Project	MONOCLE H2020 (grant 776480)	Start / Duration	1 February 2018/ 48 Months
Dissemination	PUBLIC	Nature	REPORT
Date	25/06/2019	Version	1.0

Multiscale Observation Networks for Optical monitoring of Coastal waters, Lakes and Estuaries

Deliverable 6.1

Report on value-adding approaches and algorithms for single sensor and inter-sensor data processing

Project Description

Funded by EU H2020 MONOCLE creates sustainable *in situ* observation solutions for Earth Observation (EO) of optical water quality in inland and transitional waters. MONOCLE develops essential research and technology to lower the cost of acquisition, maintenance, and regular deployment of in situ sensors related to optical water quality. The MONOCLE sensor system includes handheld devices, smartphone applications, and piloted and autonomous drones, as well as automated observation systems for e.g. buoys and shipborne operation. The sensors are networked to establish interactive links between operational Earth Observation (EO) and essential environmental monitoring in inland and transitional water bodies, which are particularly vulnerable to environmental change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

Project	MONOCLE H2020 (grant 776480)	Start / Duration	1 February 2018/ 48 Months
Dissemination	PUBLIC	Nature	REPORT
Date	25/06/2019	Version	1.0

Deliverable Contributors:	Name	Organisation	Role / Title
Deliverable Lead	Thomas Jackson	PML	WP6 Lead
Contributing Author(s)	Liesbeth De Keukelaere	VITO	Contributor
	Stefan Simis	PML	Contributor
	Steef Peters	Water Insight	Contributor
	Olivier Burggraaff	U Leiden	Contributor
	Norbert Schmidt	DDQ	Contributor
	Frans Snik	U Leiden	Contributor
	John Wood	Peak Design	Contributor
	Jaume Piera	CSIC	Contributor
	Marc Ventura Oller	CSIC	Contributor
	Robrecht Moelans	VITO	Contributor
Reviewer(s)	Claire Neil	U Stirling	
Final review and approval	Stefan Simis	PML	Scientific Coordinator

Document History:

Release	Date	Reason for Change	Status	Distribution
0.1		Initial outline	Draft	Internal
0.2-0.9		Drafts and Review	Draft	Internal
1.0		First release	Final	Public

To cite this document:

Jackson, T., et al. (2019). D6.1 Report on value-adding algorithms and approaches for single sensor and inter-sensor data processing. Deliverable report of project H2020 MONOCLE (grant 776480). doi: 10.5281/zenodo.3258795

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

1. Executive Summary

This document highlights a number of techniques and approaches that are being tested in MONOCLE in order to increase the quality and utility of the data collected by the suite of sensors developed and used in the project. Many of the techniques discussed here could be used for other sensors in the field of water quality monitoring.

The approaches can be split into two categories: 1) techniques/approaches that will optimise data processing procedures and 2) those that will allow increased information retrieval from a given dataset or combination of datasets. Following a discussion of both of these areas, we formulate intended tests and demonstrations to be performed within MONOCLE alongside a set of data requirements.

2. Scope

The aim of this document is to summarize potential value-adding approaches and algorithms for single sensor and inter-sensor data processing within the context of the MONOCLE project and sensor networks in general. This document serves as an introduction and guide to methodologies that can be used with sensor networks (such as those in MONOCLE) to optimise efficiency of data collection and improve the utility and value of the data. The document marks the start of integrative data analysis in the MONOCLE project and is provided primarily as a reference for project partners to implement these techniques, as well as a public document for those interested to follow project progress.

3. Introduction

Our capacity to generate data has increased exponentially in recent times as the connectivity of data networks has increased, the cost and size of computing systems has decreased, and as the creation/uptake of data standards has allowed an increase in software and hardware compatibility. The volume and diversity of data that can now be collected present both new opportunities and challenges to the science community. We can now create smarter networks of sensors with increased autonomy, greater levels of cross-communication and higher levels of data resolution and accuracy. Increased network connectivity and diversity increases the number of potential data formats, number of communication channels and complexity of data quality control. However, the adoption of modern data handling techniques is required to maximise the utility of the data provided, avoid excessive data transfer and avoid the contamination of high-quality datasets with spurious results.

In this document we discuss a number of approaches that can be used alongside sensors and sensor network usage in order to tackle data handling challenges and maximize their synergistic use. We discuss these approaches in the context of aquatic optics and biogeochemical parameters as collected by the MONOCLE system, though many of the principles and algorithms are applicable across all fields of data driven science.

Section 4 of this document deals with techniques for optimising data collection and transmission in order to reduce data volumes and improve the quality and utility of the data provided, without losing essential information. Section 5 covers techniques for the enhancement of datasets through automated processing and multi-variate analysis for the purposes of enhanced information retrieval and data quality control. Individual sensor capabilities are covered in more detail in their respective user guides/documentation, while sensor connectivity and system structure are covered in D5.2 'System architecture and standards report'. Results presented here are for illustrative purpose only and may not reflect the final performance of the MONOCLE sensor system.

4. Optimised data processing and transfer

Over the last two decades, the characteristics used to describe big data have increased from the "Three V's" of Delany (2001) to six (Ur Rehman et al 2013) or even ten V's (Khan et al 2018). Of these ten V's (Volume, Velocity, Variety, Value, Variability, Veracity, Validity, Volatility, Viability, and Viscosity) the primary points of interest and for this document are Volume, Value, Veracity and Validity.

Some of the first scientists to encounter issues of data volumes too large for transmission were astronomers. An illustrative example is the Kepler satellite, capable of recording a 95 megapixel image every six seconds but possessing only a 550KBps download speed (Kock 2010). Here, the solution consisted of onboard data reduction by reducing the bandwidth by a factor of 300 through temporal averaging, selection of interesting targets and onboard processing of only the relevant pixels before transmission.

In the satellite calibration/validation landscape which MONOCLE developments apply to, similar challenges present themselves. In terms of a reduction of operational cost, buffered data storage and transmission using long-range radio protocols may benefit from data volume reduction and pre-processing. At the same time, essential information on the variability of the signal should be retained from any given measurement to preserve quality of information.

