Low-pass filter application for noise removal in water quality data with high temporal resolution

Elisa Coraggio (✉ elisa.coraggio@bristol.ac.uk)
University of Bristol

Claire Gronow
University of Bristol

Theo Tryfonas
University of Bristol

Dawei Han
University of Bristol

Research Article

Keywords: water quality, wireless sensor networks (WSN), noise removal, signal processing, low pass filters

Posted Date: August 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3284948/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Low-pass filter application for noise removal in water quality data with high temporal resolution

Elisa Coraggio 1 *, Claire Gronow 1, Theo Tryfonas 1, Dawei Han 1.

1 Department of Civil Engineering, University of Bristol, Bristol, United Kingdom;
elisa.coraggio@bristol.ac.uk (E.C.); claire.gronow@bristol.ac.uk (C.G.); theo.tryfonas@bristol.ac.uk (T.T.);
d.han@bristol.ac.uk (D.H.)

*Corresponding author: elisa.coraggio@bristol.ac.uk;

ABSTRACT

Large datasets with high temporal resolution are becoming widely available through the use of wireless sensors and other low-effort, automated data collection techniques. The higher the sampling frequency is, the more obvious and significant the noise will be due to the highly unrealistic oscillations in the observations. Machine learning techniques work well with large amounts of data, but it is essential to ensure that the data collected is as clean as possible from noise; otherwise, the machine learning algorithm will struggle to predict the actual data and instead attempt to reproduce the noise. This study explores the use of four low pass filters: Butterworth, Chebyshev I, Chebyshev II and Savitzky-Golay filter for removing noise from water quality dataset with high temporal resolution. This study describes how the filters are implemented and gives advice on how to evaluate the filters’ capability to reduce noise and preserve signal features. The method is applied to five water quality parameters based on a water quality dataset with a 5 minutes resolution collected in an urban surface water body in Bristol, United Kingdom. Based on the results of this study, it has been found that for the analysed water quality parameters (conductivity, water temperature, dissolved oxygen, and fDOM) Butterworth filter with a cut-off frequency between 2.33E-05 Hz~12 hours and 4.5E-04 Hz~6 hours is the filter that allows the best compromise between noise removal and signal preservation.

KEYWORDS: water quality, wireless sensor networks (WSN), noise removal, signal processing, low pass filters

1. INTRODUCTION

In recent years, rapid developments have been made in wireless sensor networks (WSN) for various applications in environmental monitoring, such as air quality monitoring, climatological monitoring, forest monitoring, water quality and water distribution systems monitoring (Watras et al., 2014; Pule et al., 2017; Post et al., 2018; Munir et al., 2019). The application of WSN for water quality is being extensively studied, employing different sensors capable of continuously collecting and transmitting water quality measurements at high temporal resolution (Perelman and Ostfeld, 2013).

WSNs have the advantage of being able to offer freedom in the selection of the sampling frequency for monitoring water quality parameters. Systems monitored at low frequency have been recognised to be prone to uncertainty (Birgand et al., 2013); and are also unsuitable for the data-hungry machine learning tools for timeseries predictions that are becoming more prevalent (Strobl and Robillard, 2008). Thus, reasons for seeking high-frequency data from WSNs are to reduce uncertainty and improve our current understanding of behaviours and trends of observed water quality parameters, and explore the application of machine learning tools to improve water quality prediction models.

Recognised challenges in WSN applications for water quality monitoring include security concerns, network communication, energy management and data storage and post-processing. However, current research on data processing for WSN-based WQM systems typically does not focus on issues specific to the WQM process that arise from the particularly challenging water environment, including biofouling, sensor drift, and underwater
signal propagation, that add noise to the collected data. This noise becomes more significant when the data collected has a high temporal resolution. (Tapparello et al., 2017)

Counterintuitively, in timeseries prediction models, high temporal resolution datasets can reduce the model’s performance by increasing the noise and reducing the noise-to-signal ratio in the analysed signal.

In time series forecasting, the presence of noisy data can significantly damage the quality of the final predictions and the overall performance of the prediction model. This is a typical issue in the domain of time series forecasting, where the order of the data points cannot be mixed due to the crucial role of the temporal dependency between points in temporal sequence.

More attention should be paid to the data post-processing to improve the quality of the collected data, especially by utilising noise reduction techniques.

The precedent for ‘cleaning up’ large datasets in the field of hydrology started when Schreiber and Grassberger (1991) explored noise reduction approaches with the aim of finding chaotic behaviour in hydrological daily timeseries. In their study, after reducing the noise, the clean signal was considered for further analysis (i.e. prediction). This work has encouraged others to build on the idea of improving model’s predictions by cleaning the noisy signal.

In this framework, low pass filters have been applied in the past in the field of hydrology by Kawamura et al., (1998) on monthly sea level pressure data.

Building on this work, Elshorbagy, Simonovic and Panu (2002), discouraged smoothing the original hydrological time series by reducing the noise and using them for further analysis. The reason for questioning this approach is that the removed component of the original signal classified as noise might contain the noise, but it may also carry part of the signal. However, it can be highly beneficial for the water resources field to investigate methods of eliminating noise from hydrologic data, as long as it can be verified without a doubt that the eliminated component is indeed noise.

