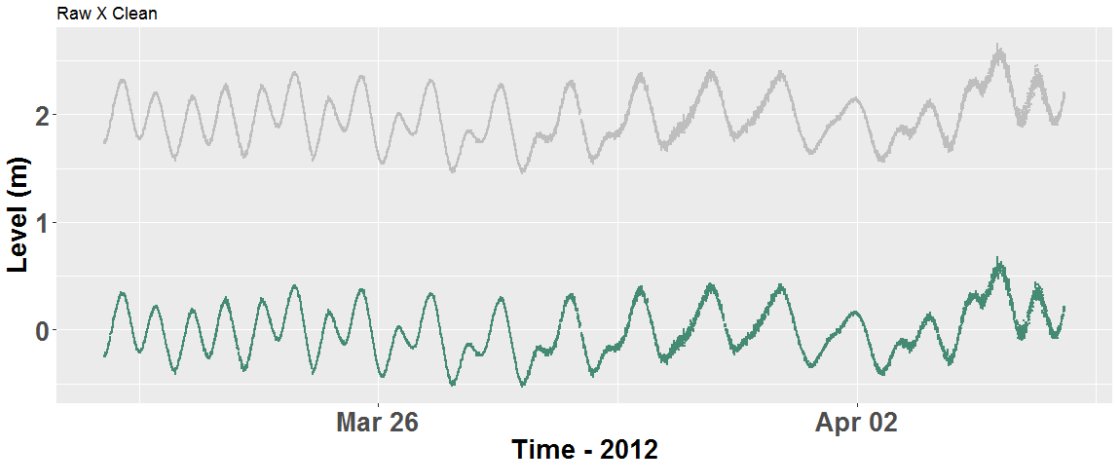
Zeileis, A., Gronthendieck, G., Ryan, J.A., Ulrich, J.M., Andrews, F. (2017). zoo: S3 Infrasctructure for Regular and Irregular Time Series (Z's Ordered Observations). URL: http://zoo.R-Forge.R-project.org/ .

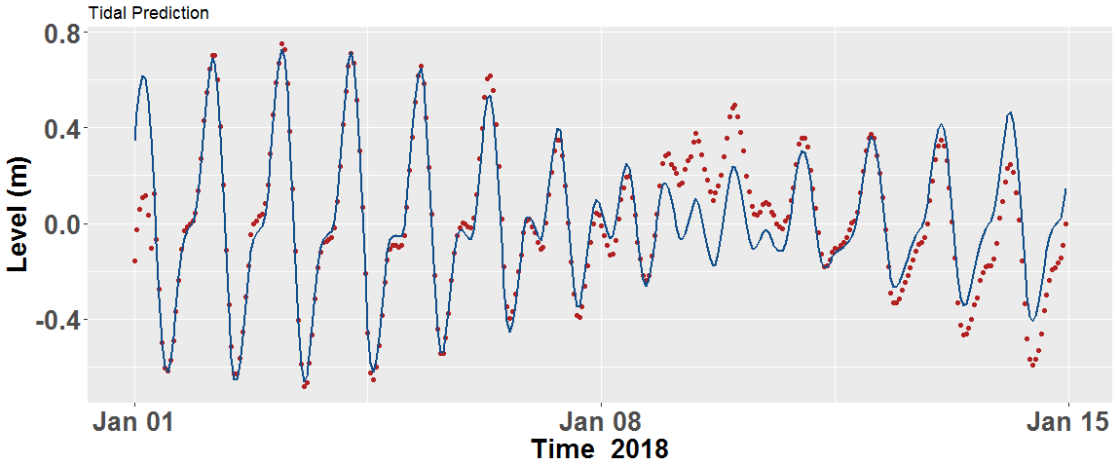
8. Annex

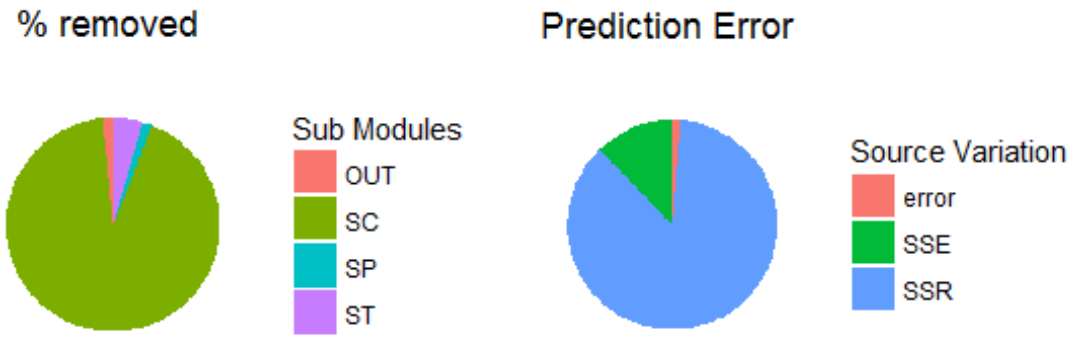
8.1. Station Summary

Abas

```
## [1] "Station: abas  Sensor: rad  Country: Japan"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:19:00.000"

## [1] "MSl: 0  Amp: 1.59  Points Removed in QC: 1.17 %"

## [1] "RMSE tide: 0.09  Set harmonics: h69"
```

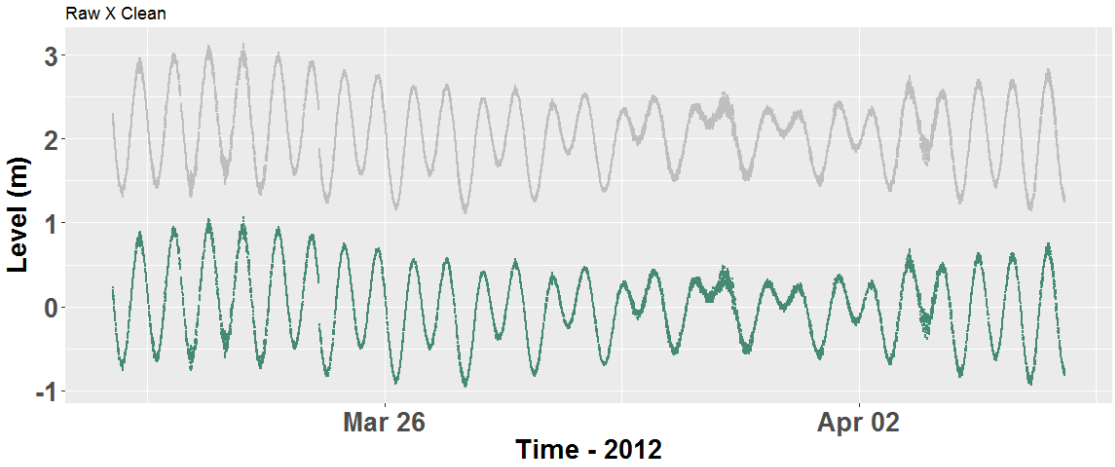


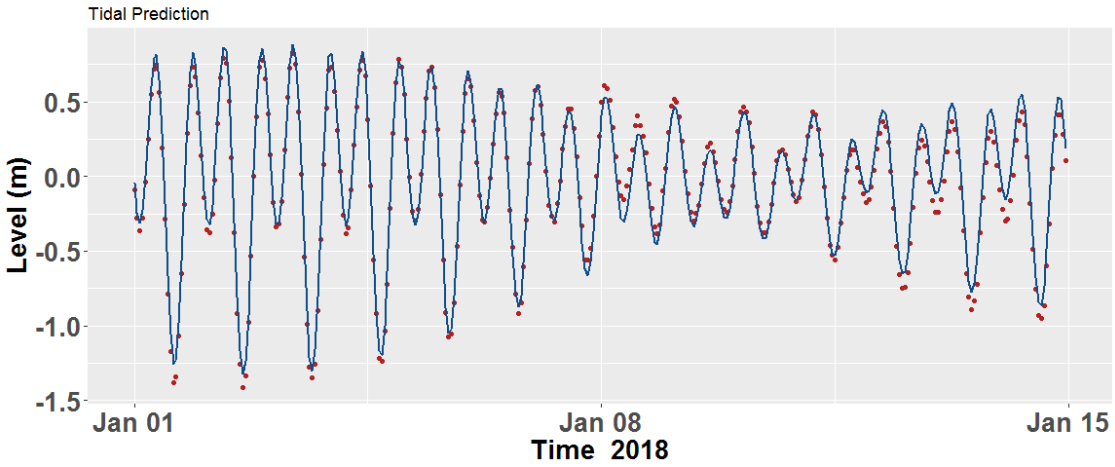


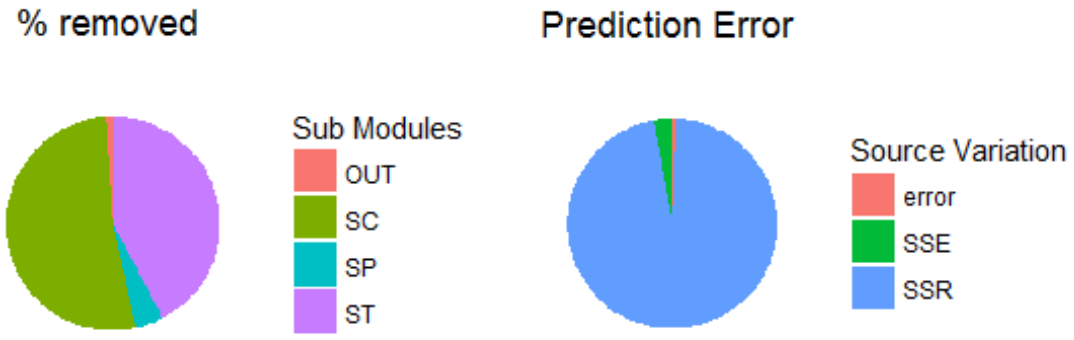


Abur

```
## [1] "Station: abur  Sensor: rad  Country: Japan"

## [1] "First obs: 2018-04-03 05:17:00  Last obs: 2018-04-05 22:19:00.000"

## [1] "MSl: 0  Amp: 2.67  Points Removed in QC: 0.09 %"

## [1] "RMSE tide: 0.08  Set harmonics: h37"
## [2] "RMSE tide: 0.08  Set harmonics: h37"
```

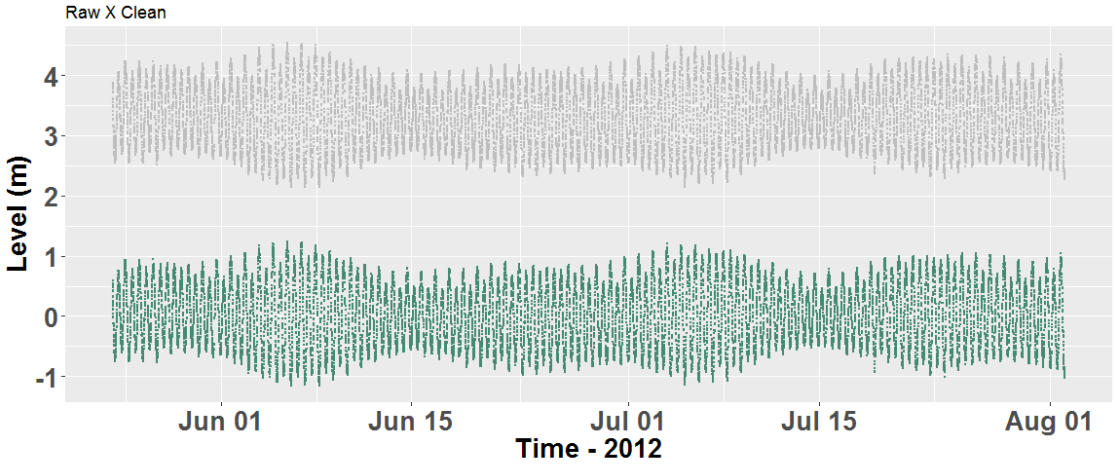


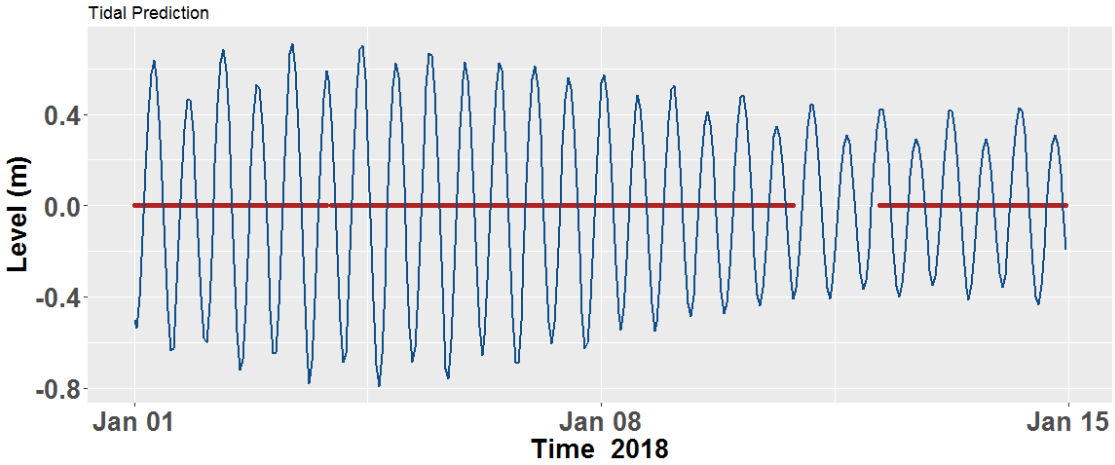


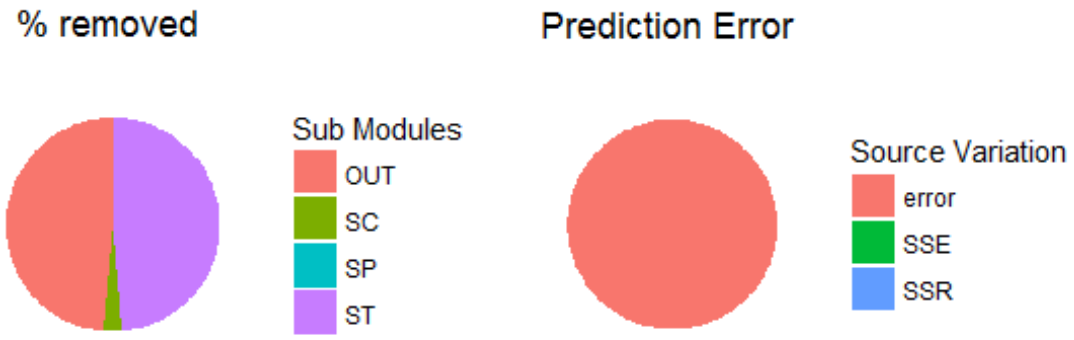


Acaj

```
## [1] "Station: acaj  Sensor: bub  Country: El Salvador"
## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:21:00.000"
## [1] "MSl: 0  Amp: 2.63  Points Removed in QC: 18.75 %"
## [1] "RMSE tide: 0.32  Set harmonics: h69"
```