A number of approaches will be tested within MONOCLE to reduce the data volumes that are transferred from sensor networks to data stores and on to users. These approaches are described in the sections below.

Local (sensor side) vs remote processing and hardware optimisation

The first potential benefit of sensor side processing to consider is the ability to reduce the volume of data to be transmitted.

(Spectro)radiometers

Within MONOCLE, Water Insight and Plymouth Marine Laboratory develop methods to extract representative reflectance spectra from repeated water-leaving reflectance measurements taken with MONOCLE radiometry sensor suite (e.g. Simis and Olsson 2013, Groetsch et al. 2016). This can be seen as a two-step procedure in which anomalous/corrupted measurements are first filtered from a dataset before further (optional) automated selection of, or statistical generation of, a representative spectrum from multiple measurements. Simple averaging could be used to create representative spectra, but other methods, such as spectral weighting while accounting for measurement uncertainty, may provide a more robust result (increasing data veracity).

Additionally, hardware configurations can aid data reduction for some sensors. The utilisation of a solar tracking radiometry platform for commercially available spectroradiometers (SO-RAD by PML) will reduce the transmission of unusable or contaminated data. This system uses algorithms to:

- predict whether solar elevation is above a suitable measurement threshold (nominally 30°) from GPS input (time, position). This uses the PyEphem library for python.
- predict whether suitable azimuth viewing angles can be obtained within sensor rotation limits, from any GPS input (time, position) and dual-GPS position logging yielding the ship bearing. Bearing is obtained using Vincenty's formulae and has the form: $\theta = \arctan(\sin(\Delta lon).\cos(lat2) / \cos(lat1).\sin(lat2) - \sin(lat1).\cos(lat2).\cos(\Delta lon))$

WISP-M

The Water Insight WISP-M system currently processes data from Level 0 to Level 2, reducing data volume requirements for users. This processing involves changing the wavelength grid and reducing the ensemble of measurements. Much of this processing is currently performed ‘server-side’ but a development version of the WISP-M sensor-side processing is under consideration to provide users with a direct estimation of water quality parameters in situations where the data transfer speed is not optimal. The WISP-M processing steps (and data file sizes) are given below.

A WISP-M measurement (as currently performed on the WISPstation instrument) contains 10 spectral measurements from each of the 8 channels. The internal spectrometer operates 2048 pixels covering a wavelength range of 200-1100 nm. So the theoretical sampling interval is about 0.44 nm.

The instrument is configured to operate as follows:

1) A channel is opened and the optimal integration time is determined. Then (very rapidly) 10 consecutive measurements are performed while the integration time remains fixed.
2) This process is repeated for each of the 8 channels.
3) The data is packed into a serialised JSON L0 file and send to the WISPcloud server

L0 data

The WISPcloud server receives the data as textfile (serialised json) and converts the measurements to counts (16 bits unsigned integers) first.

Estimated volume of the json file:

Data	=	8 channels * 2048 numbers * 12 bytes * 10 repetitions per channel
Darkcurrent data	=	8 channels * 18 numbers * 12 bytes * 10 repetitions per channel
(including metadata)	=	about **2 MB** per datafile

Estimated volume of the data as unsigned integers:

Data	=	8 channels * 2048 * 2 bytes * 10 repetitions
Darkcurrent data	=	8 channels * 18 numbers * 2 bytes * 10 repetitions per channel
(including metadata)	=	about **330 KB** per datafile

Next the data is joined with the wavelength table on the 200-1100 nm grid.

L0A processing

At this stage the data is at L0A and some quality checks are applied (the checks are only applied to the final working range of 350-900 nm):

-Error: The level0A spectrum is NOT a flat line
-Error: The level0A spectrum does NOT contain negative values
-Error: The level0A spectrum is NOT saturated
-Warning: The level0A spectrum contains zero values
-Warning: The Level0A spectrum is NOT underexposed

Next the data is corrected for dark current and the non-linearity correction is applied. At this stage the data is at L0B and the datatype changes from 16 bits unsigned integer to 32 bits float.

L0B processing

At L0B the following quality checks are first applied (for the working range of 350-900 nm):

-Error: The L0B spectrum is NOT a flat line
-Error: The L0B spectrum does NOT have extreme negative values
-Warning: The L0B spectrum contains negative values
-Warning: The L0B spectrum contains zero values
-Warning: The dark current is high

The data is interpolated to a regular 1 nm grid on a 200-1100 nm range, bringing the data to L1A.

L1A processing

At L1A first some standard quality checks are applied again:

-Error: The L1A spectrum is NOT a flat line
-Error: The L1A spectrum does NOT have extreme negative values
-Warning: The L1A spectrum contains negative values
-Warning: The L1A spectrum contains zero values

The L1A spectrum is corrected for the integration time and the radiometric correction is applied.

Now each spectrum is reduced (interpolated) to the working grid of 350-900 nm with a fixed interval of 1 nm. This is achieved by applying a moving average of 9 pixels including corrections for edge pixels. This brings the spectrum to L1B.

Note 1: L1B spectra are the lowest level spectra that can be extracted from WISPcloud using the API.
Note 2: Since the instrument only has one spectrometer, all the channels have the same radiometric correction

L1B processing

Besides the before mentioned quality checks, at this level it is also checked if the integration time is within reasonable boundaries, otherwise a warning is given.

Standard Level1B processing starts by averaging all L1A spectra from 1 channel within the set of 10 repetitions that have not been flagged with an error.

This reduces the data to 2 downwelling irradiance (Ed), 2 downwelling radiance (Ld) and 2 upwelling radiance (Lu) spectra per observation at 1 nm resolution over a range of 350-900 nm as 32 bits float with a volume of **13.2 Kb.**

L2R processing

The objective at L2 is to calculate the best above water remote sensing reflectance spectrum, so a selection of channels is made. Between the two Ed channels, the one with the highest average value is selected. Between the two viewing directions of the instrument, the set of channels with an azimuthal difference (sun azimuth – sensor azimuth) closest to the optimal 138 degrees is chosen (Mobley 2015).