In recent literature, different denoising techniques have been applied to improve the performance of water quality prediction models based on machine learning algorithms; Lu and Ma (2020) propose an advanced denoising method (CEEMDAN) to pre-process raw data to improve the performance of Extreme Gradient Boosting (XGBoost) and Random Forest (RF) water quality prediction models using an hourly dataset. The model achieves better accuracy in short-term forecasting of the water quality data. Xu and Liu (2013) combine BPNNs and wavelet transform to build a prediction model for water quality using an hourly dataset. It achieves a high training speed and strong robustness. Ahmed et al. (2019) implement enhanced wavelet denoising techniques, and adopt a prediction model based on them to improve the accuracy of the water quality prediction effectively. These filters reduce the impact of noise in the time series on prediction accuracy.

Although these studies present solutions to improve the performance of WQ ML prediction models, they were all based on limited numbers of data points (below 2000 data points), which does not represent a high time resolution dataset. They do not capture long-term features of a water quality timeseries; therefore, the validity of their results on a high temporal resolution dataset has been questioned by Bi et al. (2021).

Bi et al. (2021) successfully used the Savitzky-Golay (SG) filter on a 4 hours time resolution dataset with around 10,000 data points testing the filter with different window sizes. Comparing SG filters to the Moving Median and Moving Average filters, SG filters have been proven being able to retain well the characteristics of a time series while removing its noise.

This work proposes the use of noise filtering solutions based on signal processing techniques testing them on high time resolution WSN data with the aim of keeping the desired high temporal resolution but reducing the noise-to-signal ratio.

These techniques are successfully applied in several signal-rich fields: electronics and telecommunication, radio, television and audio recording, radar, control systems, music synthesis, image processing, and computer graphics, but have limited application in environmental monitoring and no applications in water quality.

This work defines an approach to identify noise in a time series, explores the use of four different filters, outlines the method to identify the optimum filter for five water quality parameters (water temperature, turbidity, conductivity, dissolved oxygen, dissolved organic matter – fDOM), and investigates issues related to finding a suitable evaluation criteria that allows to account for trade-offs when selecting the optimum filter.

2. METHODS AND DATA
In signal processing, a filter is a process that removes specific unwanted components of a signal; it can be tailored to remove unwanted frequencies or frequency bands. In this study, four low-pass filters have been used to denoise water quality timeseries. The clean signal obtained by filtering the timeseries using Butterworth, Chebyshev I and Chebyshev II filters has been compared with the denoised signal obtained by applying the Savitzky-Golay filter used by Bi et al. (2021). The filters used in this work are linear continuous-time filters. In this category of filters, different types of filters are characterised by the type of pass band (Table 1).

### Table 1 - Type of filters in signal processing.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low pass filter</td>
<td>Passes low frequency band</td>
</tr>
<tr>
<td>High pass filter</td>
<td>Passes high frequency band</td>
</tr>
<tr>
<td>Band pass filter</td>
<td>Passes selected range of frequency band</td>
</tr>
<tr>
<td>Band stop filter</td>
<td>Stop selected range of frequency band</td>
</tr>
</tbody>
</table>

The main steps of this methodology are the following:

- Transform the original timeseries from time domain to frequency domain using FFT.
- Identify the threshold for the noise in the signal’s frequency spectrum.
- Select the filter’s key parameters.
- Denoise the signal using the different filters.
- Convert the denoised signal back into the time-domain, where needed.
- Evaluate and compare the filter performance.

### 2.1 Low pass filters

Low-pass filters attenuate high frequencies and allow frequencies lower than a cut-off frequency to pass. An ideal low pass filter would enable complete transmission in the pass band, complete attenuation in the stop band, and a sharp vertical cut between the two bands so that as little data as possible is ‘lost’. Each family of filters can have different orders. The higher the order, the more sharp is the cut-off, the more the filter will approach the “ideal” filter. In reality, ideal filters do not exist. Accepting that some noise will always remain in the data, the filter should be applied with the aim of reducing the amount of noise in the data as much as possible, trying to minimise the loss in signal.

Four types of low-pass filters are used in this work (Figure 1):

- Butterworth filter – has no gain ripple in the pass band and in the stop band. It is characterised by a slow cut-off.
- Chebyshev filter (Type I) – has no gain ripple in the stop band. It is characterised by a moderate cut-off.
- Chebyshev filter (Type II) – has no gain ripple in the pass band. It is characterised by a moderate cut-off.
- Savitzky–Golay filter – based on least-squares polynomial approximation
The Butterworth filter is a signal processing filter that aims to have a passband with a frequency response as flat as possible. This type of filter is commonly known as a maximally flat magnitude filter. It was first described in 1930 by the British engineer and physicist Stephen Butterworth in a published paper titled "On the Theory of Filter Amplifiers". The Butterworth filter has a slower roll-off rate around the cut-off frequency compared to Chebyshev filters, but it does not produce any ripple. Due to its maximal flat frequency response in the passband, the Butterworth filter is widely used as an anti-aliasing filter in audio processing as a noise reduction tool and in radar to design the radar target track display. (Wu et al., 2010; Liu, Sabrina and Hardson, 2023)