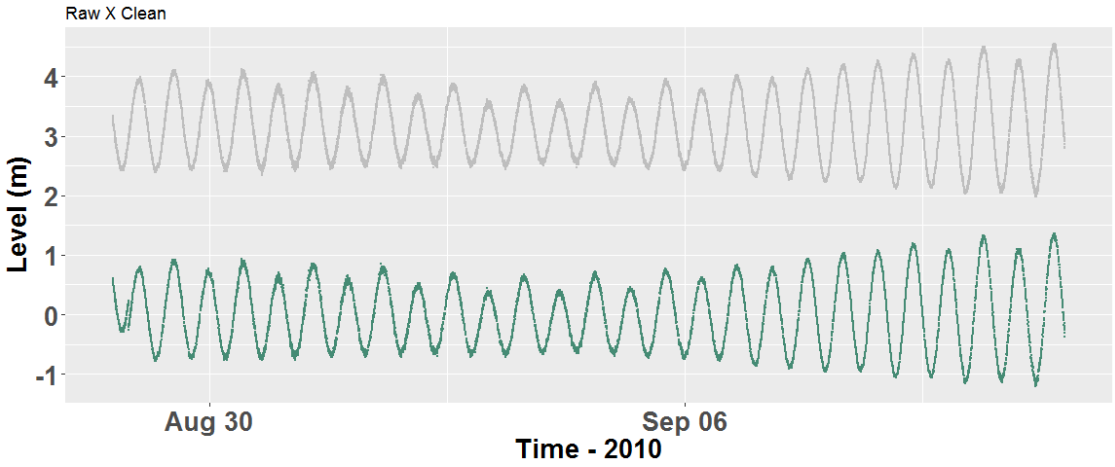


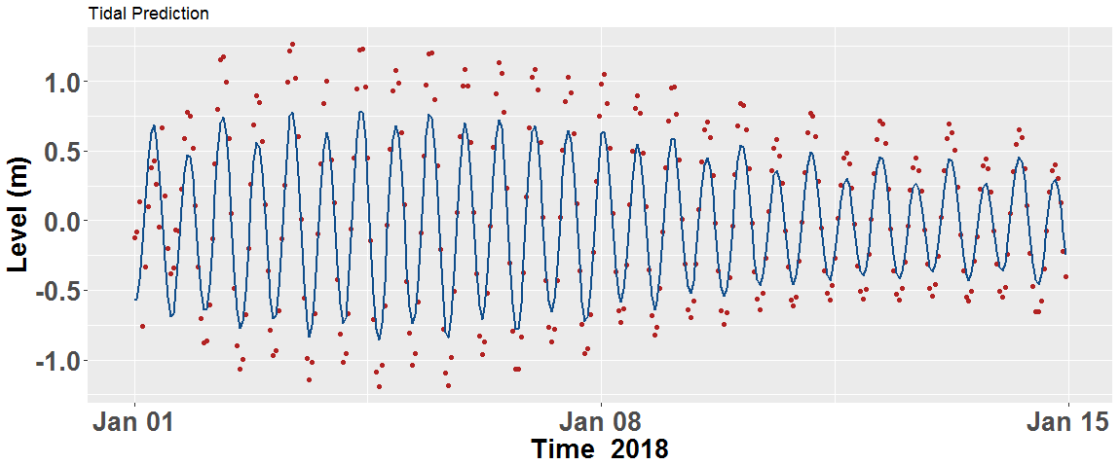


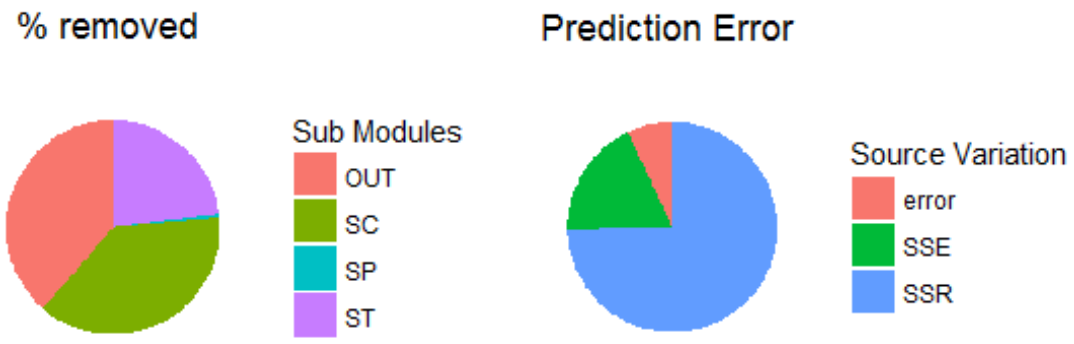


Acaj

```
## [1] "Station: acaj  Sensor: prs  Country: El Salvador"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:21:00.000"

## [1] "MSl: 0  Amp: 3.17  Points Removed in QC: 2.16 %"

## [1] "RMSE tide: 0.2  Set harmonics: h37"
```

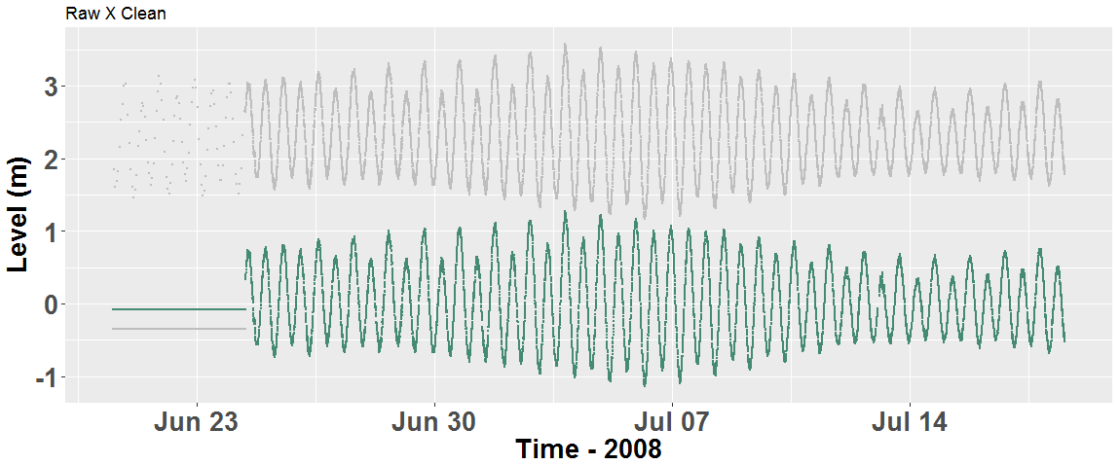


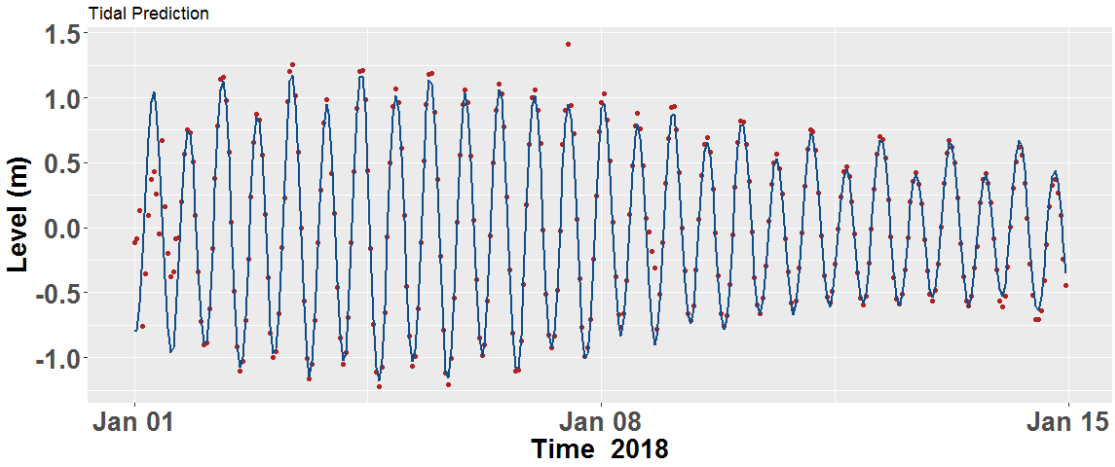


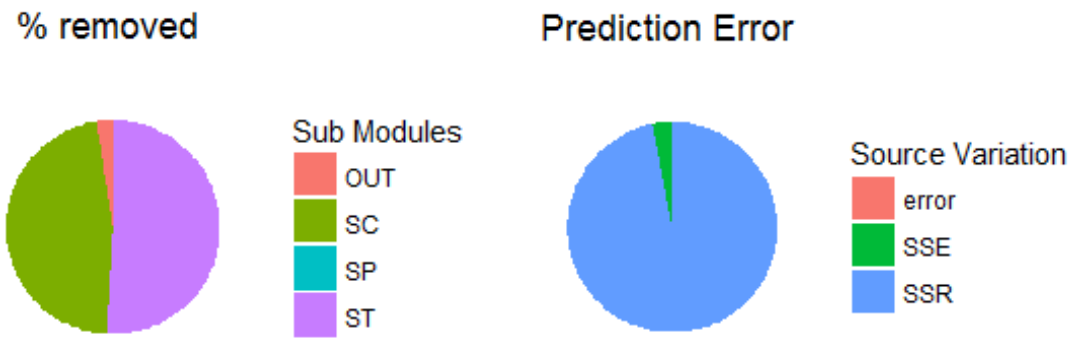


Acaj

```
## [1] "Station: acaj  Sensor: rad  Country: El Salvador"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:21:00.000"

## [1] "MSl: 0  Amp: 3.31  Points Removed in QC: 1.71 %"

## [1] "RMSE tide: 0.1  Set harmonics: h37"
```

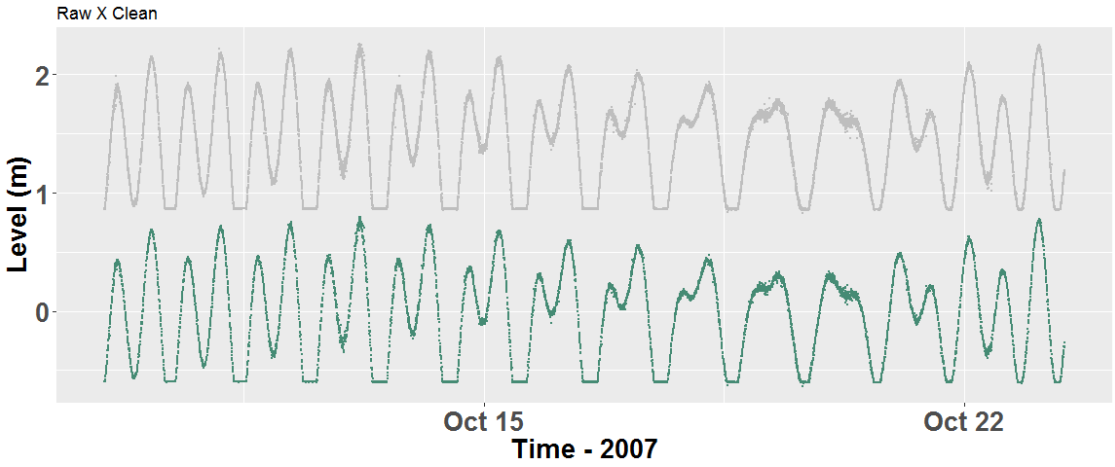


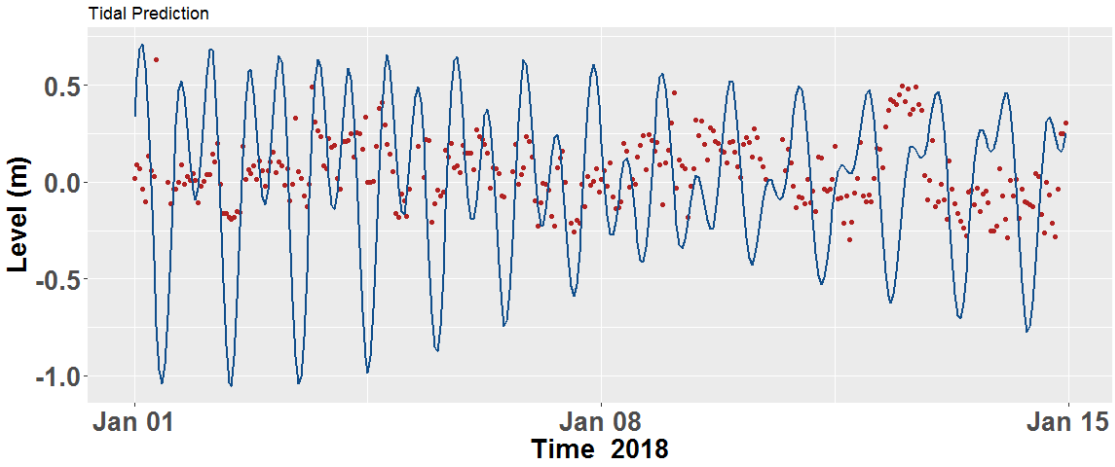


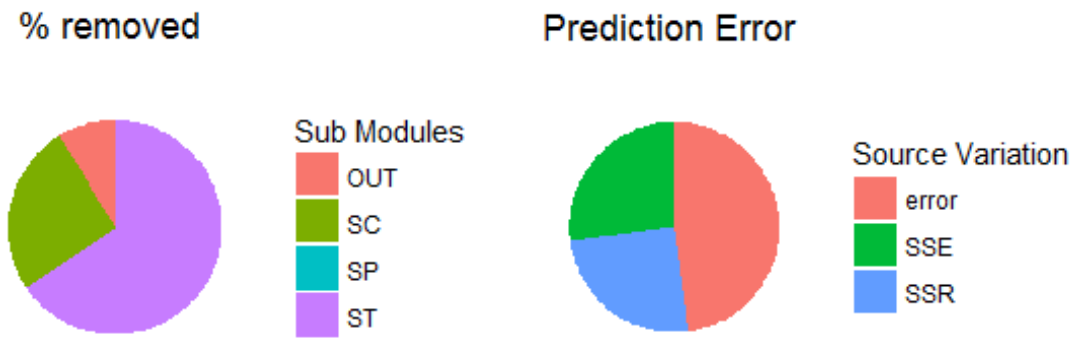


Aden

```
## [1] "Station: aden  Sensor: pr1  Country: Yemen"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 21:57:00.000"

## [1] "MSl: 0  Amp: 2.87  Points Removed in QC: 8.4 %"

## [1] "RMSE tide: 0.43  Set harmonics: h7"
```

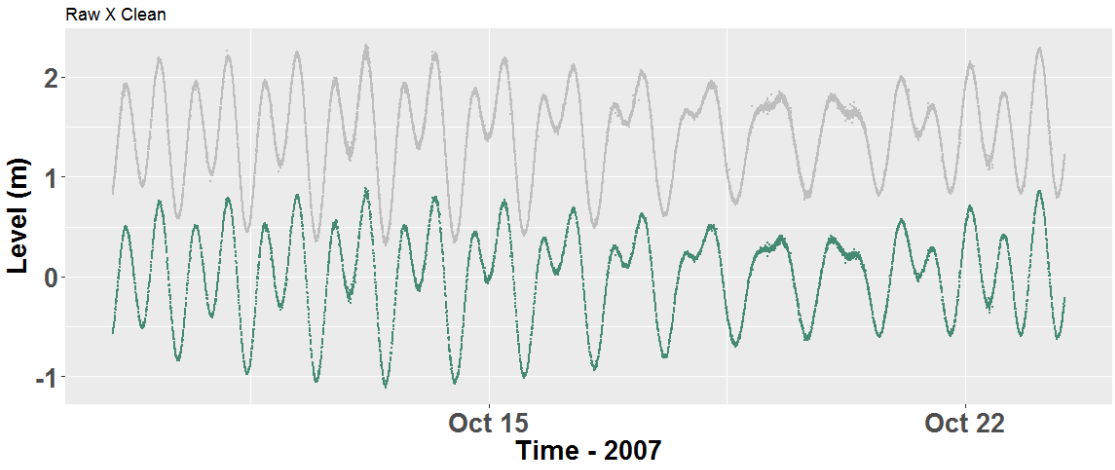


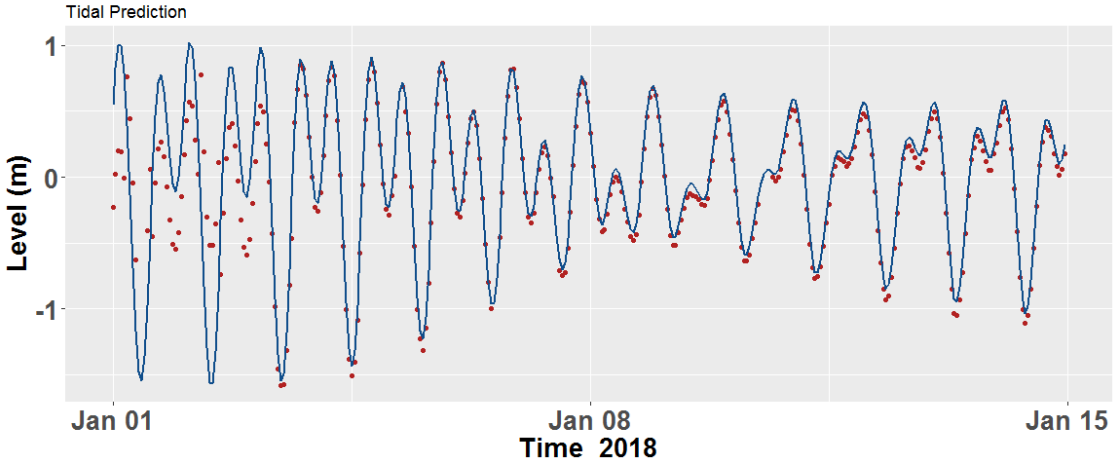


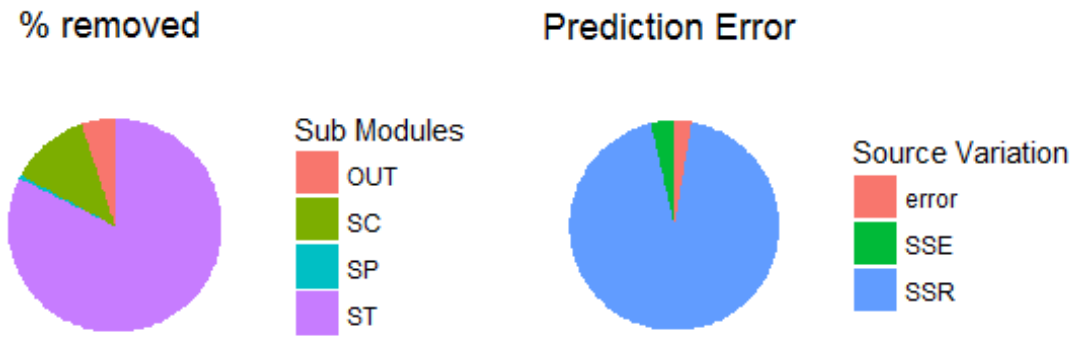


Aden

```
## [1] "Station: aden  Sensor: pr2  Country: Yemen"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 21:57:00.000"

## [1] "MSl: 0  Amp: 3.15  Points Removed in QC: 3.74 %"

## [1] "RMSE tide: 0.11  Set harmonics: h37"
## [2] "RMSE tide: 0.11  Set harmonics: h37"
```