For standard processing a fixed rho is used with a value of 0.028.

$$Rrs = \frac{L_u - 0.028 * L_d}{E_d}$$

The resulting Rrs has a volume of **2.2 Kb**. These data are now three orders of magnitude smaller than the initially transferred json file and are within a data volume range that is acceptable for many end users.

Experimental L1B to L2R processing in the context of MONOCLE

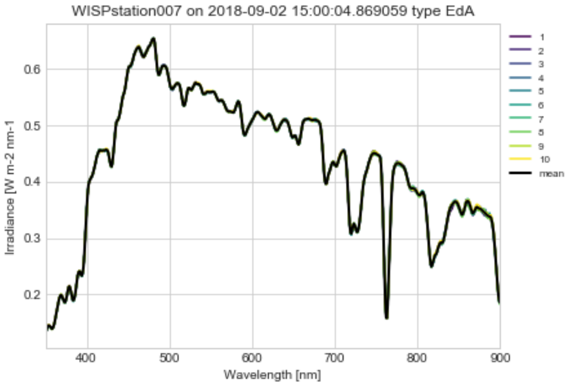
Currently the following additional processing options are being investigated:

1) Changing the fixed rho value to a dynamic value based on the viewing geometry (to conform with Mobley 2015).
2) Correcting for sun glint using the similarity spectrum (Ruddick, 2006)
3) Correcting for sun glint using the Groetsch et al. approach (2016)
4) Correcting for sun glint using the Simis and Olsson approach (2013).

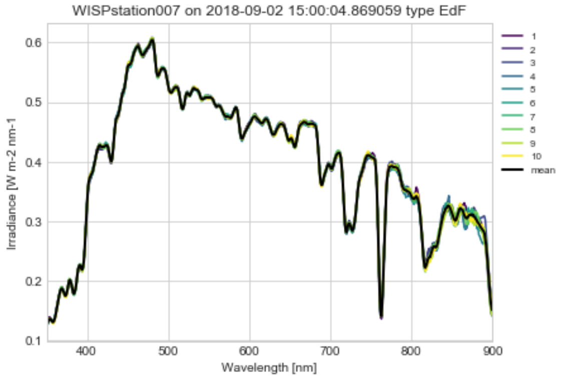
We are also looking into eliminating spectra from the set-of-10 by ignoring spectra that are further away from the mean than e.g. 3 X the standard deviation.

Methods suggested by Brando et al, (2016) to select a percentile of the 10 repetitive measurements will also be investigated further, although a first glance at the data suggests that this will not have a big impact on most measurements. Only during low light conditions with long integration times, the Lu spectra may be strongly variable because of wind, waves and sunglint variations. For an example of the sets of 10 measurements and resulting R_{rs} see

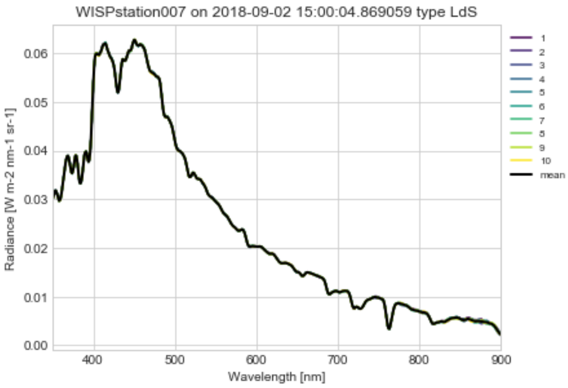
Figure 1.



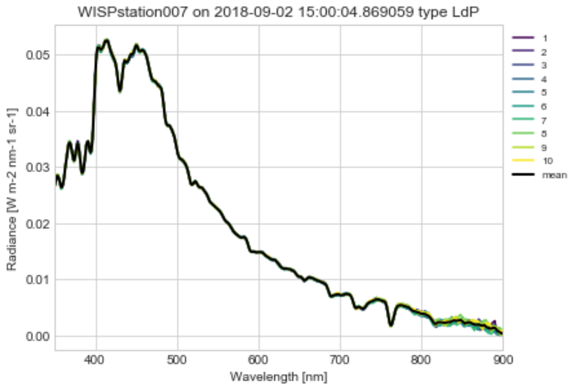
Set of 10 Ed measurements at the Aft sensor



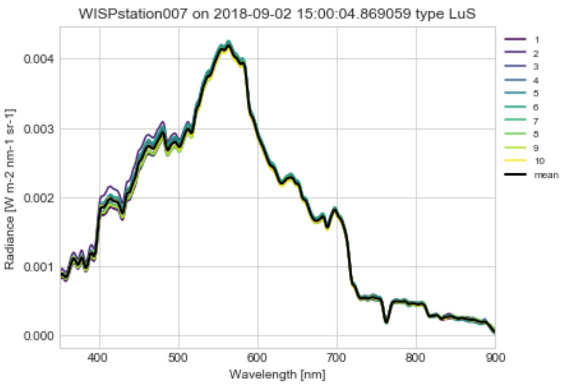
Set of 10 Ed measurements at the Fore sensor



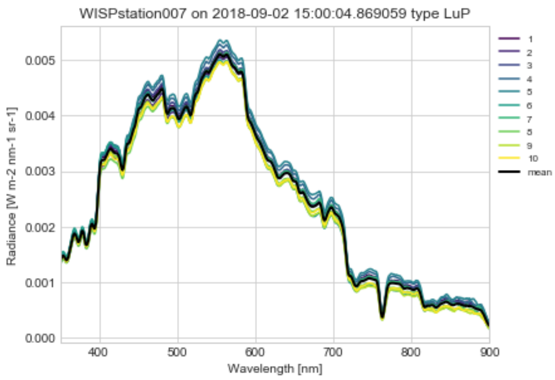
Set of 10 Ld measurements at the Starboard sensor