Chebyshev filters have a steeper roll-off than Butterworth filters and have either passband ripple (type I) or stopband ripple (type II). Chebyshev filters have the characteristic of minimising the error between the idealised and the actual filter features over the filter range but with ripples in the passband or stopband. The mathematical characteristics of these types of filters are derived from Chebyshev polynomials. Type I Chebyshev filters are usually referred to as “Chebyshev filters”, while type II filters are usually called “inverse Chebyshev filters”. Chebyshev Type I filters are equi-ripple in the passband and monotonic in the stopband. As such, Type I filters roll off faster than Chebyshev Type II and Butterworth filters, but at the expense of greater passband ripple. Chebyshev Type II filters are monotonic in the passband and equi-ripple in the stopband, making them a good choice for DC and low-frequency measurement applications, such as bridge sensors (e.g. loadcells). (Moore and Jorgenson, 1993; Hongbin et al., 2006; Moschas and Stiros, 2011, 2014; Peng et al., 2018)

Filters created using the Type II method have a slower roll-off compared to those made using the Chebyshev Type I method, but a quicker roll-off compared to those made with the Butterworth method.

In 1964, Savitzky and Golay were interested in smoothing noisy data obtained from chemical spectrum analysers, and they demonstrated that the use of least squares smoothing helps to minimise noise without altering the shape or height of waveform peaks (in their case, Gaussian-shaped spectral peaks). (Savitzky and Golay, 1964)

The Savitzky-Golay (SG) filter is a type of digital filter that can be utilised to smooth out a set of digital data points. The data is smoothed without distorting the signal tendency utilising a technique called convolution. This technique involves fitting consecutive sets of nearby data points with a low-degree polynomial using the linear least squares method. This filter is widely acknowledged as a simple and efficient method for denoising.

To ensure optimal performance of the SG filter, it is crucial to set the window length and polynomial degree appropriately. These settings should match the scale of the signal being processed. This is especially crucial for signals with a high rate of change, as the filter’s effectiveness may be limited if not properly calibrated. (Dombi and Dineva, 2020)

Compared with other smoothing filters, e.g. convolution with Gaussian or multi-pass moving-average filtering, Savitzky–Golay filters have an initially flatter response and sharper cut-off in the frequency domain, especially for high orders of the fit polynomial.

When dealing with data that has a limited signal bandwidth, using Savitzky-Golay filtering can improve the signal-to-noise ratio more effectively than other filters. This means that peak heights of spectra are better preserved compared to similar noise removal filters. There are some drawbacks to using Savitzky-Golay filters. These include the relatively weak suppression of certain high frequencies (poor stopband suppression) and the need for artefacts when using polynomial fits for the first and last points (Schmid et al., 2022).
Thanks to their peak shape preservation property, SV filters are widely used in electrocardiogram processing. When generalised to two dimensions, they can also be used in image processing, such as ultrasound and synthetic aperture radar.

When designing a low pass filter, it is important to identify the key parameters for the filters, the window size for the Savitzky Golay filter and the cut-off frequency for Butterworth, Chebyshev I and Chebyshev II filters. The Savitzky-Golay filter is a window-based filter; therefore, the window size is a key parameter for designing this filter.

For Butterworth, Chebyshev I and Chebyshev II filters, the key parameter is the cut-off frequency. To define the cut-off frequency, the frequency spectrum should be analysed to reveal the characteristics of the signal.

2.2 Defining the cut-off frequency of a signal

A timeseries of data collected by wireless sensors is a discrete signal consisting of a series of values acquired at a constant rate. The number of measurements between any two time periods is finite; the assumption is that the variable measured remains unchanged throughout each non-zero region of time.

When a sequence is sampled at regular intervals, a discrete-time signal is obtained, which is associated with a sampling rate.

Time signals recorded from natural phenomena are corrupted by noise. The measured time signal will always contain some noise due to random influences and inaccuracies, which can never be eliminated completely (Schouten et al., 1994). In water quality timeseries, the inaccuracies may be caused by the measuring device and analogue-to-digital conversion, the biofouling of the sensor, currents present in the water, and other external influences.

To analyse and remove noise from a timeseries, time domain information and the frequency content of the signal should be analysed. The Fast Fourier Transform (FFT) can be used to calculate the signal’s spectrum (frequency content), defined as the signal amplitude vs. its corresponding frequency (Figure 2). FFT is a mathematical algorithm that calculates the Discrete Fourier Transform (DFT) of a discrete signal, which converts the signal into its frequency components.

Figure 2 – FFT transforms the signal from the time domain to the frequency domain.

In the frequency spectrum, the component with the lowest frequency represents the mean value of the signal. Therefore, any noisy signal obtained from a sensor contains a useful low-frequency signal and some noise that occupies the entire frequency range.
Since the useful signal contains low-frequency components, the high-frequency components above a certain frequency, the cut-off frequency, are considered noise. While there is a theoretical threshold where the data can be divided into signal or noise, such a point does not exist in real life, which makes it difficult to select the appropriate cut-off frequency. The selection of a cut-off frequency is difficult because the term itself implies the existence of a specific point where the filter stops undesired frequencies and allows desired ones to pass through. However, this exact point, given that it is only theoretical, does not exist in real life. The most important task when designing a filter is to thoroughly understand the signal that will enter the filter and the signal that should come out of the filter. This understanding is primarily based on the physics of the parameter being filtered. To select a cut-off frequency, it is necessary to have a general idea of the frequencies that should be allowed to pass through and those that the filter should block.