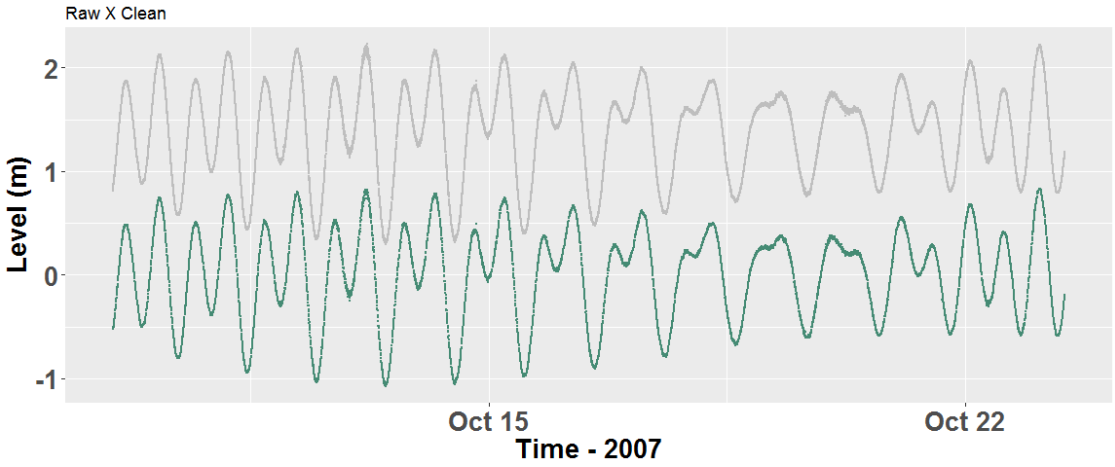


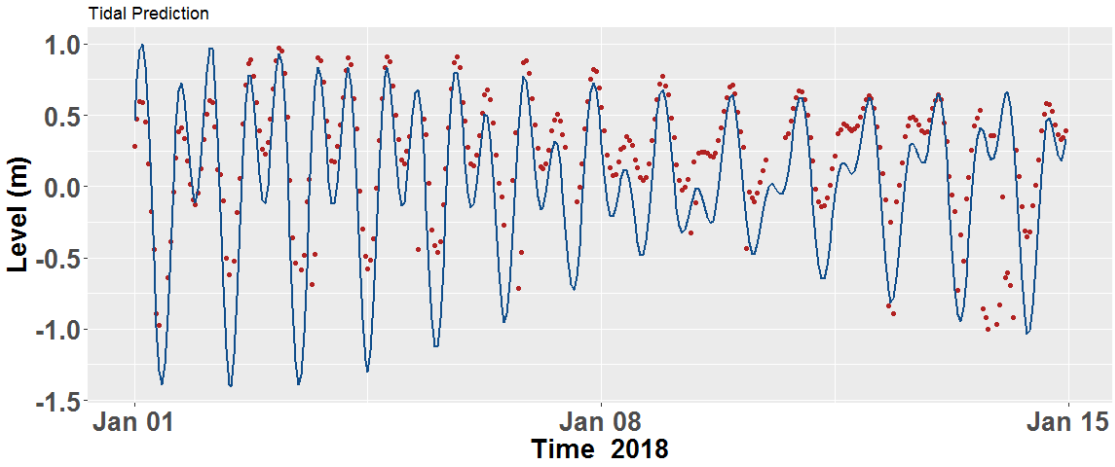


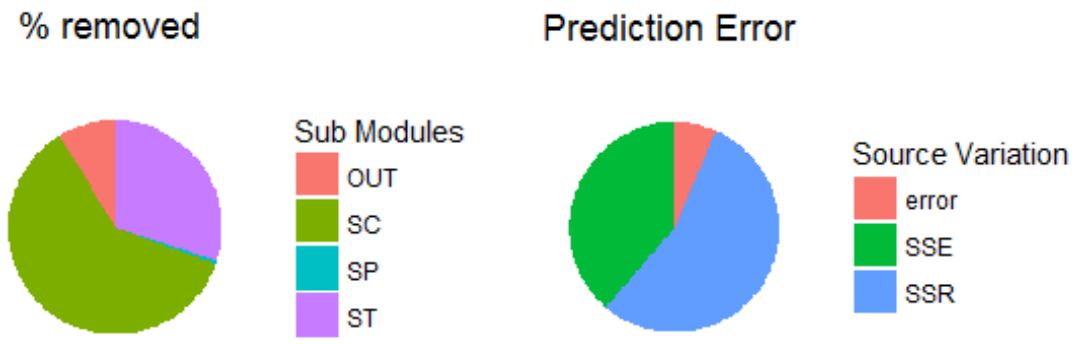


Aden

```
## [1] "Station: aden  Sensor: rad  Country: Yemen"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 21:57:00.000"

## [1] "MSl: 0.01  Amp: 3.65  Points Removed in QC: 17.03 %"

## [1] "RMSE tide: 0.47  Set harmonics: h7"
```

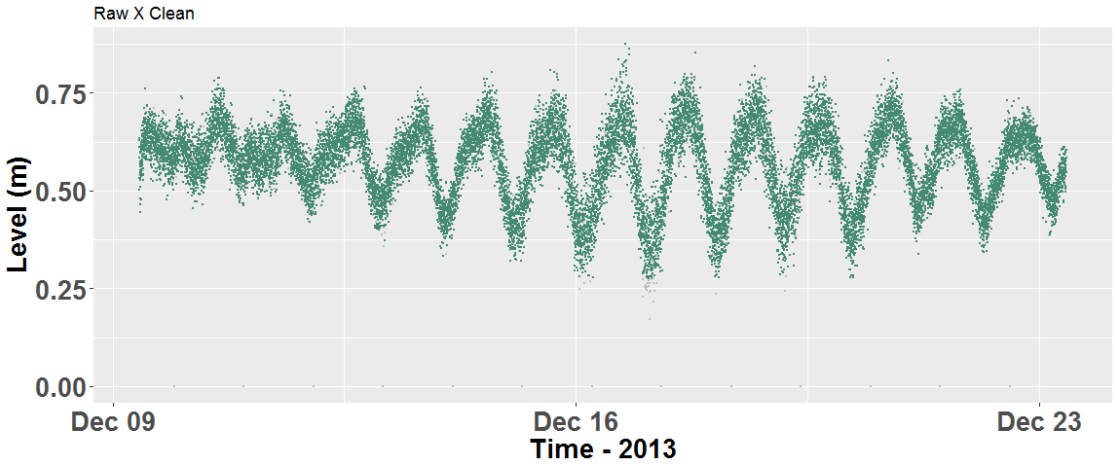


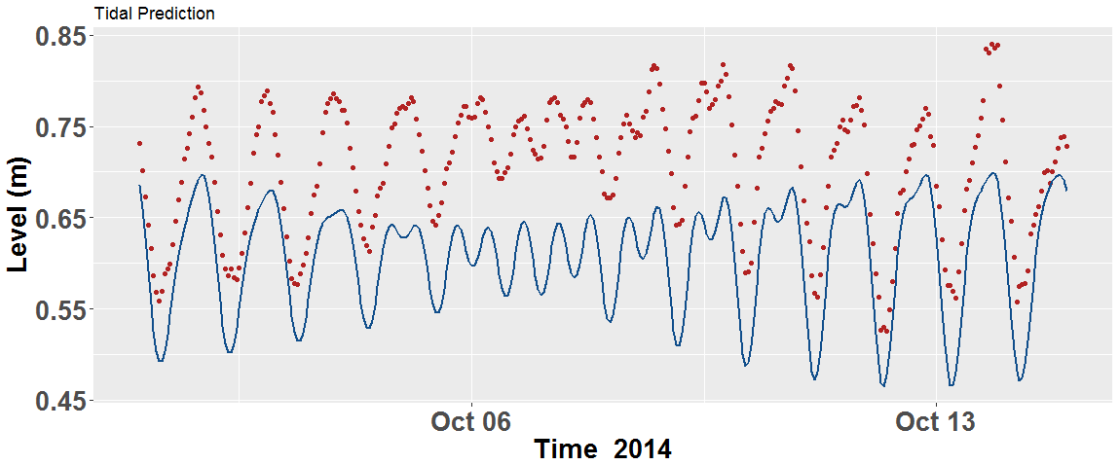


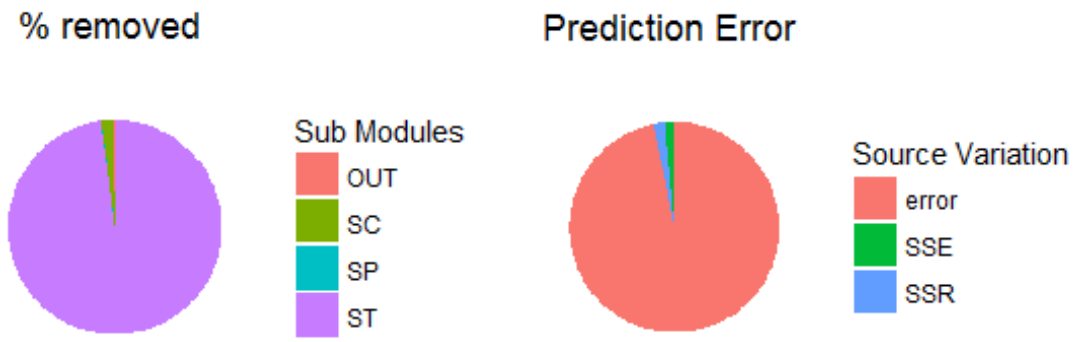


Bass

```
## [1] "Station: bass  Sensor: prs  Country: St. Kitts & Nevis"

## [1] "First obs: 2018-02-08 06:16:00  Last obs: 2018-02-10 13:08:00.000"

## [1] "MSl: 3.51  Amp: 42.62  Points Removed in QC: 73.63 %"

## [1] "RMSE tide:   Set harmonics: h7"
```

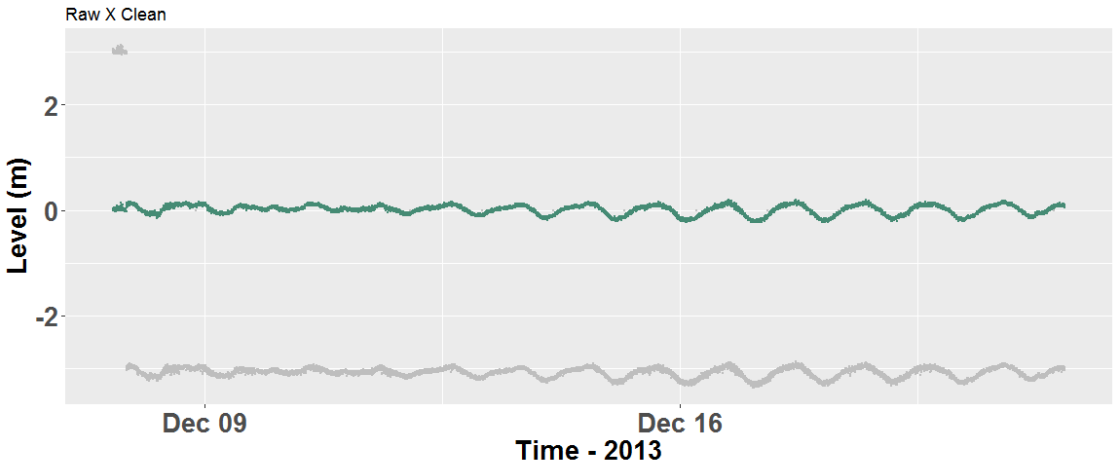


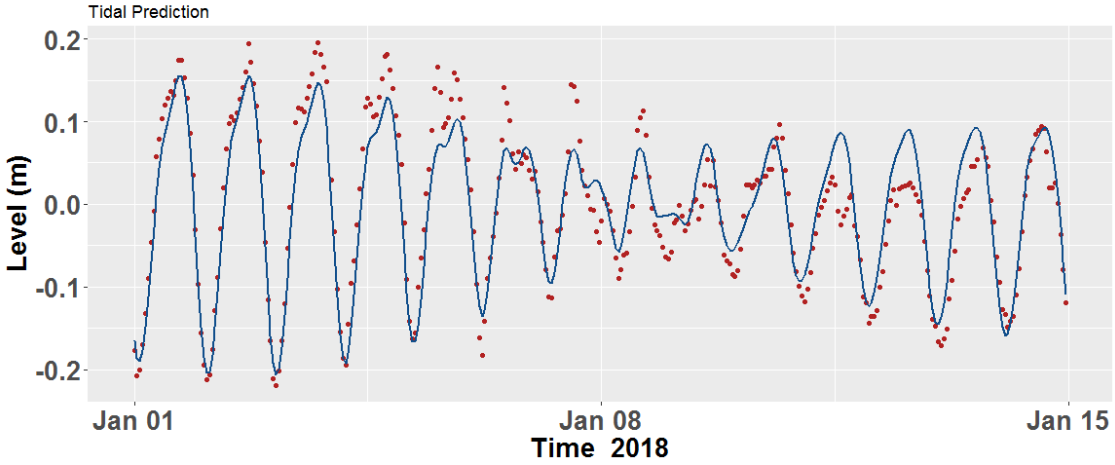


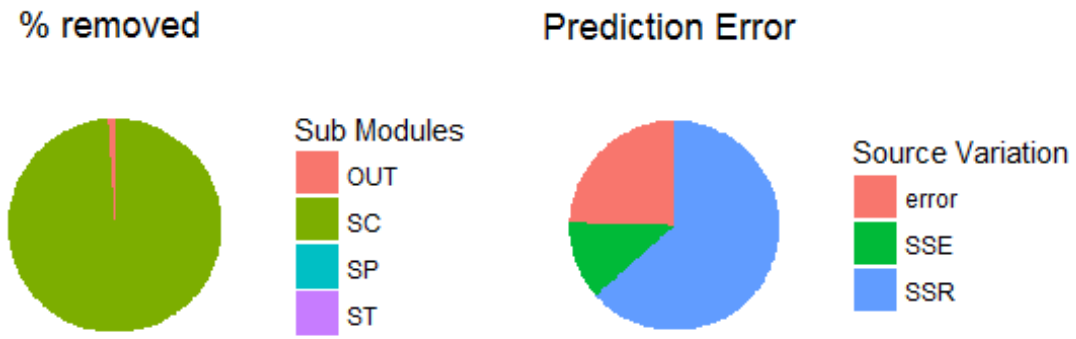


Bass

```
## [1] "Station: bass  Sensor: rad  Country: St. Kitts & Nevis"

## [1] "First obs: 2018-02-08 06:16:00  Last obs: 2018-02-10 13:08:00.000"

## [1] "MSl: 0  Amp: 0.52  Points Removed in QC: 20.67 %"

## [1] "RMSE tide: 0.03  Set harmonics: h37"
## [2] "RMSE tide: 0.03  Set harmonics: h37"
```