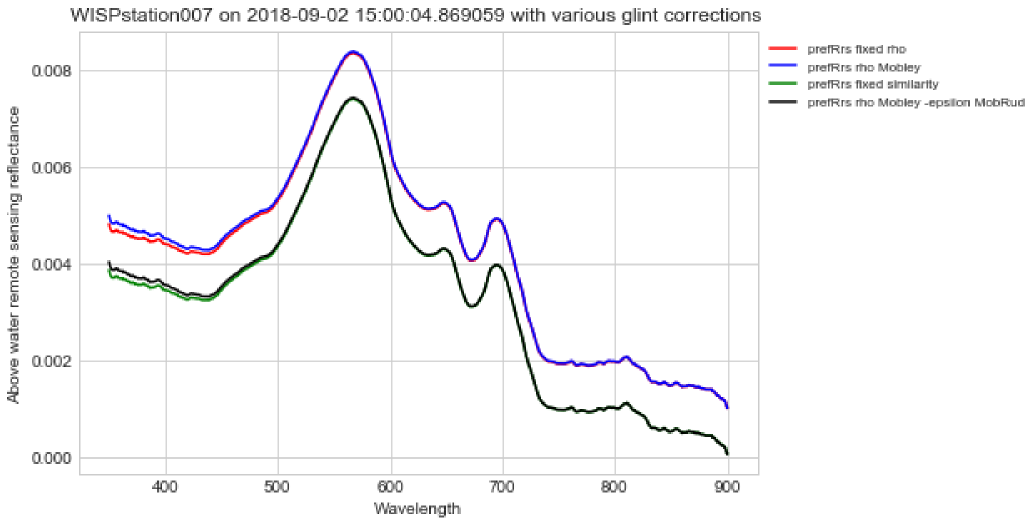
Set of 10 Ld measurements at the Portside sensor



Set of 10 Lu measurements at the Starboard sensor



Set of 10 Lu measurements at the Portside sensor



Above water Remote sensing reflectance based on 4 different calculations:
Standard fixed rho = 0.028 (red)
Rho according to the Mobley 2015 LUT (blue)
NIR adjusted spectrum (fixed rho) using the Ruddick similarity approach (green)
NIR adjusted spectrum (Mobley rho) using the Ruddick similarity approach (black)

Figure 1: Examples of repeated WISP spectral measurements and calculated R_{rs} Spectra.

Drone imaging

Within MONOCLE, Drone-acquired image data are processed by VITO from raw radiance information to reflectance and derived products. Most of this processing is performed server-side (between the sensor acquisition and provision of products to users) but image acquisition methods can be used to reduce the amount of corrupted or contaminated data that are initially acquired.

The image acquisition protocols, to avoid corrupted or contaminated data, are defined in D3.1 'Operational Protocols For Acquisition And Deployment V1'. There are three main considerations in the data acquisition phase:

- Avoid sun glint, i.e. the reflection of direct sun light into the field of view of the sensor, as much as possible. Tilt the camera slightly (15°) during flight operations and look away from the sun.
- Data collected by the GPS/IMU system onboard of the drone is important to perform a geometric correction. Since water is a dynamic medium subjected to influence from waves, tides, floating and settling particles and more, the typically used georeferencing technique (i.e. structure for motion, Westoby et al. 2012) applied over land are not valid here. To know which part of the water surface your drone image captures, you fully rely on the Global Positioning System (GPS) and Inertial Measurement Unit (IMU) , which collect information on the latitude, longitude and flying height of the drone, as well as the roll, pitch and yaw of the camera.
- Finally, to correct for the incident radiance, either an irradiance sensor or spectral reference panels, with known albedo, are required.

The raw data collected by airborne drone imagery contains raw information expressed in Digital Numbers (DN) and is subjected to distortions from the camera as well as the atmosphere in between the target and the sensor. Figure 2 shows an example acquired at Loch Leven (Scotland, UK), an enhanced uncorrected true colour image which suffers from vignetting effects (i.e. darkening towards the edges of the image), sun glint effects at the bottom of the image and cloud shadow in the middle of the image.

Server-side processing for the drone imagery to convert raw airborne drone data into meaningful bio-physical data consists of three main steps. These are radiometric correction, georeferencing and turbidity-algorithm implementation (Raymaekers et al., 2017), as shown in shown in the schematic

Figure 3. The different steps are discussed more in detail in the next paragraphs.

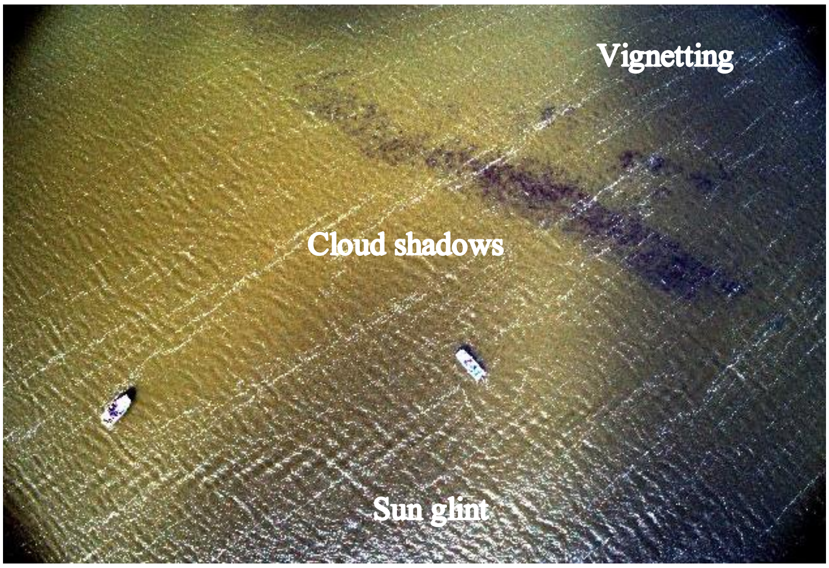


Figure 2: Example of an uncorrected true color image captured with the MicaSense RedEdge camera. The image shows vignetting effects towards the edges, sun glint effect at the bottom and cloud shadows in the middle of the picture