### 2.3 Evaluation criteria

When evaluating the performance of filters in literature, the most commonly used metrics are RMSE and Signal to Noise Ratio (SNR); the latter mostly used when removing noise in images (Zhang et al. (2012), Muppalla et al. (2018), Devi and Patil (2020) and Bi et al. (2021)). These criteria work well where there is a reference signal that is assumed to be the cleaned signal. In the absence of this assumption, the denoised signal is compared to the original signal. In this work, three evaluation criteria, reported below, have been used to assess the performance of the filters.

The **Root Mean Square Error (RMSE)** measures the standard deviation of the differences between the original signal and the denoised signal.

\[
RMSE = \sqrt{\frac{1}{N} \sum_{t=0}^{n} (y_t - \hat{y}_t)^2}
\]  

(1)

The higher the RMSE, the higher is the difference between the original signal and the denoised signal. In addition to this standard evaluation criteria, this work has assessed filters’ effectiveness by examining the average standard deviation of the denoised signal in a selected time interval where it is assumed that the signal remains constant. Based on this assumption, any variation in the signal in this time interval is defined as noise. The **Constant Signal Average Standard Deviation** is defined as

\[
\text{Constant Signal Average Standard Deviation} = \frac{1}{N} \sum_{k=0}^{N-1} |x[k] - \mu|
\]  

(2)

where \(N\) is the number of values in the data set, \(\mu\) is the mean, and \(x[k]\) is the signal represented as a function of the discrete-time variable \(k\). This evaluation criterion cannot be applied to compare the Savitzky–Golay filter with the other low-pass filters. The selected window, where the signal is assumed to be constant, is smaller than the windows equivalent to the selected cut-off frequencies. Furthermore, the filter’s effectiveness is assessed using a basic prediction model to determine the significance of the noise the filter attempts to reduce. A **trend prediction model** is used to predict the signal at different timesteps:

\[
\hat{x}(t + 1) = x(t) - \Delta x(t)
\]

\[
\hat{x}(t + 2) = x(t) - 2\Delta x(t)
\]

\[
\ldots
\]

\[
\hat{x}(t + n) = x(t) - n\Delta x(t)
\]

where \(\Delta x(t) = x(t) - x(t - 1)\)

The RMSE is calculated by comparing the predicted timeseries with the original noisy timeseries at different timesteps and cut-off frequencies for the four filters. After each evaluation, the spectrum of the denoised signal is analysed to visually check that it resembles as much as possible the spectrum of noise. Additionally, the physics of the variable is taken into consideration to
determine what can be defined as an acceptable significant change in the variable to establish the type of filter and cut-off frequency that is most suitable for each variable. (Coraggio et al., 2022)

2.4 Data

Figure 3– Bristol City Council Sampling Sites (Bristol City Council, 2019; Ordnance Survey, 2019) and monitoring stations used in this study

The data used in this study is part of the dataset collected using EXO2 water quality sondes. (Coraggio et al., 2022) These are multiparameter sondes equipped with up to seven sensors capable of measuring a wide range of variables (YSI, 2019), found to be largely successful in the field by other researchers (Snazelle, 2015; Snyder et al., 2018). For this study, turbidity, fDOM (fluorescent dissolved organic matter, a surrogate for dissolved organic content), conductivity, water temperature, and dissolved oxygen (DO) were considered sufficient to capture changes in the area of study, Bristol floating harbour (UK) due to weather, pollutant, and tidal stimuli. The dataset contains data from three monitoring sites (shown in Figure 3) collected between 2018 and 2020, with occasional interruptions due to calibration of the sensors, loss of battery and sensor failures. At each site, data were taken every 5 minutes, with a signal frequency of 0.0033 Hz.

For this work, only data collected between 15/11/2019 and 25/03/2020 have been used. The selected timeframe is the largest section of collected data that had only a one day interruption due to battery loss, which has been filled using past recorded data with similar trend. The timeseries comprehends almost 37500 data points, representing the winter water quality variation in the study area, dataset characteristics are reported in Table 2.

Table 2 – Dataset characteristics

<table>
<thead>
<tr>
<th></th>
<th>Conductivity</th>
<th>fDOM</th>
<th>Dissolved Oxygen</th>
<th>Turbidity</th>
<th>Water Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>n. data points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>476.72</td>
<td>35.30</td>
<td>10.88</td>
<td>11.23</td>
<td>7.89</td>
</tr>
<tr>
<td>std</td>
<td>97.42</td>
<td>5.13</td>
<td>0.48</td>
<td>5.55</td>
<td>0.75</td>
</tr>
<tr>
<td>min</td>
<td>328.50</td>
<td>26.39</td>
<td>9.35</td>
<td>1.46</td>
<td>6.01</td>
</tr>
<tr>
<td>25%</td>
<td>417.20</td>
<td>30.74</td>
<td>10.63</td>
<td>6.64</td>
<td>7.40</td>
</tr>
<tr>
<td>50%</td>
<td>455.70</td>
<td>34.59</td>
<td>10.97</td>
<td>9.99</td>
<td>7.81</td>
</tr>
<tr>
<td>75%</td>
<td>497.70</td>
<td>39.95</td>
<td>11.24</td>
<td>15.36</td>
<td>8.34</td>
</tr>
<tr>
<td>max</td>
<td>834.40</td>
<td>46.99</td>
<td>11.80</td>
<td>32.23</td>
<td>9.89</td>
</tr>
</tbody>
</table>

3. RESULTS
3.1 Selection of the cut-off frequencies

Noise removal techniques are designed to improve the quality of the original signal, attempting to reduce as much noise as possible. Four noise reduction filters, discussed in sections 2.2 and 2.3, have been applied to the original water quality timeseries.