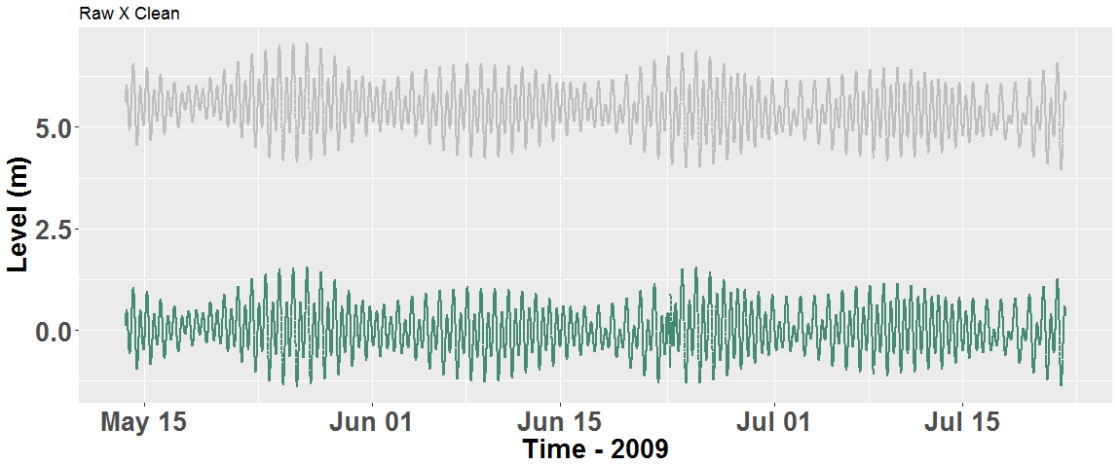


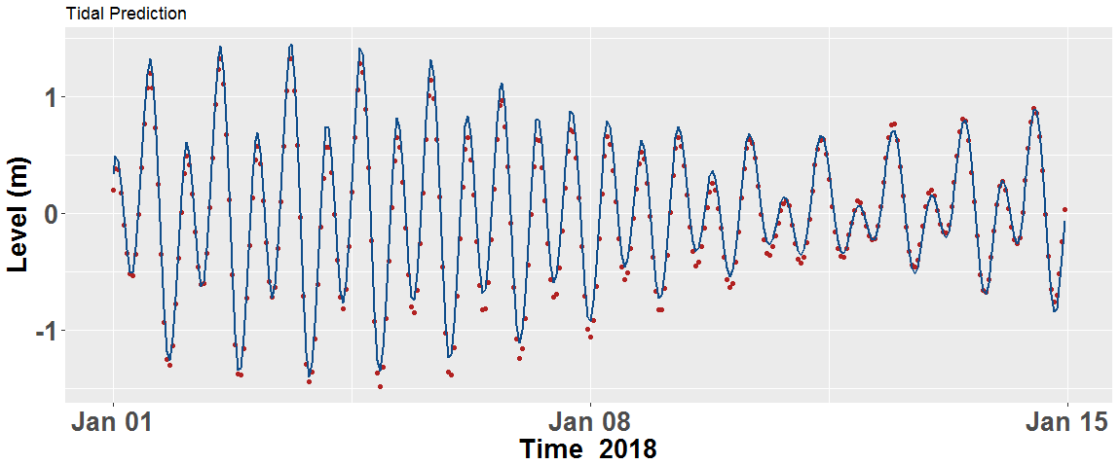


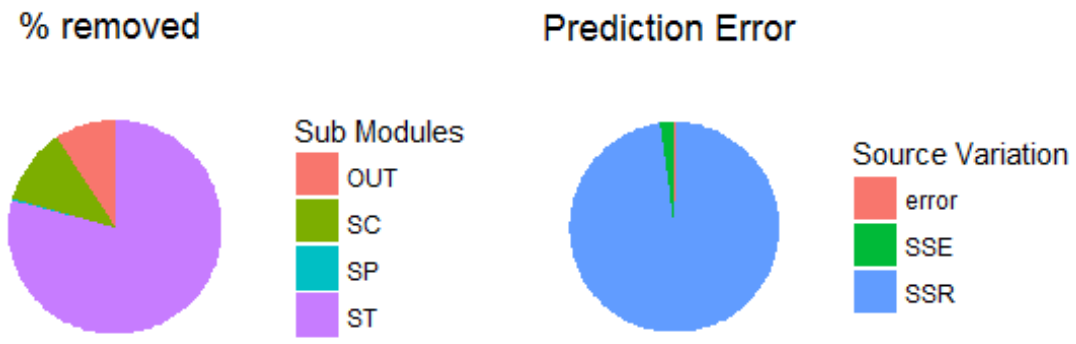


Beno

```
## [1] "Station: beno  Sensor: enc  Country: Indonesia"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:22:00.000"

## [1] "MSl: 0  Amp: 3.31  Points Removed in QC: 0.62 %"

## [1] "RMSE tide: 0.09  Set harmonics: h69"
```

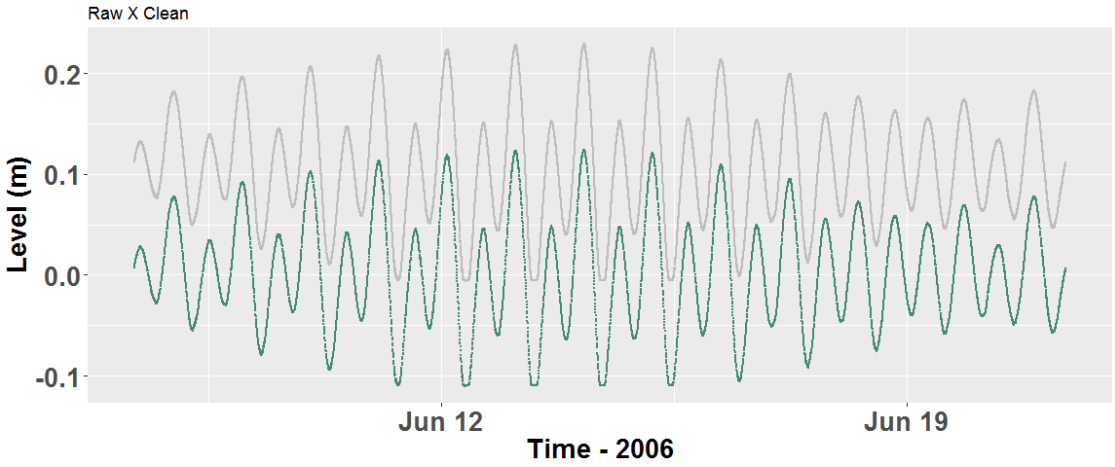


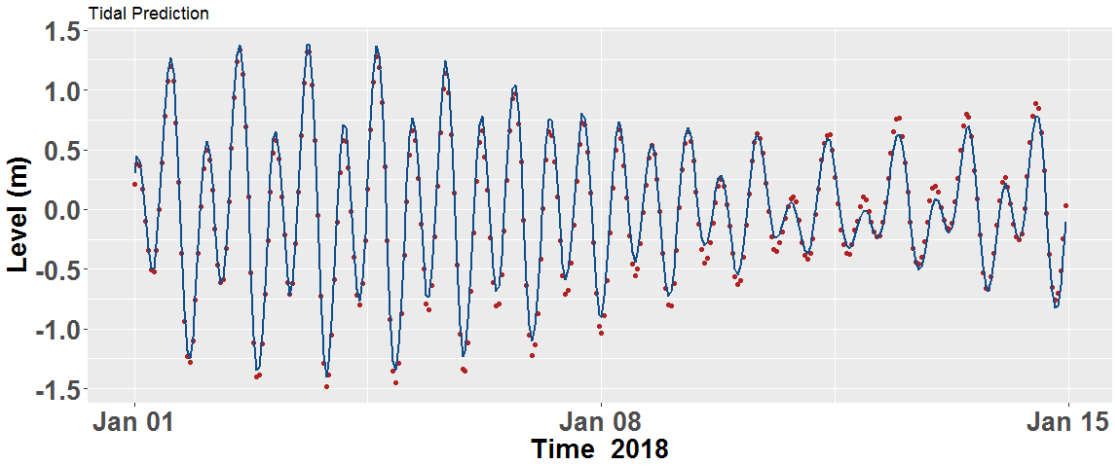


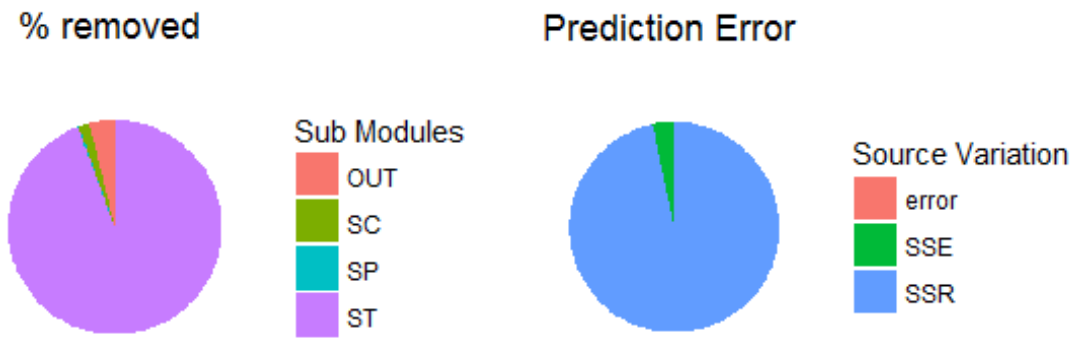


Beno

```
## [1] "Station: beno  Sensor: prs  Country: Indonesia"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:22:00.000"

## [1] "MSl: 0  Amp: 3.92  Points Removed in QC: 15.83 %"

## [1] "RMSE tide: 0.1  Set harmonics: h69"
```

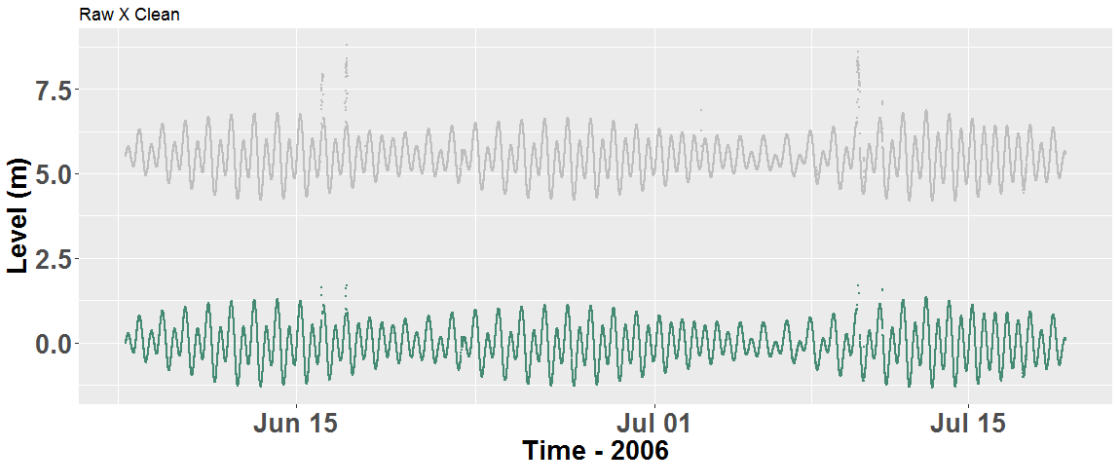


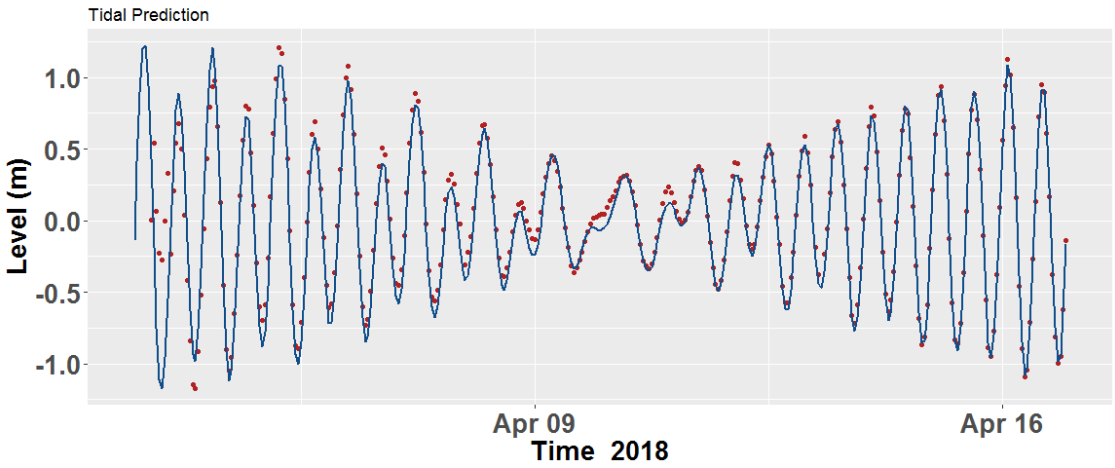


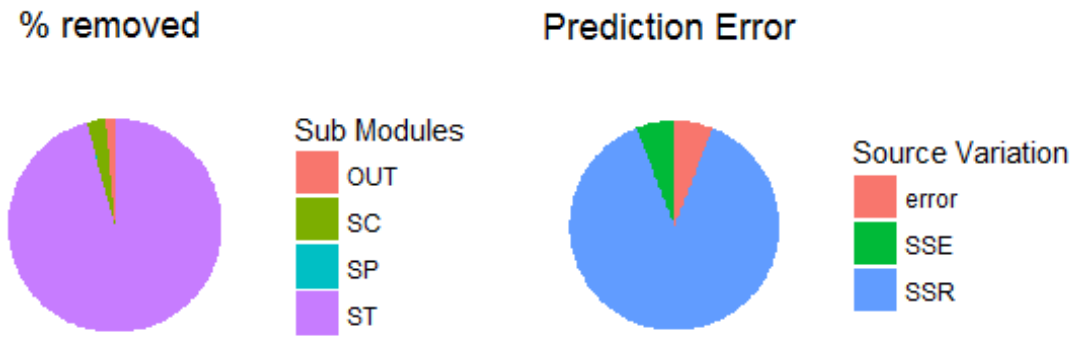


Beno

```
## [1] "Station: beno  Sensor: rad  Country: Indonesia"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:22:00.000"

## [1] "MSl: 0  Amp: 3.36  Points Removed in QC: 2.82 %"

## [1] "RMSE tide: 0.15  Set harmonics: h37"
## [2] "RMSE tide: 0.15  Set harmonics: h37"
```

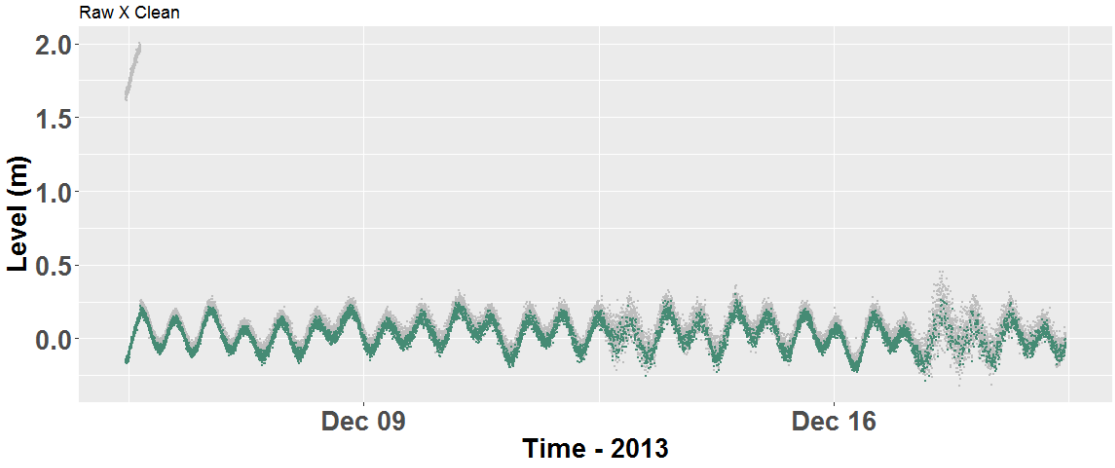


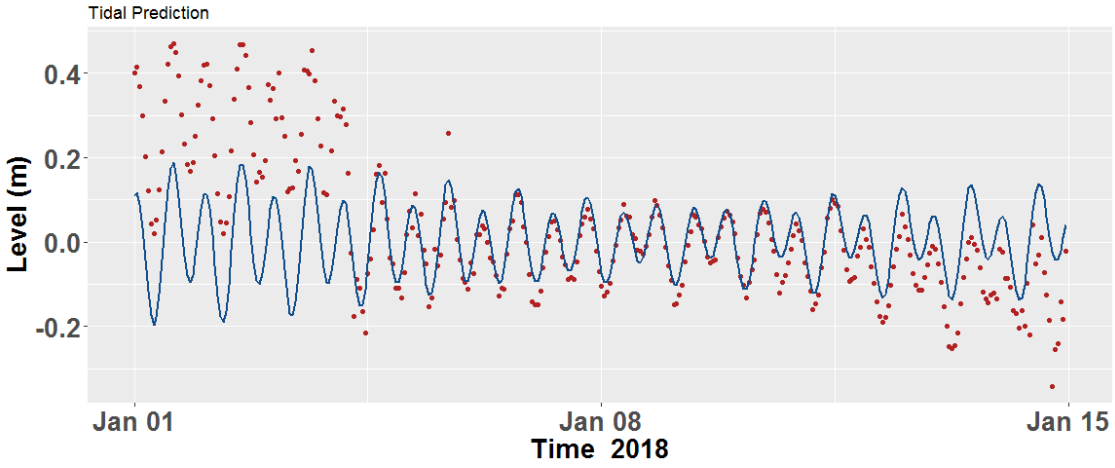


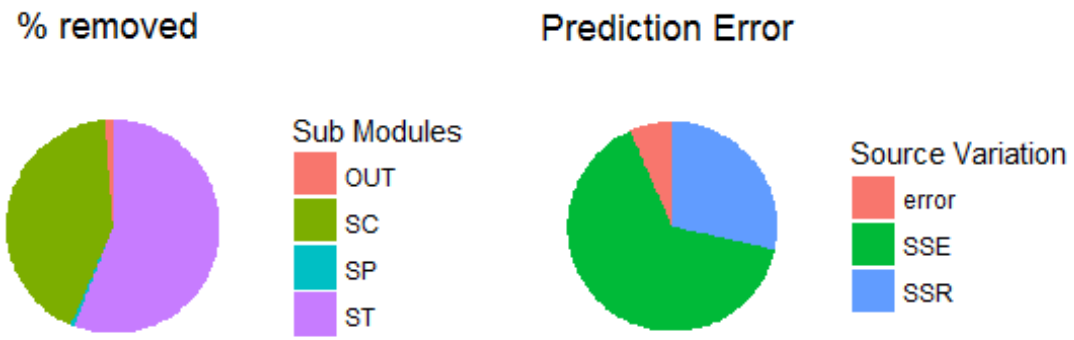


Geor

```
## [1] "Station: geor  Sensor: prs  Country: Cayman Islands"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:19:00.000"

## [1] "MSl: 0.01  Amp: 1.12  Points Removed in QC: 57.72 %"

## [1] "RMSE tide: 0.12  Set harmonics: h7"
## [2] "RMSE tide: 0.12  Set harmonics: h7"
## [3] "RMSE tide: 0.12  Set harmonics: h7"
```