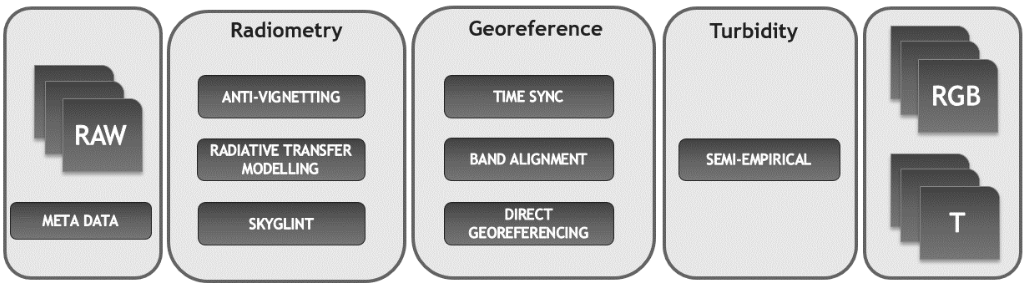


Figure 3. Schematic overview of the drone image processing chain to convert raw data into meaningful bio-physical units and true-color products

The radiometric correction step converts raw drone imagery from digital numbers to water leaving reflectance. The latter quantity is of interest because the light has travelled through the water column and thus bears information regarding (optical) characteristics like turbidity or chlorophyll-a concentration. The first step performs an anti-vignetting of the image. Vignetting is the darkening of the image towards the edges and can be corrected for by normalising with a calibrated reference image. Secondly, the radiance signal received by the sensor can be converted into physically meaningful water-leaving reflectance through radiative transfer modelling. The at-sensor radiance ($\mathrm{L_{at-sens}}$) is the sum of the atmospheric radiance ($\mathrm{L_{atm}}$), the specular reflection at the water surface ($\mathrm{L_{spec}}$) and the water-leaving radiance ($\mathrm{L_w}$):

$$\mathrm{L_{at-sens}} = \mathrm{L_{atm}} + \mathrm{L_{spec}} + \mathrm{L_w}$$

The specular reflection consists of two components: direct reflection of sun light, also called sun glint ($\mathrm{L_{r,sun}}$), and scattering of the atmosphere to the water surface and reflected into the detector, i.e. sky glint ($\mathrm{L_{r,sky}}$):

$$\mathrm{L_{spec}} = \mathrm{L_{r,sun}} + \mathrm{L_{r,sky}}$$

When processing drone images, two assumptions are made:

1. Drones fly at limited height (esp. compared to satellites), so $\mathrm{L_{atm}}$ can be neglected
2. The camera of the drone is slightly tilted to avoid sun glint and thus the $L_{r,sun}$ component can be ignored. This is however a simplification of reality, since (i) the pixels of a frame camera have different viewing angles and (ii) waves at the water surface can lead to occurrence of sun glint within the image.

The simplified radiative transfer formula is:

$$\mathrm{L_{at-sens}} = \mathrm{L_{r,sky}} + \mathrm{L_w}$$

The sky glint contribution is modelled with the iCOR image processing tool (De Keukelaere et al., 2018) adapted for drone imagery. iCOR is an image-based atmospheric correction tool which relies on Moderate-Resolution Atmospheric Radiance and Transmittance Model – version 5 (MODTRAN5) (Berk et al., 2006) Look-Up-Tables (LUTs) to solve the radiative transfer equation based on a set of input parameters. The input parameters are: height, solar and viewing angles and simulated cloud type and coverage (open sky, cumulus, stratus, etc.).

The quality of interest is water leaving reflectance ($\mathrm{\rho_w}$), which is an optical property of water and can be related to bio-physical parameters. $\mathrm{\rho_w}$ is expressed as:

$$\mathrm{\rho_w} = \frac{\mathrm{L_w}}{\mathrm{E_d}}\,\pi$$

with $\mathrm{E_d}$ the downwelling irradiance. The value for downwelling irradiance can be obtained from either spectral reference targets present in the field or an irradiance sensor mounted on the drone. An irradiance sensor allows continuous monitoring of changing light conditions, but is very sensitive to viewing angle and is not straightforward to process due to difficulties in the separation of direct and diffuse sun light on the sensor. This separation is not measured in situ but calculated in post-processing. Due to strong dependence of this separation on cloud cover, sensor orientation and time of the day a post-processing approach is not ideal, but is currently the nest option available. Future advances, such as reduction in the weight and capability of irradiance sensors, are expected in the years to come and these would allow the separation of direct and diffuse irradiance to be measured in-situ. Another solution to measure irradiance is the use of spectral reference panels, which have a known and calibrated reflectance value (albedo) over the complete panel and can be fixed on a boat or at the shore-side. The drone has to fly over these panels, and only the light conditions at the moment of the overpass are captured. When the measured radiance from the camera can be coupled with the known reflectance value, other measured radiance values can be translated into reflectance through interpolation. This method has satisfactory results in conditions with uniform cloud but it is less appropriate under strong variations in cloud cover.

Figure 4: Two types of spectral reference panels placed at the shore-side (left) and fixed on a boat (right), with known spectral behaviour. The light target has an albedo of 36%, while the albedo of the darker target is 12%

Images are subsequently georeferenced (a geometric correction is applied). Time-synchronized triggering between the camera and auxiliary sensors is of utmost importance, since the drone location or attitude may alter within a fraction of a second. Each spectral band of the camera is aligned before the images are projected based on position, altitude and orientation of the drone and camera, which are recorded by the GPS/IMU. In contrast to land application, no fixed recognizable features are present in water bodies, precluding use of structure-for motion techniques (Westoby et al. 2012). The precision of the drone auxiliary sensors determines the geometric accuracy of the final product which is projected through the so-called direct georeferencing technique. A GPS system provides information on latitude, longitude and height of the camera, while an IMU sensor captures the roll, pitch and yaw of the camera. Application of translation, rotation, projection on a flat surface (water) and image warping, results in an image that is georeferenced in space.

Figure 5 summarises the geometric correction step.