Figure 4 - Frequency spectra of the five water quality parameters. The selected cut-off frequencies are highlighted with differently coloured vertical dashed lines.

Figure 4, shows the frequency spectra of the five water quality parameters obtained by applying the FFT to the detrended original timeseries. The cut-off frequencies tested in this work have been selected based on the peaks in the frequency spectrum for the parameters.

In Figure 5, the frequency spectrum for water temperature is shown. The signal is characterised by an initial peak near the 0 Hz frequency that represents the average of the signal, a second peak near 1.18e-5 Hz that corresponds to the parameter’s daily variations and a third peak near 2.33e-5 Hz that corresponds to the 12 hours variation. These frequencies, representative of the signal characteristics, have been selected as cut-off frequencies in the comparison work, together with the 4.5e-6 Hz, 6 hours variation frequency that does not correspond to a peak in the frequency spectrum, and it is used as the lowest cut-off frequency in the comparison.
The selected four cut-off frequencies have been used in the low-pass filters to remove the noise in the signal. For the SG filter, the filter’s window has been selected as the time equivalent of the cut-off frequencies to represent the characteristics of the water quality signals and allow a fair comparison between filters.

3.2 RMSE evaluation

The filtered signal has been evaluated using RMSE (Table 3) to understand which cut-off frequency and filter are more appropriate for each of the variables analysed. According to Table 3, the lowest RMSE for the five water quality parameters corresponds to the Chebyshev II filter, with a 4.50E-05 Hz ~ 6 hrs cut-off frequency.
### Table 3 - Evaluation of filtered signal using RMSE

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Cut-off frequency [Hz]</th>
<th>Butterworth filter</th>
<th>Chebyshev I filter</th>
<th>Chebyshev II filter</th>
<th>RMSE</th>
<th>Savitzky-Golay filter</th>
<th>Window (timesteps)</th>
<th>Time equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Water Temperature</strong></td>
<td>7.00E-07</td>
<td>0.279407</td>
<td>0.322063</td>
<td>0.219051</td>
<td>0.230181</td>
<td>4609</td>
<td>16 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18E-05</td>
<td>0.048775</td>
<td>0.053378</td>
<td>0.033372</td>
<td>0.044252</td>
<td>289</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33E-05</td>
<td>0.028528</td>
<td>0.039153</td>
<td>0.022309</td>
<td>0.025843</td>
<td>145</td>
<td>12 hrs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-05</td>
<td>0.019250</td>
<td>0.026668</td>
<td><strong>0.013760</strong></td>
<td>0.018460</td>
<td>73</td>
<td>6 hrs</td>
<td></td>
</tr>
<tr>
<td><strong>Dissolved Oxygen</strong></td>
<td>7.00E-07</td>
<td>0.206973</td>
<td>0.239790</td>
<td>0.175031</td>
<td>0.175972</td>
<td>4609</td>
<td>16 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18E-05</td>
<td>0.057511</td>
<td>0.060154</td>
<td>0.047665</td>
<td>0.056907</td>
<td>289</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33E-05</td>
<td>0.051140</td>
<td>0.053382</td>
<td>0.042637</td>
<td>0.050798</td>
<td>145</td>
<td>12 hrs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-05</td>
<td>0.046269</td>
<td>0.048013</td>
<td><strong>0.038335</strong></td>
<td>0.045904</td>
<td>73</td>
<td>6 hrs</td>
<td></td>
</tr>
<tr>
<td><strong>Conductivity</strong></td>
<td>7.00E-07</td>
<td>45.41</td>
<td>55.49</td>
<td>43.76</td>
<td>37.94</td>
<td>4609</td>
<td>16 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18E-05</td>
<td>7.87</td>
<td>10.21</td>
<td>5.84</td>
<td>7.56</td>
<td>289</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33E-05</td>
<td>5.96</td>
<td>7.30</td>
<td>4.46</td>
<td>5.81</td>
<td>145</td>
<td>12 hrs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-05</td>
<td>4.85</td>
<td>5.51</td>
<td><strong>3.60</strong></td>
<td>4.70</td>
<td>73</td>
<td>6 hrs</td>
<td></td>
</tr>
<tr>
<td><strong>fDOM</strong></td>
<td>7.00E-07</td>
<td>2.100732</td>
<td>2.422963</td>
<td>1.774804</td>
<td>1.856595</td>
<td>4609</td>
<td>16 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18E-05</td>
<td>0.248018</td>
<td>0.307092</td>
<td>0.181419</td>
<td>0.239608</td>
<td>289</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33E-05</td>
<td>0.191553</td>
<td>0.217943</td>
<td>0.143840</td>
<td>0.188324</td>
<td>145</td>
<td>12 hrs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-05</td>
<td>0.156181</td>
<td>0.172597</td>
<td><strong>0.118241</strong></td>
<td>0.151355</td>
<td>73</td>
<td>6 hrs</td>
<td></td>
</tr>
<tr>
<td><strong>Turbidity</strong></td>
<td>7.00E-07</td>
<td>2.783055</td>
<td>3.139150</td>
<td>2.174678</td>
<td>2.591965</td>
<td>4609</td>
<td>16 days</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.18E-05</td>
<td>0.907390</td>
<td>0.961985</td>
<td>0.704473</td>
<td>0.908752</td>
<td>289</td>
<td>1 day</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.33E-05</td>
<td>0.776449</td>
<td>0.805890</td>
<td>0.599784</td>
<td>0.768193</td>
<td>145</td>
<td>12 hrs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.50E-05</td>
<td>0.672377</td>
<td>0.704850</td>
<td><strong>0.509165</strong></td>
<td>0.663910</td>
<td>73</td>
<td>6 hrs</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6 – Water temperature signal filtered using Chebyshev II filter using different cut-off frequencies