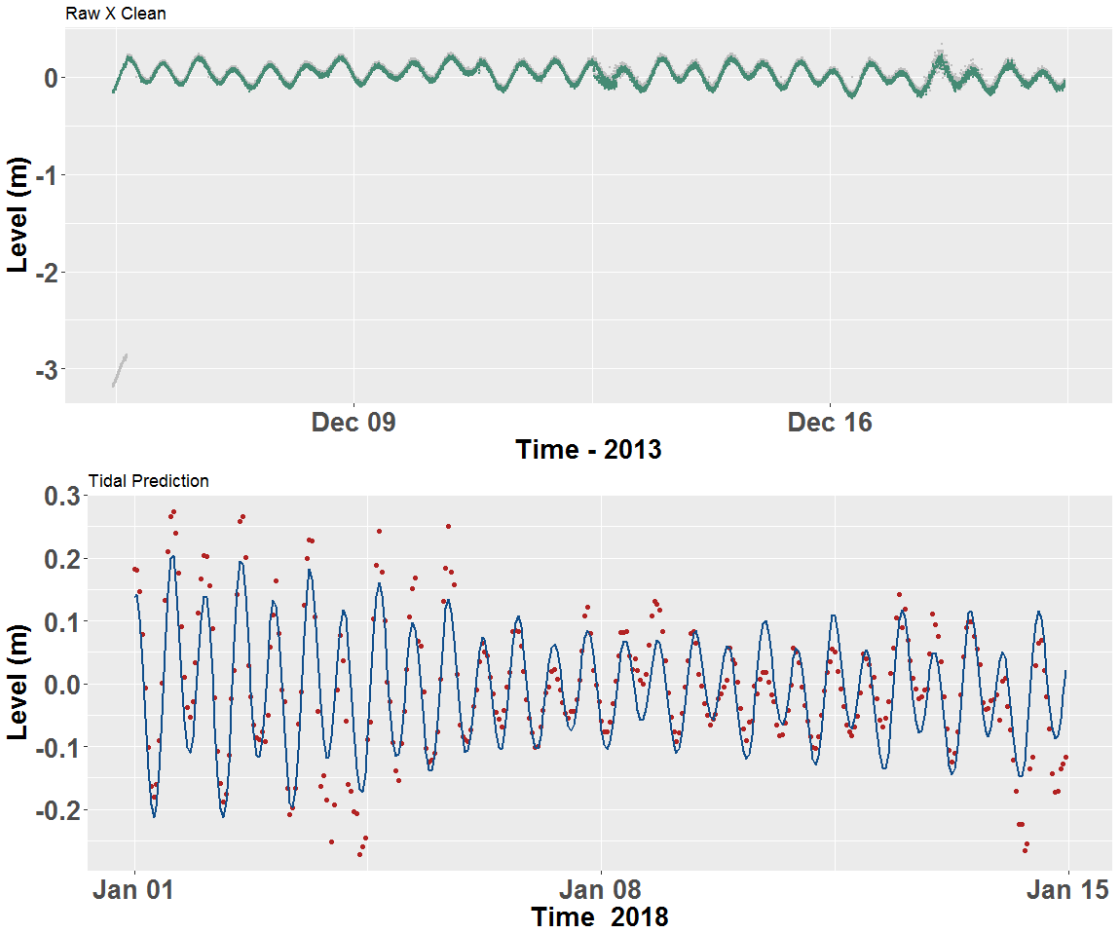


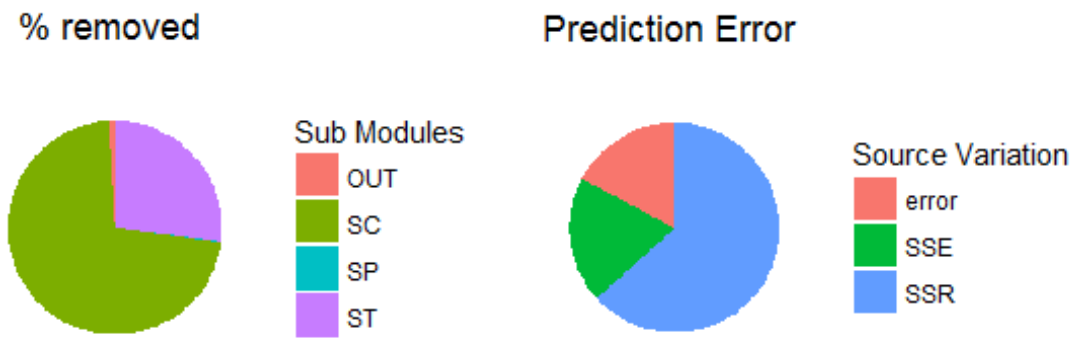




Geor

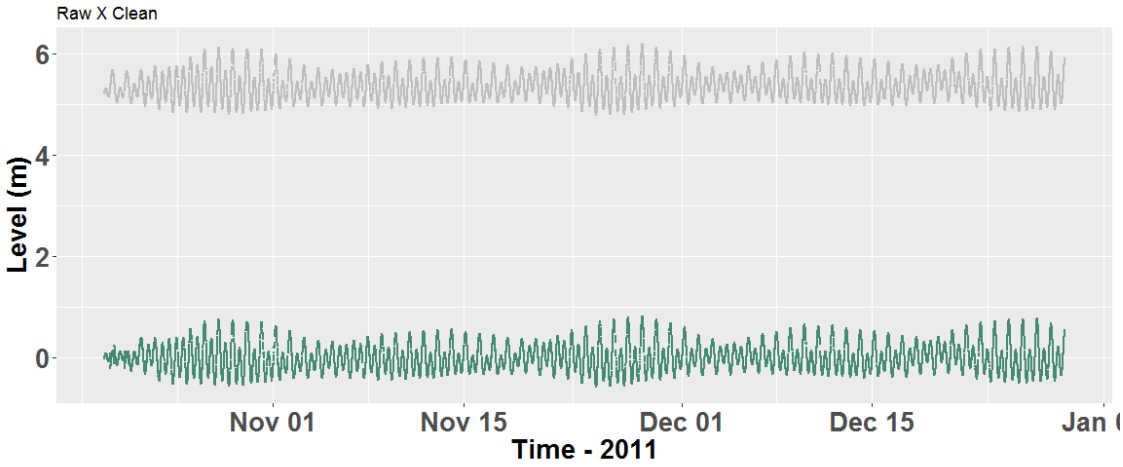
```
## [1] "Station: geor  Sensor: rad  Country: Cayman Islands"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:19:00.000"

## [1] "MSl: 0  Amp: 0.72  Points Removed in QC: 22.12 %"

## [1] "RMSE tide: 0.05  Set harmonics: h37"
## [2] "RMSE tide: 0.05  Set harmonics: h37"
```

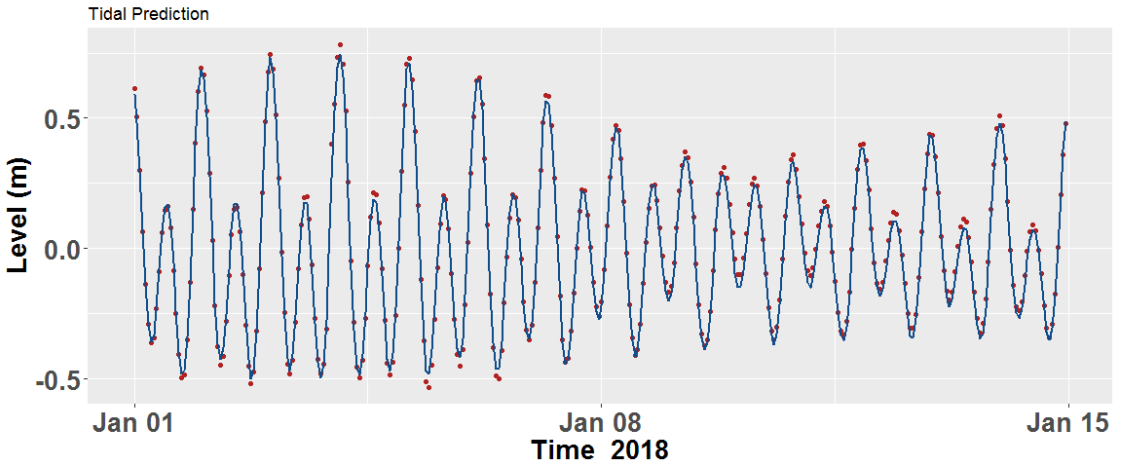


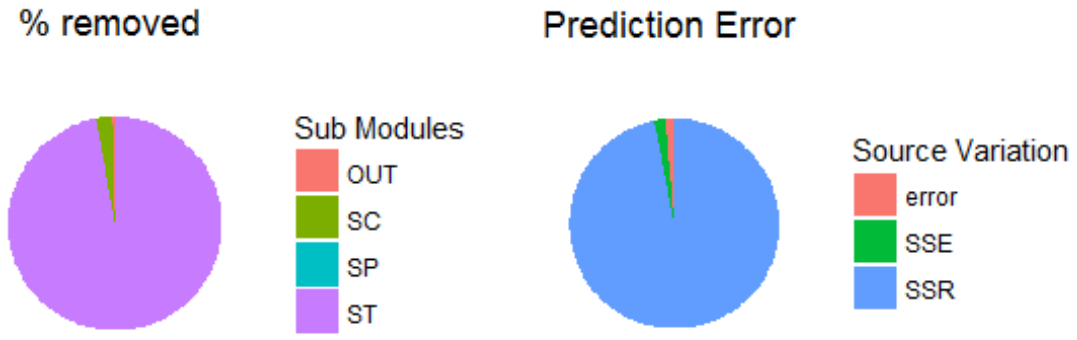


Mata

```
## [1] "Station: mata  Sensor: enc  Country: PerÃº"

## [1] "First obs: 2018-04-03 05:20:00  Last obs: 2018-04-05 22:20:00.000"

## [1] "MSl: 0  Amp: 1.62  Points Removed in QC: 18.64 %"

## [1] "RMSE tide: 0.04  Set harmonics: h37"
## [2] "RMSE tide: 0.04  Set harmonics: h37"
```

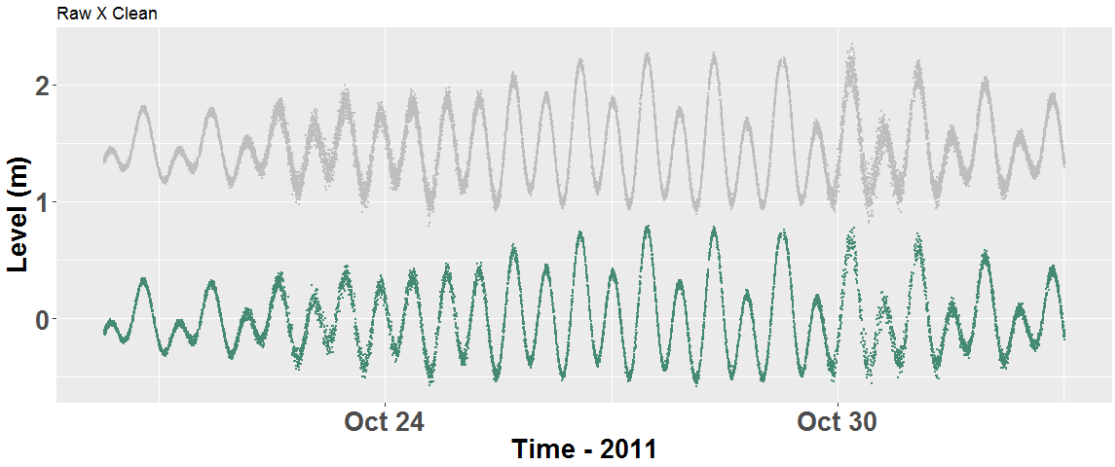


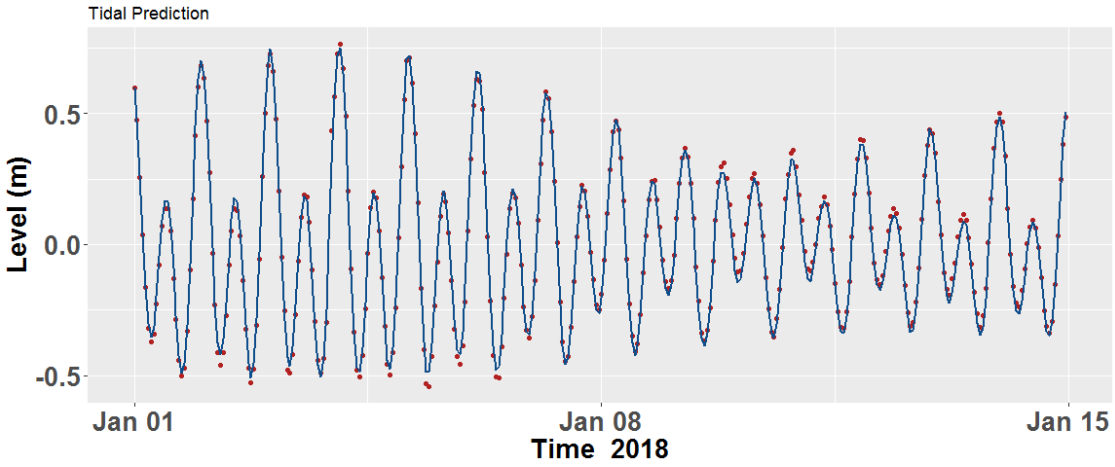


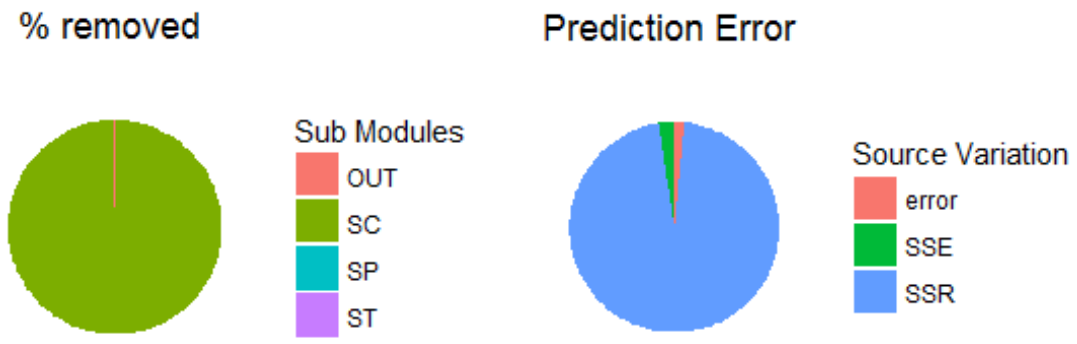


Mata

```
## [1] "Station: mata  Sensor: prs  Country: PerÃº"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:20:00.000"

## [1] "MSl: 0  Amp: 1.61  Points Removed in QC: 3.15 %"

## [1] "RMSE tide: 0.04  Set harmonics: h69"
```