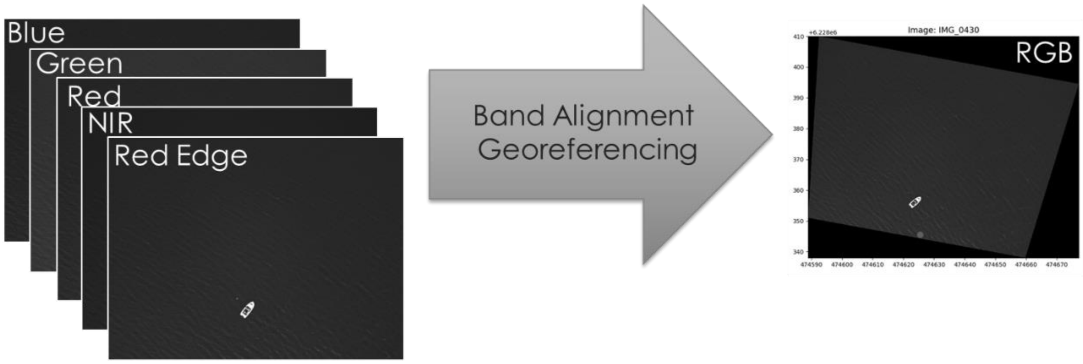


Figure 5: Schematic overview of the geometric correction

From water leaving reflectance we can create a suite of derived products, including a true-colour mosaic, turbidity, total suspended matter or chlorophyll maps. Images are captured in high resolution, centimetre spatial scales, however for most applications such high granular resolution over water bodies is not necessary. Therefore, the data can be reduced by down sampling the images to a coarser spatial resolution. Currently this down sampling is performed in the final stage of the server-side processing, before upload to the MONOCLE back-end.

iSPEX

The iSPEX system developed by the University of Leiden utilises the local processing capability of mobile devices (smartphones) in order to provide sensor calibration and process RAW images (very large files) into products for transmission. The local calibration process includes dark current and read noise, flat-fielding, and spectral calibration. RAW images can be up to 32 MB each, depending on camera resolution and bit depth. While the full measurement protocol for iSPEX has not been developed yet, it will likely require the user to take multiple images for calibration purposes and to reduce noise, for example due to waves on the water. Multiple exposures would then be averaged and calibration images subtracted on-board the smartphone to prevent transferring gigabytes of files. The observed spectrum would be cut out of the image and reduced in dimension by averaging over multiple pixel rows to further reduce the data size. The goal is to have data files smaller than 1 MB per iSPEX observation.

KdUSTICK

The latest version of KdUINO is referred to as KdUSTICK and applies the same concept as the original version, but with measurements in 3 different bands (with 3 colour light-frequency converters) as shown in Figure 6.

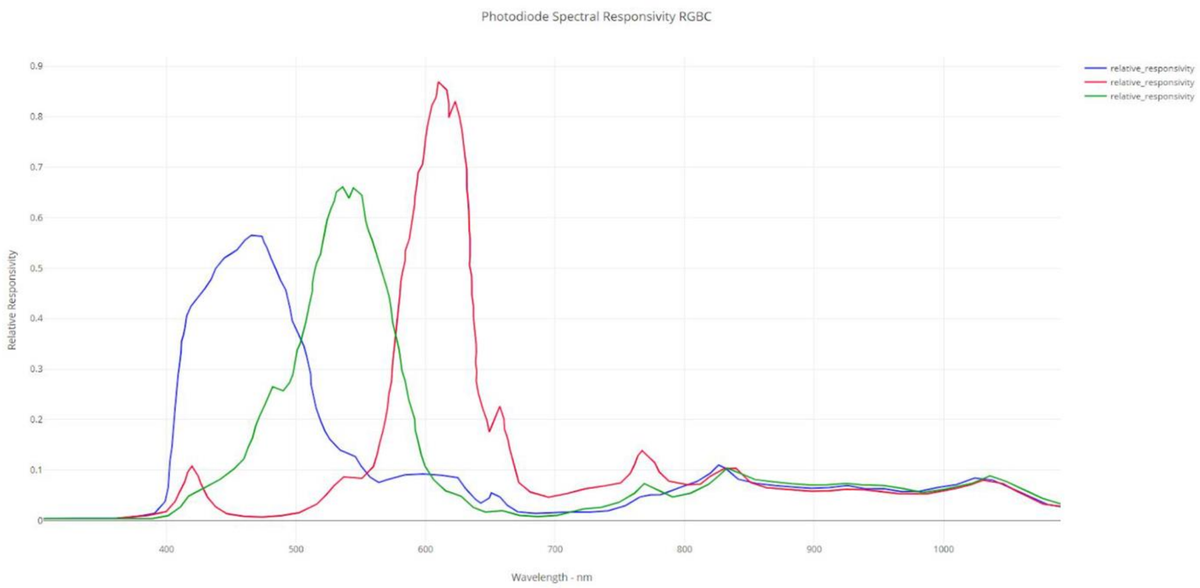


Figure 6 Estimated wavelength sensitivity for the 3 sensors used in KdUSTICK

The KdUSTICK can record time series of downwelling light attenuation ($K_{d)}$ measurements and save it in local storage or send it in real time using Internet of Things technologies (Sigfox currently being tested). As with the reflectance sensor processing, it is possible for the sensor system to derive a statistical representation of the timeseries and transmit that at a reduced data volume, rather than the complete time series. An example of such statistics are the mean value and the variance obtained from a K_d histogram over a representative time interval. The complete time series could be provided following a request to the sensor if the bandwidth and a scientific justification is available.

Additionally, the KdUSTICK obtains irradiance measures (Ed) from the red, green, blue and clear light sensing elements. This means that multiple K_d products (K_d(red), K_d(green), K_d(blue) and K_d(clear)) can be derived as well as the coefficient of determination corresponding to each of them. Unless explicitly requested, the intention is to transmit only K_d(clear) and corresponding coefficient of determination in near real-time.