Figure 6 shows the filtered signal using a Chebyshev II filter with different cut-off frequencies for a one-day segment of the water temperature timeseries. The Chebyshev II filter with a low cut-off frequency (orange line in Figure 7) fails to capture the daily fluctuations in the water quality signal. The cut-off frequency with the lowest RMSE is represented by the purple line in Figure 7. The higher the cut-off frequency, the more the filtered signal resembles the original signal.

Figure 7 shows the filtered signal using a cut-off frequency of 4.05e-5~ 6 hours (that has the lower RMSE) for water temperature on a one-day timeseries window applying four different filters. Data filtered with Chebyshev II resembles the original signal the most at this fixed cut-off frequency. The Chebyshev I filtered signal, instead, underestimates the peaks and overestimates the lowest points of the original signal.

Figure 7 – Water temperature signal filtered using different filters with the same 4.5e-05 Hz~ 6 hours cut-off frequency/window
3.3 Evaluation of the filters on a time interval with constant signal

For this evaluation method, the mean of the standard deviation of the signal is calculated in a section of the time series where the signal is assumed to be constant. This method has been applied to the water temperature time series. Specifically, a time interval between 6 am and 9 am has been selected from the main time series, as the water has released all the heat accumulated during the previous day, and the solar radiation and air temperature are not strong enough yet to reheat the water body.

The mean and standard deviation of the filtered water temperature signal are reported in Table 4. The lowest standard deviation corresponds to the signal filtered using the Chebyshev I filter using a cut-off frequency of 7E-07~16 days. The filtered signal that has an average value most similar to the original’s signal average value is the signal obtained using the Chebyshev II filter with a cut-off frequency of 1.18E-05~1 day.

Figure 8 shows the filtered signal in the selected constant-signal window using different cut-off frequencies and filters.

![Figure 8](image)

**Figure 8 – Selected timeframe with the assumption of having a constant water quality signal. The signal is filtered using three low pass filters with different cut-off frequencies.**

**Table 4 - Mean and standard deviation of the filtered signal in a timeframe where the signal is assumed to be constant.**

<table>
<thead>
<tr>
<th>Cut-off frequency</th>
<th>Butterworth</th>
<th>Chebyshev I</th>
<th>Chebyshev II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td>7E-07</td>
<td>-0.01453</td>
<td>4.65E-12</td>
<td>-0.01453</td>
</tr>
<tr>
<td>1.18E-05</td>
<td>-0.01238</td>
<td>3.45E-05</td>
<td>-0.01352</td>
</tr>
<tr>
<td>2.33E-05</td>
<td>-0.00658</td>
<td>0.000564</td>
<td>-0.00942</td>
</tr>
<tr>
<td>4.5E-05</td>
<td>-0.00045</td>
<td>0.0002056</td>
<td>-0.00211</td>
</tr>
</tbody>
</table>
3.4 Evaluation of the filters using a trend prediction model.

A trend prediction model is used to predict water temperature testing different lead times. With an increase in lead time, the error in the prediction model will increase, leading to high RMSE for prediction based on noisy data.

Figure 9 provides an example of this behaviour. The purple line represents the RMSE of the predicted timeseries based on the original signal. As the lead-time increases, the RMSE shows a rapid increase. On the other hand, the blue line shows the signal filtered with the lowest cut-off frequency, which removes most of the noise (as well as a part of the signal itself). In this case, the RMSE shows a slow change as the lead-time increases.

Figure 9 The evaluation of the Water Temperature predictions based on a trend model. Each subplot shows how the RMSE changes as lead time varies at different cut-off frequencies for the four analysed filters. The purple line represents the RMSE of the prediction model built on the original non-filtered timeseries.

Analysing the behaviour of the four filters at different cut-off frequencies, Figure 10 shows that for the lowest analysed cut-off frequency (7e-07 Hz), the RMSE does not change significantly as the lead time increases. However, as the cut-off frequency increases, the RMSE increases at a faster rate when the lead time is increased. At a fixed cut-off frequency, filters that perform better at short lead times perform worse at longer lead times.