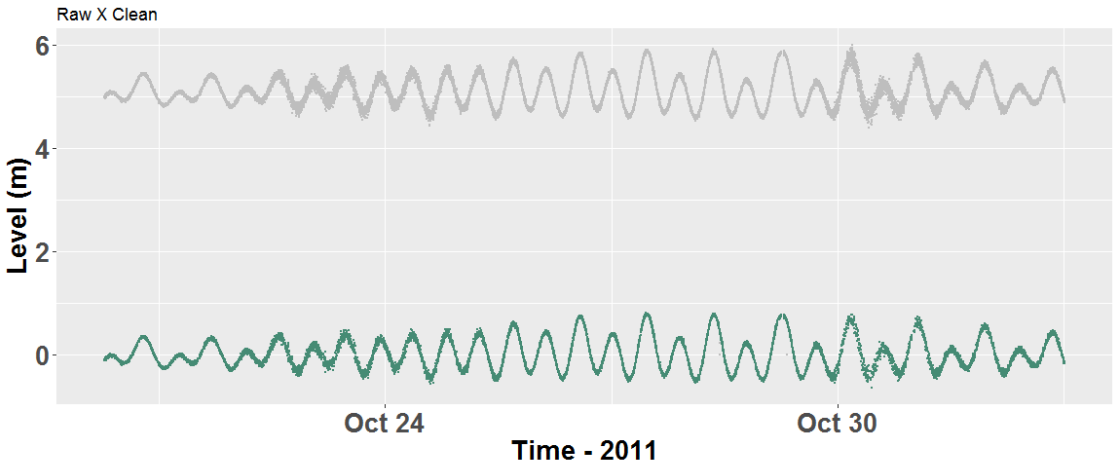


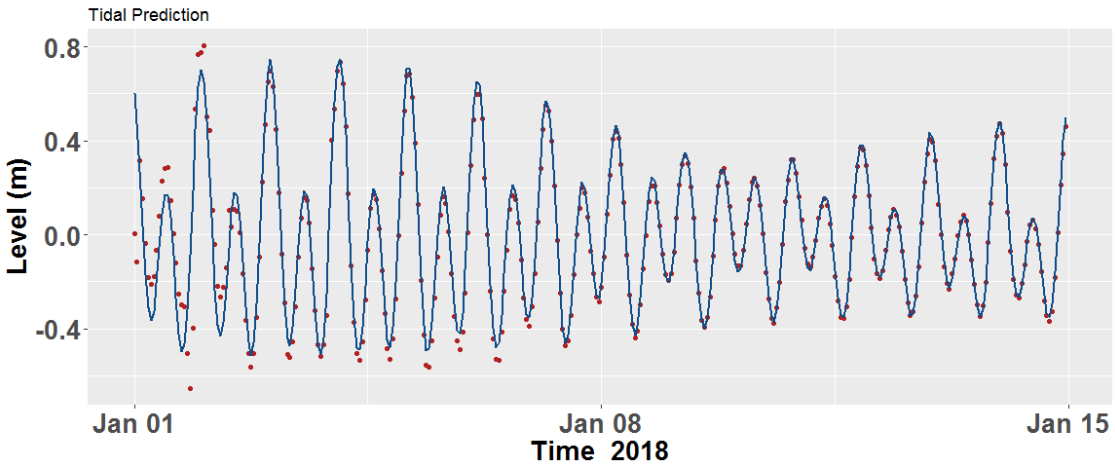


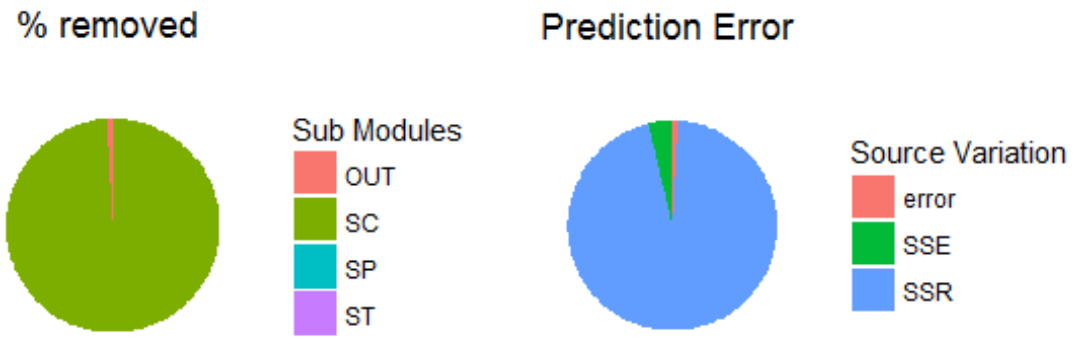


Mata

```
## [1] "Station: mata  Sensor: rad  Country: PerÃº"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:20:00.000"

## [1] "MSl: 0  Amp: 1.57  Points Removed in QC: 4.76 %"

## [1] "RMSE tide: 0.05  Set harmonics: h37"
## [2] "RMSE tide: 0.05  Set harmonics: h37"
```

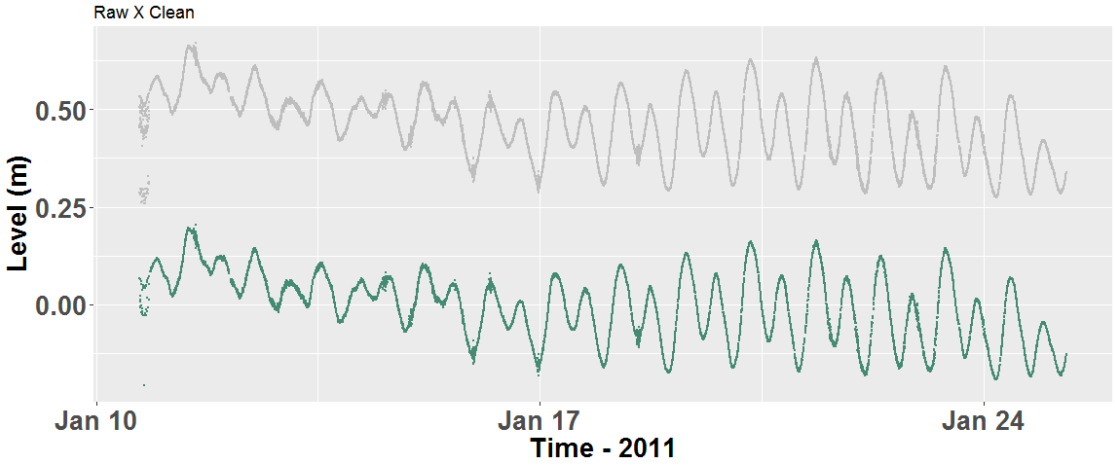


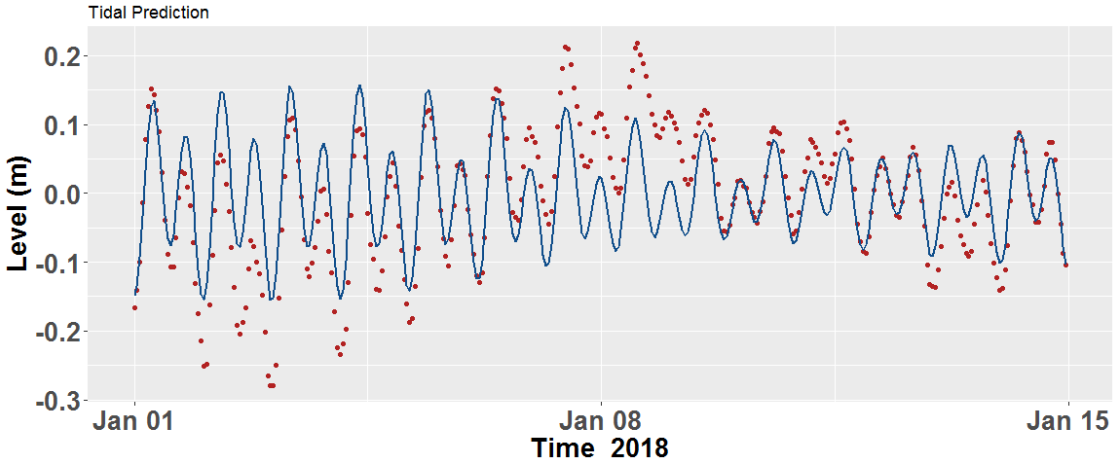


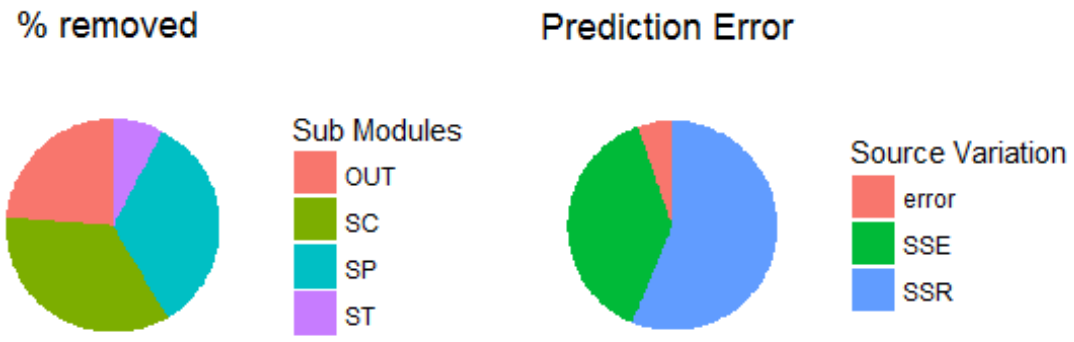


Nice

```
## [1] "Station: nice  Sensor: rad  Country: France"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:18:19.000"

## [1] "MSl: 0  Amp: 0.61  Points Removed in QC: 0.09 %"

## [1] "RMSE tide: 0.06  Set harmonics: h7"
## [2] "RMSE tide: 0.06  Set harmonics: h7"
## [3] "RMSE tide: 0.06  Set harmonics: h7"
```

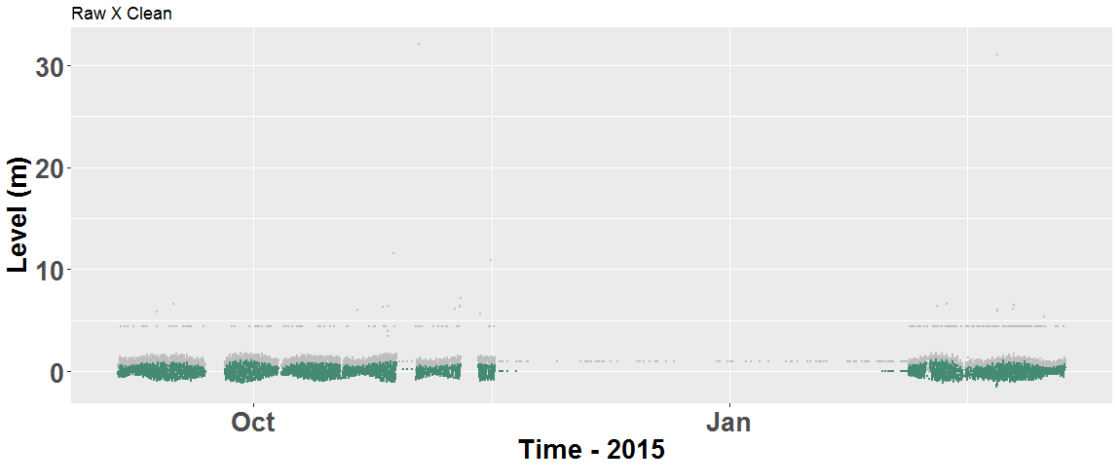


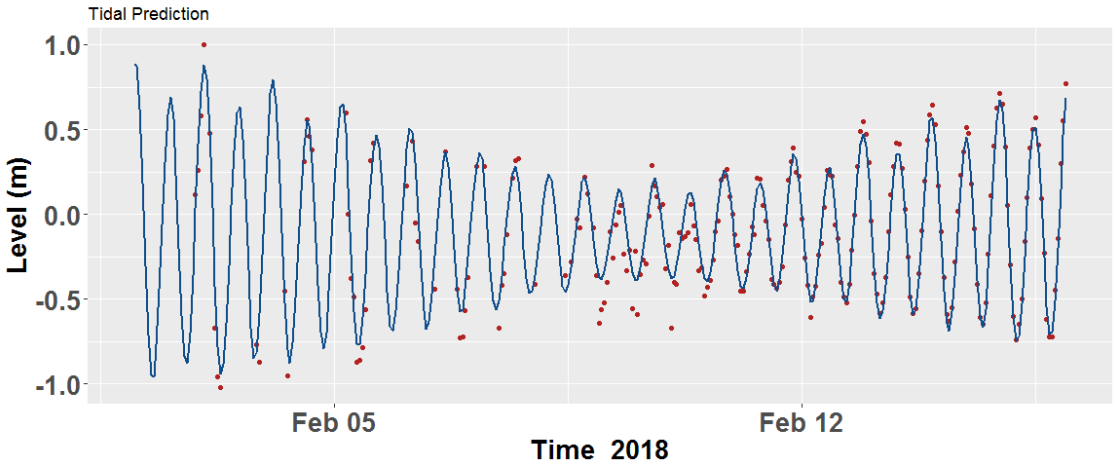


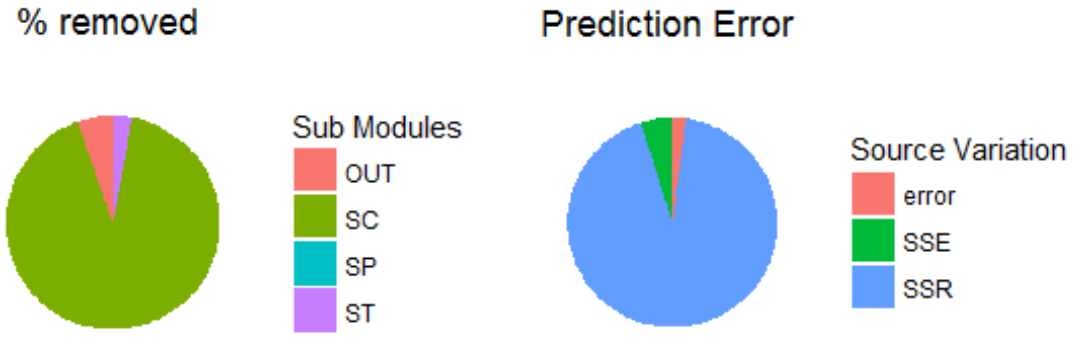


Noua

```
## [1] "Station: noua  Sensor: rad  Country: Mauritania"

## [1] "First obs: 2018-04-03 05:20:00  Last obs: 2018-04-05 21:50:00.000"

## [1] "MSl: -0.07  Amp: 2.76  Points Removed in QC: 25.49 %"

## [1] "RMSE tide: 0.11  Set harmonics: h37"
```

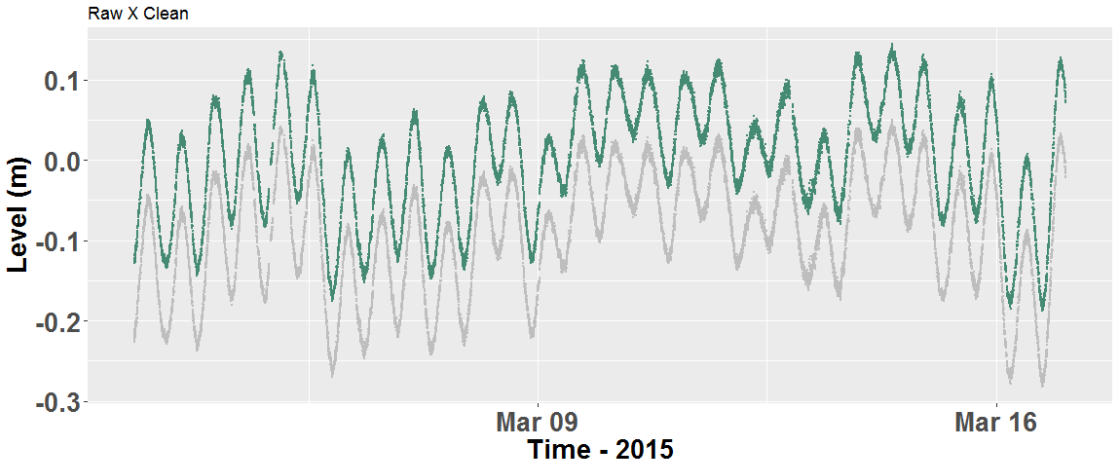


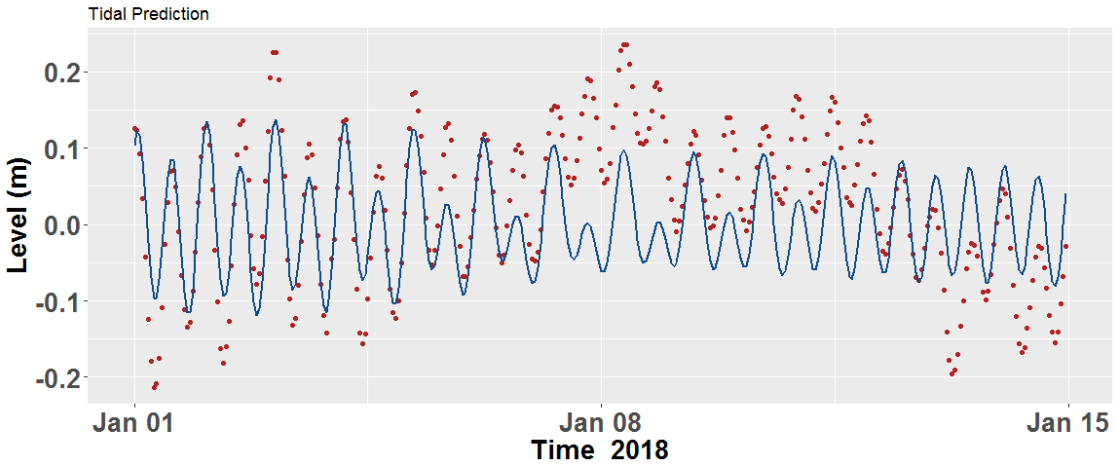


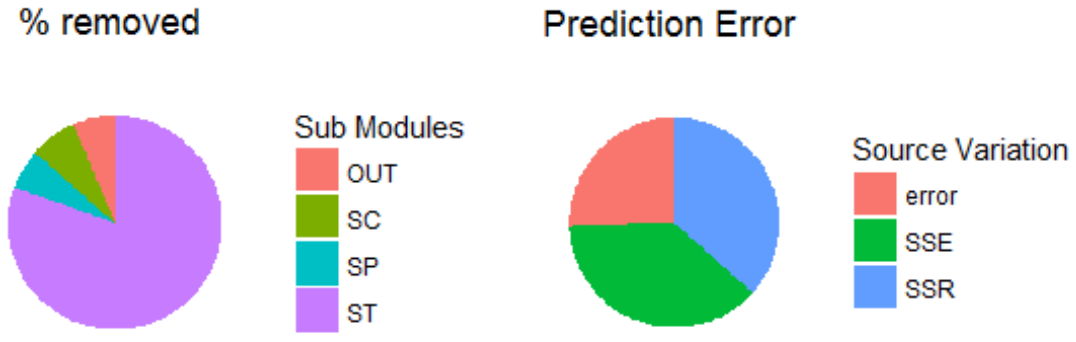


Pumo

```
## [1] "Station: pumo  Sensor: prs  Country: "

## [1] "First obs:   Last obs: "

## [1] "MSl: 0.01  Amp: 6.01  Points Removed in QC: 24.54 %"

## [1] "RMSE tide: 0.06  Set harmonics: h7"
```