Variable sampling rates and signal to noise ratio (SNR) optimisation

By building on the Sensor Planning Service of the Open Geospatial Consortium, there is the possibility of sending requests to sensors for a change in data provision. Thus, it is also possible that requests could be sent through the sensor network for changes in the sampling/sensing protocol in response to user requirements. Alternatively, the sampling strategy could include automated adaptation to local conditions or satellite coincidence. Instruments within the MONOCLE system that are candidates for this approach include Kduino/ KdUSTICK, WISP-M, SO-RAD and CLAM devices.

Increasing integration times for sensing can increase the SNR but it is not the only option. Due to the nature of the sensors contained in the MONOCLE system, we can consider both hardware and software (processing) approaches to optimise SNR.

KdUSTICK

The KdUSTICK can perform sampling at a variety of rates using both different integration times in the colour sensor (per measurement) and also specifying how long we want to take measurements for (total time series lengths). This means that sampling rates could be increased at times of increased variability in the local environment (such as rainfall events). As described in Darecki *et al.* (2011), it is possible to resolve high-amplitude, short-duration light flashes caused by wave focusing at near-surface depths in sunny conditions (Figure 7). This requires sampling rates of up to 1 kHz, much higher than those required for the routine estimation of K_d. Therefore, it is sensible to have short bursts of high frequency measurement in order to check for wave focusing alongside longer measurement periods designed to maximise the SNR for the K_d estimates.

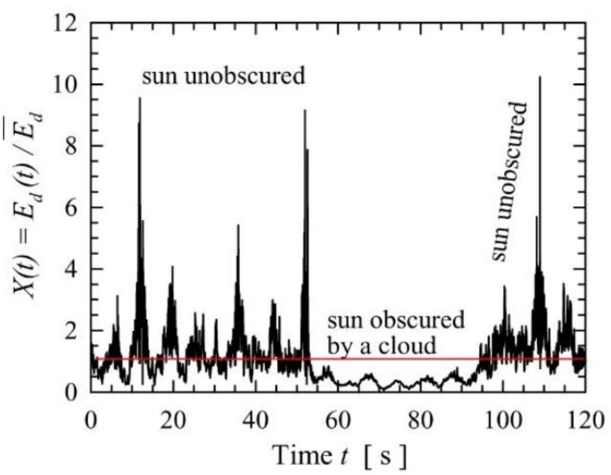


Figure 7 Example from Darecki *et al.* (2011) of near-surface light fluctuations. Measurements in the open ocean waters (south of the Hawaiian Islands) September 3, 2009 at 10:20 AM local time illustrating variable sky conditions with intermittent cloud cover.

The light sensors currently used within the KdUSTICK, which use pulse frequencies (i.e. light-to-frequency converter), have approximately 20% deviation error in their measurement. This error is only significant for calculation of K_d in very transparent waters ($K_d < 1$ m^{-1}) but in these cases it might be required to increase the SNR, if possible. Also, long time integrations in light sensors are important to mitigate light variability in the water column caused by the surface wave effect and to correctly estimate the water transparency. We can minimize wave-induced light fluctuations by averaging pulses per unit time, see Figure 8. The optimal time for integration of the light sensor measurement has not yet been determined but the current settings control the sensor integration time to cycles of one minute.

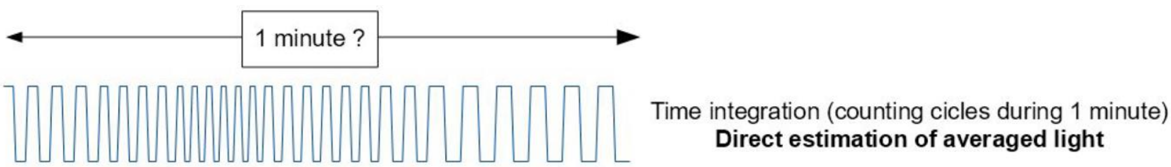


Figure 8 Schematic diagram of how we obtain averaged pulses.

WISP

The WISP device firmware features the ability to vary sampling rates. This is currently accessible through direct interaction with the sensor or by remote access. The maximum frequency of sampling is bounded by the duration of the measurement itself and the time it takes to send the data.

Currently a safe margin is implemented to allow the instrument to switch from 1 channel to the next, combined with a time margin to allow the sending of the data to the database over 3g/4g. Variable properties are the duration of the measurement itself (depending on the ambient light conditions) and the speed of data transfer. Practical experience indicates that a safe interval between two consecutive measurements is between 2 and 3 minutes. This is a short enough window that for many applications, such as matching measurements to satellite overpass times, it should be possible to take a measurement with minimal time offset.

In the design of a spectrometer system, there is always a trade-off between the speed of measurement and the quality (in terms of signal to noise ratio) are interlinked. For the WISP-M we have chosen a spectrometer configuration with an enhanced sensitivity (Signal/Noise = 300:1) and a slit of 100 um. The system is using smart software (programmed in C) to quickly optimise the integration time to obtain a signal level that is just below saturation level.

Radiometers

The SO-RAD platform will adopt the Sensor Planning Service to ensure maximum data rates during satellite overpasses or periods of special interest (e.g. nearby citizen science measurements). This is not in the scope of data or signal optimization but part of synergy between in situ and satellite observations demonstrated in WP7. Depending on the sensor system mounted on the SO-RAD, optimization of the SNR is possible through optimized integration times. In current tests, TriOS RAMSES sensors are used with automated integration time. Data which use < 10 % (under-saturated) or the full (over-saturated) dynamic range are discarded. The TriOS sensors do not support high-frequency (multiple sub-second exposures) due to limitations in data transfer rate.

CLAM

The primary use of variable sampling rates for the CLAM sensor would be to increase the signal to noise ratio by integrating measurements over a longer sampling interval. There are hardware configuration changes that will optimise the SNR between clear and turbid waters (primarily changes to the pathlength of the sensor).