This behaviour can be explained by recognising that the RMSE evaluates the difference between the prediction made on the filtered dataset and the original (noisy signal). Therefore, at short lead times, the more the predicted time series is similar to the original model, the better the RMSE. This is not a desired behaviour for the filter. The filtered time series should not resemble as much as possible the original time series. If this happens, it means that the noise has not been effectively removed.
At the same time, filters that perform worse at short lead times will perform better at longer lead times. These are filters that remove the most noise from the signal; hence they differ more from the original signal at a shorter lead time.

![Graphs showing RMSE changes for different cut-off frequencies](image)

**Figure 10** The evaluation of the Water Temperature predictions based on a trend model. Each subplot shows how the RMSE changes as lead time varies at a fixed cut-off frequency for the four analysed filters.

Figure 11, shows on the top of the plot the frequency spectrum of the signal with vertical lines representing the selected cut-off frequencies and the frequency response of the Butterworth filter for each of the cut-off frequencies. In the bottom plot, the blue line represents the original noisy signal and the other curves are the filtered timeseries colour coded according to the cut-off frequency used for each of them.

From the visual inspection of the filtered signal against the original signal, in Figure 11, it can be noticed that the signal filtered with a cut-off frequency of 4.5E-04 Hz is able better represent the peaks in the original signal if compared to the signal filtered with a cut-off frequency of 2.33E-05 Hz.
### Table 5 - Summary of the best filter according to the three evaluation criteria adopted in this work.

<table>
<thead>
<tr>
<th>Filter</th>
<th>RMSE</th>
<th>Constant signal window – standard deviation</th>
<th>Trend prediction model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butterworth</td>
<td></td>
<td></td>
<td>4.5E-04</td>
</tr>
<tr>
<td>Chebyshev I</td>
<td></td>
<td>7.00E-07</td>
<td></td>
</tr>
<tr>
<td>Chebyshev II</td>
<td>4.50E-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG filter</td>
<td></td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Table 5 summarises the filter and cut-off frequency/window that has been selected as optimum combination according to each evaluation criteria for the water quality signal. RMSE tends to select as optimum, the filter and cut-off frequency that allows the largest amount of noise to pass through. When the filtered data have been evaluated in a window where the signal was assumed to be constant, the optimum combination was the filter and cut-off frequency that removed the largest amount of noise, including part of the signal. With the trend prediction model and a subsequent visual check of the filtered signal, a balance between these behaviours has been found when using the Butterworth filter with cut-off frequency of 4.5E-04 Hz.

### DISCUSSION

When selecting the appropriate cut-off frequency for filtering water quality parameters with low pass filters, it is helpful to examine the peaks in the frequency spectrum of the analysed parameter. These peaks indicate the frequencies that define the signal and should be retained. The peaks in the frequency spectrum are the frequencies that characterise the signal; therefore, they are the frequencies that should be preserved. These frequencies, when transformed back in time intervals, are, according to Coraggio et al. (2022) significant sampling frequencies for water quality parameters and can be used as windows for the SG filter.

When evaluating the signal filtered with different filters and using different cut-off frequency/ windows, the “best fit” obtained by evaluating the filtered signals following the evaluation criteria used in literature, RMSE, find that Chebyshev II is the optimum filter for most of the water quality parameters. When looking at the daily variation of water quality in Figure 7, it is shown that Chebyshev II is the signal that resembles the original signal the most. Therefore, when the filtered signal is evaluated using RMSE, which is a metric that compares the filtered signal with the original, this filter appears to be the optimum solution.
When using RMSE as evaluation metrics, they evaluate how different the filtered signal is from the original, assuming that the original signal is the noise-free signal where instead, in this case, it is the noisy signal. As clearly shown in Figure 6 and Figure 7, using these metrics, we obtain that the more the filtered signal resembles the original signal, the lowest the RMSE. The best signal should not be evaluated on the difference between the filtered signal to the noisy signal because the aim is to reduce the noise as much as possible while preserving the "true" signal.

To address this evaluation issue, the filtered signal has been assessed in a timeframe where the water quality parameter is assumed to be constant. In Figure 8, a section of the water quality signal in a time interval between 6 am and 9 am is displayed. It has been assumed that the water temperature is constant during this interval. This time corresponds to when the water body has lost all the heat during the night, and the sun is not yet strong enough to reheat the water body. Using this evaluation metric, the lowest standard deviation corresponds to the lowest cut-off frequency combined with the Chebyshev I filter. The lowest standard deviation of the filtered signal in this time window corresponds to the Chebyshev I filtered signal with a cut-off frequency of 7E-07 Hz.

Chebyshev I has the smallest standard deviation overall when compared to the other filters applied with different cut-off frequencies, but it is also the filter that is the most far away from the mean of the original signal. Chebyshev II, does not have the smallest standard deviation for the filtered signal but the signal filtered with this filter is the one that has the average value as close as possible to the average value of the original signal.

This result could have been expected because Chebyshev I is the filter that removes as much noise as possible, given that it has no ripples in the stop band. Moreover, Chebyshev II is the filter that works the best at preserving the original signal allowing some noise to pass through. However, once more, these results are different from the behaviour expected by an optimum filter. A filter aims to retain as much signal as possible and minimise the noise.