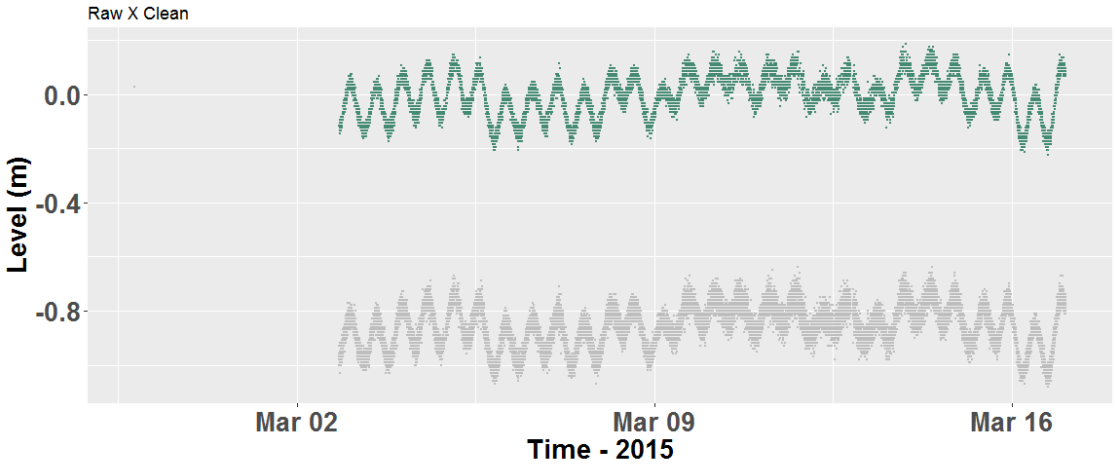


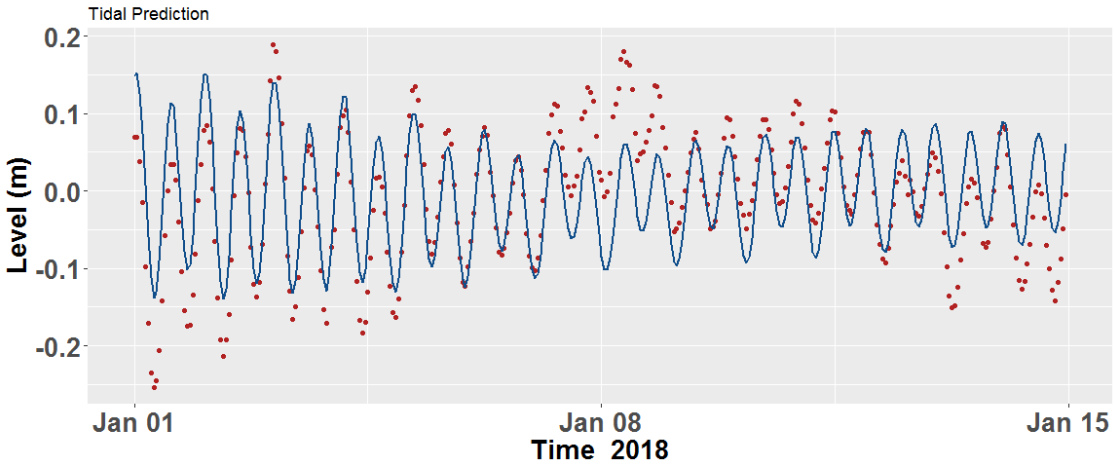


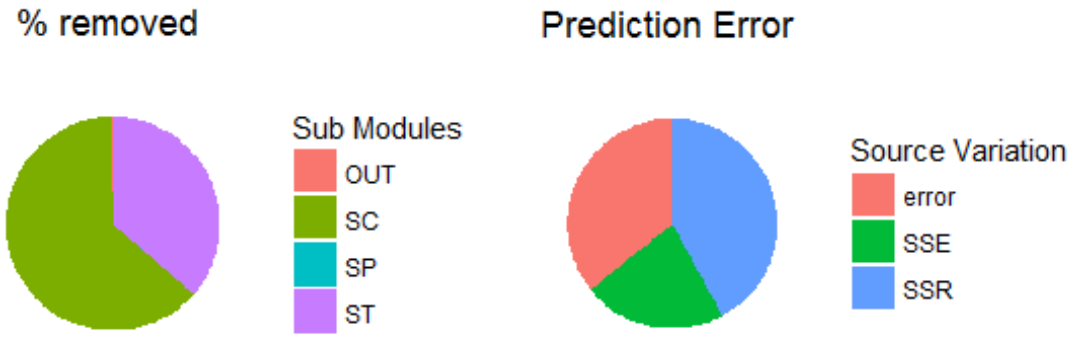


Pumo

```
## [1] "Station: pumo  Sensor: rad  Country: "

## [1] "First obs:   Last obs: "

## [1] "MSl: 0  Amp: 0.62  Points Removed in QC: 39.09 %"

## [1] "RMSE tide: 0.04  Set harmonics: h37"
## [2] "RMSE tide: 0.04  Set harmonics: h37"
```

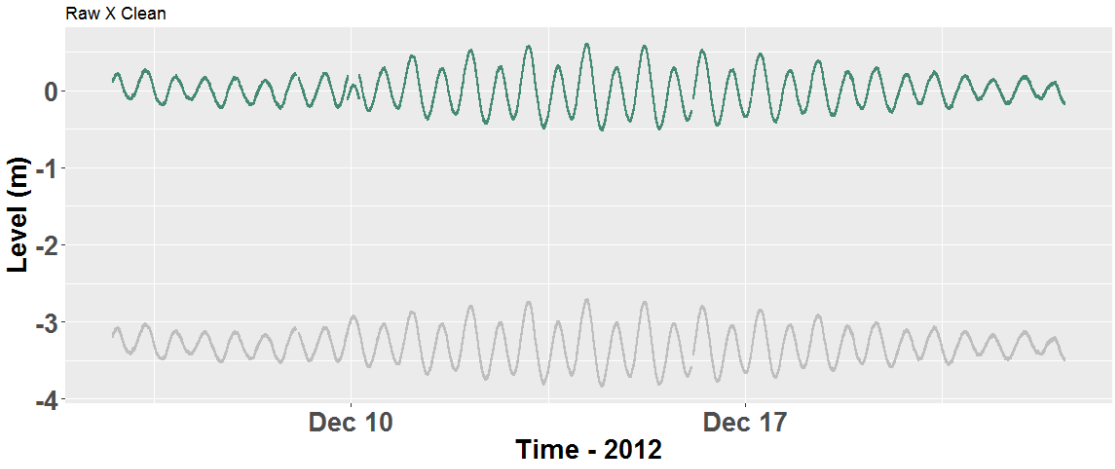


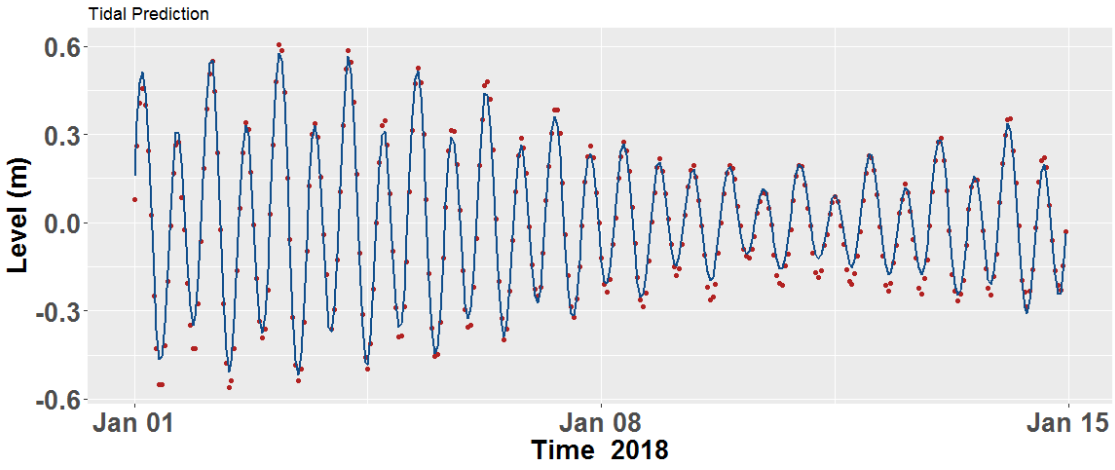


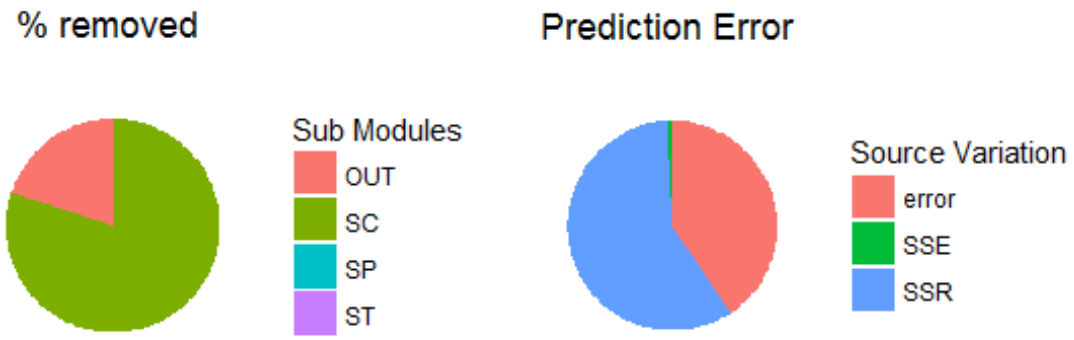


Wake

```
## [1] "Station: wake  Sensor: pwl  Country: USA"

## [1] "First obs: 2018-04-03 05:16:00  Last obs: 2018-04-05 22:20:00.000"

## [1] "MSl: 0  Amp: 1.34  Points Removed in QC: 0.16 %"

## [1] "RMSE tide: 0.04  Set harmonics: h69"
```







8.2. QC Functions

The below functions were made to apply QC to tide gauge stations. Up until the moment, the functions can be applied to any sea level time series. To be applied in a big database, some modifications to improve the processing time are still necessary.

The QC module is divided in 5 sub-modules: Breakpoints Detection, Stability Check, Outlier Detection, Speed of Change Check and Spike Detection. In each sub-module the data passes by at least one filter, and parameters can be modified in each sub-module. The main function is the QCmodule, which combines all the sub-modules and filters, and allows for filters and sub-modules to be turn on/off. After each filter, a flag is created, with 0 for suspected and removed values, and 1 for values that passed the filter.
A hourly filter function is also presented here, although this function is used for the Tidal Module. Each sub-module can also be applied separately to the dataset (instead of using the QCmodule function).

Table of Functions:

1. Breakpoint Function
2. Global Outlier Function
3. Hourly Filter Function
4. Lunar Table Function
5. Median Filter Function
6. Out-of-range Function
7. Outlier Function
8. Outlier Module Function
9. QC Module Function
10. Speed of Change Function
11. Lunar Table Function
12. Spike Detection Function
13. Spline Filter Function
14. Stability Check Function
15. Station Map Function

Breakpoint Function

Description:

Function that identifies changes in the time series. After change point identification, the time series is sliced in sets according to the change points, and then the mean is calculated for each set and removed from the data. The result is a mean referenced sea level, evolving around 0.

Usage:

breakpoint(v,t,method2 =“BinSeg”,type=“mean”)

Arguments:

v: vector with level of the time series.
t: time vector of the time series (has the same length as v).
method2: method to identify the changepoints. Default: *Binseg*. Choice between: “PELT”,

"AMOC", "SegNeigh"or "BinSeg". Given the size of the time series, BinSeg is recommendable to reduce computational cost.
type: Choose if the change point identification is made regarding the mean, variance or both. Default: *mean* Choice between "mean", "var" or "meanvar".

Details:

Function that looks for changes within the time series. A time series is considered to have a change point if, after a determined time *t*, the time series can be divided in two subsets (until *t*, after *t*), with different statistical properties, such as mean and variance (Killick & Eckley, 2014).
The function uses the R package **changepoint** for identification of the change points. For more information on the methods, please refer to Killick et al (2016).
After change point identification, the time series is sliced in sets according to the change points, and then the mean is calculated for each set and removed from the data.
The result is a mean referenced sea level, with the signals "aligned" around 0.
This is acceptable for Tide Gauge Sea Level measurements, because the resulting time series is already a relative measurement. However, this step should be applied with attention if other sources of sea level measurements are being used.

Value:

chpt: location of the change points.
flag: vector same size of v, with 1 for the position where change happens, and 0 for the rest.
changelevel: vector same size of v, with only values where change happened (rest NA) (useful for visualization).
changetime: vector with the time of the changes (same length as chpt).
level: level vectors after the mean correction.
time: time vector (same as t).

Requirements:

This function requires the R packages: changepoint (Killick et al, 2016), dplyr (Wickham, 2017).

References:

Killick, R., Eckley, I. A. (2014). changepoint: An **R** Package for Changepoint Analysis. *Journal of Statistical Software*, Vol 58, issue 3.

Global Outlier Function

Description:

Function that applies a Global Outlier Filter.
Suspected (and flagged) values are those: |values-mean| > 4*stdv. The mean and standard deviation are global values calculated for the entire time series.

Usage:

global_outlier(level, time)

Arguments:

level: level vector of the time series.
time: time vector of the time series (same length as level)

Details:
This is the first filter applied in the Outlier sub-module. It calculates a global mean and standard deviation for the time series, and uses this to identify values far from the curve. It is a gross filter. The suspected values are flagged and removed from the output time series.

Value:
clean: level vector with the suspected values replace with NA.
time: time vector of the time series (same as input).
flag: flag vector, with 1 for values that passed the filter, and 0 for suspected values.

Requirements:
This functions requires R package dplyr (Wickham, 2017).

See also:
Outlier Module Function

Hourly Filter Function

Description:
Function that creates a regular hourly grid of the data.

Usage:
hourly_filter(v,t,zone="GMT")

Arguments:
v: vector with level of the time series.
t: time vector of the time series (has the same length as v).
zone: time zone of the time vector. Default: GMT.

Details:
The Tidal Module calculates tidal prediction based a sea level time series. The minimum time frequency for the prediction is hourly values. The prediction functions accept a data set with higher frequency than hours (e.g., minute data), however the computational time and cost increase significantly with the length of the time series. On the other hand, the resolution of the prediction does not increase as much. The high frequency data is useful for other purposes, such as tsunamis and seiches studies (GLOSS, 2011). Therefore, it is better to use hourly data as input for the prediction.
This function creates a regular hourly grid based on the start and end date of the time series. Hourly values are computed by calculating the median within each hour. In case of a data gap larger than 1 hour (in the input data), the output data will have a NA.
Obs: According to GLOSS (2011), the recommended filter for Tidal Analysis is the Doodson Filter, described in Pugh (1987). Future work should aim in implementing this filter here.

Value:
The function returns a data frame with:
time: Time sequence in hourly intervals, from the start date to the final date.
Level: The median filtered value within each hour.
For one year of data, the output should have 8760 rows.

Requirements:

This function requires the R packages: timeDate (Wuertz et al, 2018), and dplyr (Wickham, 2017).

References:

Global Sea-Level Observing System - GLOSS (2011). Manual on Quality Control of Sea Level Observations, Version 1.0, 38pp. Draft.

Pugh, D. T. (1987). Tides, Surges and mean sea-level. Book. John Wiley & Sons, 472.

Lunar Table Function

Description:

Function that creates a lunartable, given a start and final date

Usage:

lunartable(sdate= as.Date("2005-12-27"),fdate=as.Date("2021-01-09"))

Arguments:

sdate: the start date to calculate the lunar phases and month. Default: "2005-12-27"
fdate: the final date to calculate the lunar phases and month. Default: "2021-01-09"

Details:

For some QC filters, it is necessary to calculate the amplitude or mean in a given month. However, the tidal patterns follow the moon cycle, and not the months of the calendar. This is not a problem when the climatological values are used (i.e., values calculated based on a long time series). But once some stations have less than a year, the use of a "lunar month" can give better results.
This function was created to give the "lunar month" and the lunar phase of a day. The default creates a table with dates since 2005 until 2021. However, this can be modified if required.

Value:

The function returns a table, with the following columns:
time: Time sequence, in days, from the start date to the final date.
moon: moon phase in radians, where 0 refers to the new moon, $\pi/2$ refers to the first quarter (Waxing), π refers to the full moon and 3 π /2 refers to the last quarter (Waning) (See Lunar Package for more information).
mooname: gives the names of the moon phases referring to the radians in the moon column.
phase: gives number 1 for New moon, 2 for Waxing, 3 for Full moon and 4 for Waning.
lunarmonth: number of the lunar month, in relation to the start date. Counting from the sdate to the fdate, a lunar month is defined as passing by the 4 lunar phases.
If defaults are called, then the number of lunarmonths is 186.

Requirements:

This function requires the package Lunar (Lazaridis, 2015), and dplyr (Wickham, 2017).

Median Filter Function

Description:

Function that applies a median filter to smooth over spikes. This filter is part of the Spike sub-module.