Furthermore, this evaluation criterion cannot be applied to SG filters because these filters are based on time windows. The selected windows are larger than the three hours window where the signal is assumed to be constant; therefore, it is not possible to evaluate this filter maintaining the same characteristics on this smaller window.

Optimising the noise reduction and the signal loss is a compromise between having the fast cut and stable stop band of the Chebyshev I filter, optimising the noise reduction, and having the fast cut and stable pass band of the Chebyshev II filter, optimising that signal retention. Butterworth filters have the advantage of having no ripples on the pass band and on the stop band but a less sharp cut off. This characteristic makes them the most suitable filters for water quality parameters, where not having a sharper cut-off does not penalise the filtered signal.

This is shown in Figure 9, where the original signal and the filtered signals are used in a prediction model. For each filter, the performance of the prediction model built on the original unfiltered signal quickly decreases with the increase of the lead time. The lower the frequency, the more stable the RMSE is when increasing the lead time. This is because, at lower frequencies, the filter cuts off the noise but also part of the signal. With the increase of the cut-off frequency, the slope of the performance curve starts to increase because the filtered signal begins to contain more noise. When selecting the cut-off frequency, a compromise between these two behaviours should be identified. The compromise can be found when selecting 2.33 e-05 Hz as the cut-off frequency.

In Chebyshev II, the slope of the performance curve is more accentuated than in Chebyshev I. The change in the slope of the performance curve is due to the nature of the filters. Chebyshev II has ripples on the stop band, allowing some noise to pass, and Chebyshev I has ripples on the pass band, allowing a better noise reduction but compromising part of the signal.

Butterworth’s filtered signal is very similar to the signal filtered with Savitzky-Golay in terms of the slope of the performance curve. In Figure 10, it is possible to see that with the increase of the cut-off frequency, the difference between the Butterworth filter and SG filters becomes less marked, and the performance slope increases more rapidly. It is possible to notice that Savitzky-Golay filters perform better at shorter lead times, but Butterworth filters perform better at longer lead times. Filters that perform better are shorter lead times are filters that retain more noise, for
example, Chebyshev II. Therefore, Butterworth can be selected as the preferable filter over Chebyshev I and II filters and SG filters. A visual inspection of the filtered signal, see Figure 11, suggests that the optimum cut-off frequency is 4.5E-04 Hz. Data filtered with this cut-off frequency is able better represent the peaks in the original signal, when compared to the signal filtered with a cut-off frequency of 2.33E-05 Hz. This might be due to the slow roll-off of the Butterworth filter’s curve that, when used with a cut-off frequency of 2.33E-05 Hz, is not able to sharply cut-off the frequencies after the desired frequency.

5 CONCLUSIONS

This study investigated the cut-off frequency and the type of filter that works well for the various water quality parameters with the aim of maintaining the maximum signal/noise ratio. From the results obtained, it is clear that existing evaluation metrics used for assessing the performance of water quality filters, such as RMSE, on their own, are not suited for finding the best compromise between signal loss and noise reduction. With this work, new evaluation criteria and their suitability to find the optimum compromise between different filters for water quality timeseries have been explored. Butterworth and Chebyshev filters have the advantage over SG filters of not having to be linked to a window; therefore, they can be applied to datasets of any length. With the advantage of having no ripples on the stop band and the pass band, the Butterworth filter is the filter that has been proven to be the best compromise between noise reduction and signal retention in water quality timeseries when compared to the filter’s performance of Chebyshev I and Chebyshev II and SG filters. The frequency analysis applied to the time series of water temperature, dissolved oxygen (ODO), and IDOM has shown that the cut-off frequency that produces a filtered signal that adequately characterises the water body should be between of 2.33E-05 Hz~12 hours and 4.5E-04 Hz~ 6 hours. The goodness of a filter is best if accompanied by a visual inspection of the filtered signal, and it will depend on the subsequent steps (i.e. prediction models, peak detection, trend detection etc). For instance, a good filter for peak detection may be one which uses a cut-off frequency of 4.5E-04 Hz~ 6 hours. The same filter may not be the best choice for long-time predictions, where a slightly higher cut-off frequency is preferred.

6 DECLARATIONS

6.1 Ethics approval and consent to participate

Not applicable.

6.2 Availability of data and material

The raw data supporting the conclusions of this article will be made available by the authors upon request, without undue reservation.

6.3 Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

6.4 Fundings

This work is funded as part of the Water Informatics Science and Engineering Centre for Doctoral Training (WISE CDT) under a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/L016214/1. This work was supported in part by the Engineering and Physical Sciences Research Council’s (EPSRC) ‘UK Collaboratorium for Research in Infrastructure & Cities (UKCRIC): Urban Observatories’ grant (ref. EP/P016782/1).

6.5 Authors’ contributions:
Elisa Coraggio carried out data collection, data analysis, and wrote the paper. Dawei Han provided guidance. Claire Gronow and Theo Tryfonas provided advice on the manuscript enhancement. All authors contributed to the article and approved the submitted version.

6 REFERENCES


Bristol City Council (2019) Water quality and pollution, Floating Harbour and rivers.


YSI (2019) *EXO2 Multiparameter Sonde*.