Usage:

medfilt(v,t,n=3,method1="runmed")

Arguments:

v: Variable vector of the time series to be smoothed over (e.g. sea level).
t: Time vector of the time series (same length as v).
n: Size of the window to apply the median filter. Default: n=3.
method1: Choice of the function used to calculate the median. Default: method1 ="runmed". Choice between: "runmed" (from in-built stats package), "rollapply" (from zoo package), " fractal" (from fractal package).

Details:

Function that applies a Median Filter over the time series. The window of the median can be chosen with the parameter *n*, and the function used to calculate the median can be chosen with the parameter *method1*. After calculating the median, it also makes a Median Test: if the level is higher than the absolute values of the mean + stdv, then the value is flagged and replaced with NA. This filter is part of the Spike sub-module.

Value:

Level_clean: level vector after passing by the Median Test.
Flag: flag vector for the Median Test, with 1 for values that pass the test, and 0 for suspected and removed values.
Med: the result of the level vector after passing by the Median Filter.

Requirements:

This function requires the R packages: zoo (Zeileis et al, 2017), dplyr (Wickham, 2017) and package fractal (Constantine, 2017), in case method1="fractal".

Out-of-range Function

Description:

Function applies a filter to detect out-of-range values, based on OPPE in Gloss (2011). Out-of-range values are those beyond the seasonal limit. For a given area and month, the seasonal limit is defined as *2*stdv +- mean*. This filter is part of the Outlier sub-module.

Usage:

out_of_range(level, time, lunarmonth = NULL,lunarphase = NULL)

Arguments:

level: level vector of the time series.
time: time vector of the time series.
lunarmonth: parameter to define if the seasonal limit is calculated according to lunar month or according to calendar months. Default: NULL. This vector is resultant from the **lunartable** function.
lunarphase: parameter to define if the seasonal limit is calculated according to lunar phases,

or according to calendar months. Default: NULL. This vector is resultant from the **lunartable** function.

Details:

Function applies a filter to identify out-of-range values based on the QC applied by OPPE, described in the GLOSS QC Manual (2011). According to the manual, out-of-range values are those beyond the seasonal limit. For a given area and month, the seasonal limit is defined as 3*stdv +- mean.*
The manual suggests to use *2stdv*, however the multiplication factor was changed to 3, because tests showed that it was removing real values from the time series.
Out-of-range values are flagged, and not considered in the subsequent checks.
There is the option to calculate the seasonal limit according to lunar phase (New, Waxing, Full, Waning), lunar month (4 phases of the moon), or according to calendar months (january, february, …).
This filter is part of the Outlier sub-module.

Value:

clean: Level vector after passing the filter. Suspected values have been removed and replaced with NA.
flag: Flag vector from the filter, with 1 for values that passed the filter, and 0 for suspected values.

Requirements:

This function requires R package dplyr (Wickham,2017).

References:

Global Sea-Level Observing System – GLOSS. (2011). Manual on Quality Control of Sea Level Obsevations, Version 1.0, 38pp. Draft.

See also:

Lunar Table Function; Outlier Module Function

Outlier Function

Description:

Function that applies a filter to identify outliers according to the IOC Sea Level Monitoring Facility. An outlier is a value that subtracted from the median exceeds a tolerance value. The tolerance is calculated by 3 * | Percentile 90 - median|.

Usage:

outlier_gloss(level, time, lunarmonth=NULL)

Arguments:

level: level vector from the time series
time: time vector from the time series (same length as level)
lunarmonth: parameter to define if the seasonal limit is calculated according to lunar month or according to calendar months. Default: NULL. This vector is resultant from the **lunartable** function.

Details:

Function applies a filter to identify outliers according to the IOC Sea Level Monitoring Facility (http://www.ioc-sealevelmonitoring.org/service.php). According to the treatment, an outlier is a value, that after subtracted from the median, exceeds a tolerance value.The tolerance is calculated by 3 * | Percentile 90 - median.
Outliers are flagged, and not considered in the subsequent checks.
There is the option to calculate the tolerance according to the lunar month (4 phases of the moon) or according to calendar months (january, february, …).
This filter is part of the Outlier sub-module.

Value:

clean: Level vector after passing the filter. Suspected values have been removed and replaced with NA.
flag: Flag vector from the filter, with 1 for values that passed the filter, and 0 for suspected values.

Requirements:

This function requires R package dplyr (Wickham,2017).

Outlier Module Function

Description:

Function that compiles the 3 filters (Global Outlier, Outlier Gloss, Out-of-range filters).

Usage:

outlier_mod(level, time, GO = T, OG = T, OR = T, lunarmonth = F, lunarphase = F, clim = NULL)

Arguments:

level: level vector from the time series.
time: time vector from the time series.
GO: Option to turn on/off the Global Outlier filter. Default: GO = T. Choice between T/F. If GO= T, it applies the global_outlier filter. If GO=F, then it skips it.
OG: Option to turn on/off the Outlier Gloss filter. Default: OG = T. Choice between T/F. If OG= T, it applies the global_outlier filter. If OG=F, then it skips it.
OR: Option to turn on/off the Out-of-range filter. Default: OR = T. Choice between T/F. If OR= T, it applies the Out-of-range filter. If OR=F, then it skips it.
lunarmonth: Option to calculate outlier_gloss and out-of-range filters according to the lunar month. Default: lunarmonth=F. Choice between T/F. If lunarmonth = T, outlier_gloss and out_of_range will be applied for the lunarmonth. See functions for more details.
lunarphase: Option to calculate the out-of-range filter according to the lunar month. Default: lunarmonth=F. Choice between T/F. If lunarmonth = T, out_of_range will be applied for the lunarmonth. See function for more details.
clim: vector[12x2] with the climatological means and standard deviation calculated for a station. This is an optional input. (If clim is given, it will be applied for out_of_range. See functions for more details).

Details:

This function compiles the three filters that make up the Outlier Module. Each filter can be turned on/off. However, at least one filter should be on to apply the module.

Value:

Time: time vector correspondent to the level vector.
Level_orig: level as the raw level, BUT with -999 values replace for NA.
Level_clean: level after the outlier detection. Suspected values have been replaced with NA.
Flag: vector with 0 for values removed and 1 for values kept, combining the different steps.
df_flag: data frame with the flags for each filter applied.

Requirements:

This functions requires R package dplyr (Wickham, 2017), Lunar (Lazaridis, 2015)

See also:

Global Outlier Function; Lunar Table Function; Out-of-range Function; Outlier Function.

QC Module Function

Description:

Function that compiles the different sub-modules of the QC Module.

Usage:

QCmodule(v,t, ALL = T, BP= T, ST = T, OUT = T, SC = T, SP =T, go = T, og= T, or = T, lm = F, lp = F,cli=NULL, method1 = "runmed",
method2="BinSeg",lagi=120,ni=3,ka=3,wid=2,filtro=1,type="mean")

Arguments:

v: level vector of the time series.
t: time vector of the time series (same length as t).
ALL: Option to apply directly all the sub-modules. Default = T. If ALL=F, then at least one module should be turned off (required in the following parameters).
BP: Option to turn on/off the Breakpoint sub-module. Default: T. If =F, then this sub-module is skipped.
ST: Option to turn on/off the Stability Check sub-module. Default: T. If =F, then this sub-module is skipped.
OUT: Option to turn on/off the Outlier sub-module. Default: T. If =F, then this sub-module is skipped.
SC: Option to turn on/off the Speed of Change Check sub-module. Default: T. If =F, then this sub-module is skipped.
SP: Option to turn on/off the Spike sub-module. Default: T. If =F, then this sub-module is skipped.
go,**og**,**or**: parameters of the Outlier Module, to choose which filter to (des)active.
method1: refers to the Median Filter method applied in the Spike Module.
method2: refers to the Changepoint method applied in the Breakpoint Module.
ka,**ni**,**wid**: parameters of the Spike Module.
filtro: Option to choose if you want both Median and Spline filter (filtro=1), only the median filter (filtro=2), or only the spline filter (filtro=3) on the Spike Module. Default = 1.
lagi: a parameter of the Stability Check Module.
lm,**lp**: lunarmonth and lunarphase are options of the Outlier and Speed of Change Modules to apply treatments according to the lunarmonth or lunar phase.

cli: vector with the climatological means and amplitudes, to be used as clim in Out_of_Range function of the Outlier Module.
type: parameter of the Breakpoint Module.

For more information, look at each sub-module and filter.

Details:

Function that combines the different sub-modules and filters of the QC module. It allows to turn on/off each sub-module, and to change their parameters.
Returns two data frames: one with vectors of the same length as the input vectors. This data frame contains the time vector, the original level, the clean vector (where the suspected values have been replaced with NA), a flag vector for each module, and a final flag vector (result of the combination of the prior flags) with 0 for the values removed and 1 for the values kept.
The other data frame contains only the time and clean level vectors, however the NA's have been removed. So the length of this vectors is smaller (or equal in case of no suspected values identified) to the input vectors.
It is important however to keep the original vectors accompanied only with the flags for purposes of data storing.

Value:

xQC: Data frame with:

- **Time**: time vector of the time series (same length as input).
- **Level_orig**: original level of the time series, same as input.
- **Level_clean**: clean level after the QC module, where suspected values have been replaced by NA's (same length as Level_orig).
- **FlagNA**: flag with 0 for missing values that are registered as -999 in the original level. This flag is not considered when combining the flags of the different modules.
- **FlagST**: flag from the Stability Check sub-module.
- **FlagOUT**: flag from the Outlier sub-module.
- **FlagSC**: flag from the Speed of Change Check sub-module.
- **FlagSP**: flag from the Spike detection sub-module.
- **Flag_final**: Combination of the previous flags.

xclean: Data frame with:

- **Level**: clean level, where the suspected values have been removed. This vector has length smaller (or equal in case no suspected values were identified) to the input level.

  Time: time vector of the time series, with same length as the level vector.

  If an Error occurs after the application of a module, the function breaks, and returns a message indicating in which module was the problem.

Requirements:

This function requires R packages: dplyr, lunar, changepoint, zoo, fractal, tseries.

See also:

Breakpoint Function; Global Outlier Function; Lunar Table Function; Median Filter Function; Out-of-range Function; Outlier Function; Outlier Module Function; Speed of Change Function; Lunar Table Function

*Spike Detection Function; Spline Filter Function; Stability Check Function***.**

Speed of Change Function

Description:

Function that applies a filter to check the speed of change between two consecutives measurements.

Usage:

speed_change(level, time, lunarmonth=F)

Arguments:

level: level vector of the time series.
time: time vector of the time series.
lunarmonth: Option to calculate the amplitude based on the lunarmonth. Default = F, amplitude is calculated according to calendar month.

Details:

Function that applies a filter to check the speed of change between two consecutives measurements.
Considering two consecutive measurements h1 and h2, the difference between h1 and h2 cannot surpass a tolerance value. The tolerance value is calculated in relation to the frequency of the time series and the amplitude.
This sub-module is based on the BODC QC Manual. The filter tests if |h1-h2| > |tol*Amp|. The flag is put for h2. The tol value is calculated based on the frequency (time difference between h1 and h2).
There is the option to calculate the tolerance according to the lunar month (4 phases of the moon) or according to calendar months (january, february, …).

Value:

Level_clean: vector with the level after the speed of change filter.
Flag: vector with 0 for values that were removed and 1 values kept.

Requirements:

This function requires R packages: dplyr (Wickham, 2017), zoo (Zeileis et al, 2017) and lunar (Lazarids, 2015).

References:

British Oceanographic Data Centre - BODC. (2007). Data Quality Control Procedures, Version 3.0.(September 2007), 75pp.

See also:

Lunar Table Function

Spike Detection Function

Description:

Function to detect and remove spikes (smaller spikes that pass by the outlier module) from the signal.
Uses two filters: Median Filter and a Spline Filter.

Usage:

spike(v,t,wd=12,n=3,k=3,method1="runmed",filt=1)

Arguments:

v: level vector of the time series.
t: time vector of the time series.
wd: window, in hours, to apply the spline. Default = 12.
k: degree of the standard deviation used to dected the spikes.Default = 2.
(e.g. k= 2, a value that differs from the median by 2*sdtv is removed)
n: window size for the Median filter. Default = 3. must be an odd number!
method: methods for the Median Filter. Default = "runmed".
filt: Option to choose if you want both Median and Spline filter (filt=1),only the median filter (filt=2), or only the spline filter (filt=3). Default = 1.

Details:

Function to detect and remove spikes (smaller spikes that pass by the outlier module) from the signal.
Uses two filters: Median Filter and a Spline Filter.
It tests if: *|value|* exceeds the *|spline-stdv|*k.*
For filt=3: |value| exceeds the *|median-stdv|**k
For filt=2: with the median resultant from the median filter function. *|value|* exceeds the *|spline-stdv|*k*
For filt=1: with the spline calculated with the median result of the median filter function.

Value:

Level_clean: vector with the level after the speed of change filter.
Flag: vector with 0 for values that were removed and 1 values kept.

Requirements:

This functions requires R packages: zoo (Zeileis et al, 2017), dplyr (Wickham, 2017) and fractal (Constantine, 2017)

See also:

Median Filter Function; Spline Filter Function

Spline Filter Function

Description:

Function that applies a spline to remove spikes.

Usage:

splinefilt(v,t,wd=NULL)

Arguments:

v: vector of the variable to be smoothed over (e.g. sea level).
t: time vector of the time series, correspondent to v.
wd: the size of the window, in hours, to apply the spline filter.

Details:

This function applies a spline (piecewise polynomial) to smooth over the time series, and detect not so obvious outlier.
The filter is only actually applied in the Spike function. Here, only the smooth spline is calculated, based on the required window.

Value:

sp: the output of the function smooth.spline (from build-in Stats package).

Requirements:

This function requires the R package zoo (Zeileis et al, 2017), and the built in Stats package.

See also:

Lunar Table Function

Spike Detection Function

Stability Check Function

Description:

Function that looks if the same values is being repeated for more than 2 hours (Flat line kept for more than 2 hours). It detects mal-functioning of the sensor.

Usage:

stability_check(v,t,lag)

Arguments:

v: vector of the variable to be smoothed over (e.g. sea level).
t: time vector of the time series, correspondent to v.
lag: time interval/range we want to check our stability.

Details:

Function that looks if the same values is being repeated for more than 2 hours. It detects mal-functioning of the sensor. It tests for equal values.

Value:

Level_clean: vector with the level after the speed of change filter.
Flag: vector with 0 for values that were removed and 1 values kept.

Requirements:

This function requires the R packages zoo (Zeileis et al, 2017), tseries (Trapletti et al, 2018), and dplyr (Wichkam, 2017).

See also:

QC Module Function

Station Map Function

Description:

Function to plot directly the location of the station on the world map, given the latitude and longitude.

Usage:

station_map(lat,lon,name,save=NULL)

Arguments:

Lat: latitude of the station (as value from 0 to +-90).
Lon: longitude of the station (as value 0 to 360, or 0 to +-180).
Name: name of the station (as character).
Save: pathway to save the map (optional)

Details:

Creates a world map and plot the station on top of it.

Value:

mp: a ggplot figure. to visualize it: print(mp).

Requirements:

ggplot2 (Wckham & Chang, 2016)

8.3. Harmonics List

Table A1. Table with main 37 tidal harmonics names and brief description. Source: NOAA, 2018 https://tidesandcurrents.noaa.gov/stations.html?type=Harmonic+Constituents.

Constituent #	Name	Description
1	M2	Principal lunar semidiurnal constituent
2	S2	Principal solar semidiurnal constituent
3	N2	Larger lunar elliptic semidiurnal constituent
4	K1	Lunar diurnal constituent
5	M4	Shallow water overtides of principal lunar constituent
6	O1	Lunar diurnal constituent
7	M6	Shallow water overtides of principal lunar constituent
8	MK3	Shallow water terdiurnal
9	S4	Shallow water overtides of principal solar constituent
10	MN4	Shallow water quarter diurnal constituent
11	NU2	Larger lunar evectional constituent
12	S6	Shallow water overtides of principal solar constituent
13	MU2	Variational constituent
14	2N2	Lunar elliptical semidiurnal second-order constituent
15	OO1	Lunar diurnal
16	LAM2	Smaller lunar evectional constituent
17	S1	Solar diurnal constituent
18	M1	Smaller lunar elliptic diurnal constituent
19	J1	Smaller lunar elliptic diurnal constituent
20	MM	Lunar monthly constituent
21	SSA	Solar semiannual constituent
22	SA	Solar annual constituent
23	MSF	Lunisolar synodic fortnightly constituent
24	MF	Lunisolar fortnightly constituent
25	RHO	Larger lunar evectional diurnal constituent
26	Q1	Larger lunar elliptic diurnal constituent
27	T2	Larger solar elliptic constituent
28	R2	Smaller solar elliptic constituent
29	2Q1	Larger elliptic diurnal
30	P1	Solar diurnal constituent
31	2SM2	Shallow water semidiurnal constituent
32	M3	Lunar terdiurnal constituent
33	L2	Smaller lunar elliptic semidiurnal constituent
34	2MK3	Shallow water terdiurnal constituent
35	K2	Lunisolar semidiurnal constituent
36	M8	Shallow water eighth diurnal constituent
37	MS4	Shallow water quarter diurnal constituent