Drone imaging

Drone based sensing provides the capability to perform mapping flights as well as hover in a (nearly) fixed position. When flying in mapping mode, the drone covers a larger spatial area but each image captures a different scene. You can define a percentage of overlap for the two directions (x- and y-direction). For land applications, typically an overlap > 80% is chosen to allow stereoscopy. Since water can, in general, be considered as a flat surface, this percentage can be lowered without losing information. Setting a region of overlap for water monitoring is nevertheless useful, since it allows the exclusion of bad pixels, which are for example affected by sun glint. Out of these images one mosaic can be generated.

Performing a full flight mission in hovering mode is mainly of interest in highly dynamic environments, e.g. when a sediment plume is entering a harbour. But hovering also allows the aggregation of consecutive images which increases the SNR and the variability of pixels of short timescales may also provide information on the data uncertainty. When spectral reference targets are used to account for the irradiance (see sensor-side processing), it is recommended to hover for several minutes over these reference panels to increase the SNR.

Dimensionality reduction

Dimensionality reduction is the process of reducing the number of variables to a set of 'principle' variables. This aims to capture a high proportion of the variance in the dataset in a reduced dimensional space. Consider a set of hyperspectral reflectance measurements, for many applications the full n-dimensional space (where n is the number of wavelengths measured) is not required. Additionally, there is likely to be strong covariance of reflectance values in close wavelength proximity as the absorption peaks of common absorbing components (such as photosynthetic or photoprotective pigments) can be anywhere from 10-100 nm across (Bricaud, 2004). This is one of the reasons that it is possible to estimate chlorophyll-a concentrations from remote sensing data using information from just 2-5 wavebands.

Principle component analysis (PCA) is the process of identifying the spatial axes that contain the majority of the multidimensional variance. Some data loss is unavoidable but ideally this is minimal if sufficient variance is captured. In addition, computing and other power requirements may limit the implementation of sensor-side dimensionality reduction.

PCA might be possible for the KdUSTICK data as measurements are performed for at least 3 wavebands. An example of using PCA to classify water masses is shown in Figure 9. In the example given, 3 waterbodies can be distinguished when compared on the 2 principle component axes. This means that only those two PCA axes values would be required to assign a measurement to a waterbody or water type.

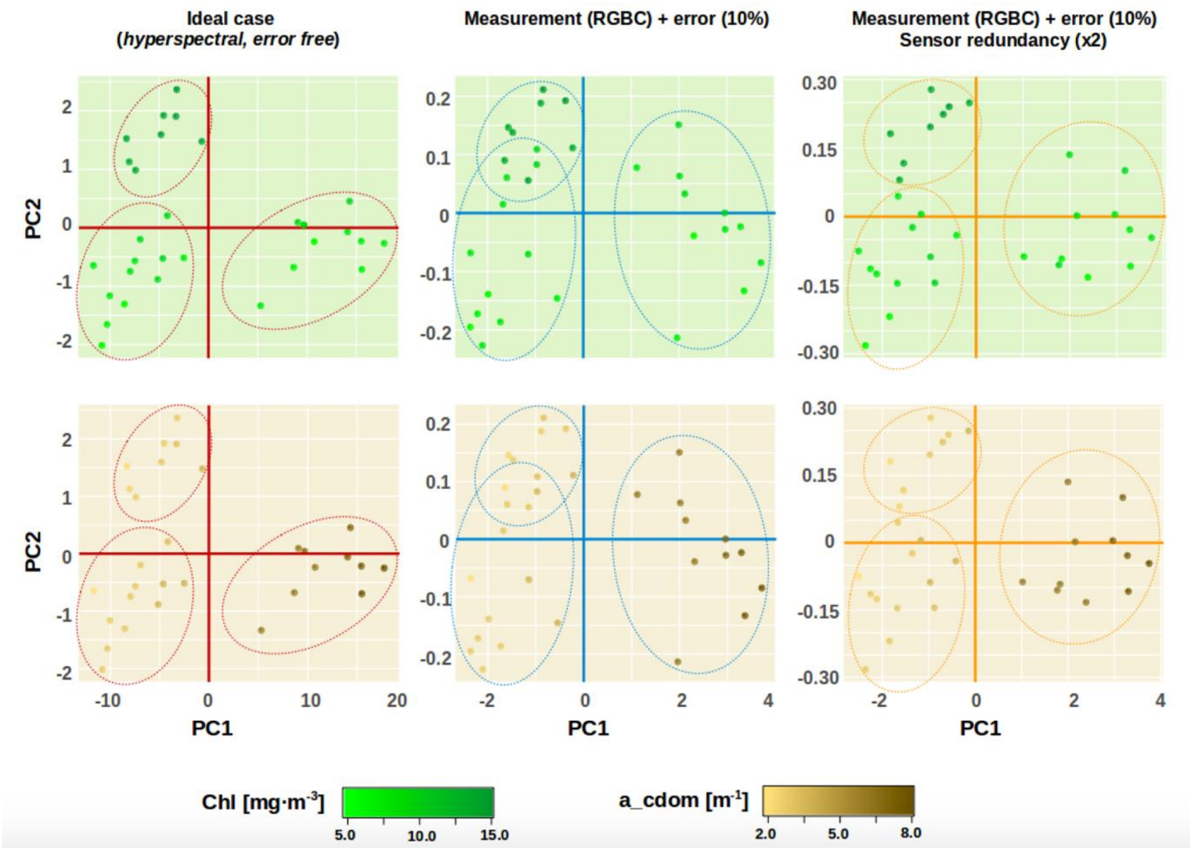


Figure 9 Example of the requirements evaluation for the new version of KdUINO using the outputs from numerical modelling. The simulations provide the data for classifying different water bodies based on the PCA analysis of the Kd measurements obtained.

Data Compression, standardised formats and access/sub-setting methods

Modern instruments have the capacity to generate enormous volumes of data at high spatial and/or temporal resolutions. This can place an unwelcome burden on networks and a cost to users to transfer and store data. Consider the case of a user who might wish to know the weekly, mean concentration of chlorophyll-*a* in a lake over the course of a year, it is much easier for them to download a summary time series than download all the data for the lake to their own machine and then perform operations upon it. This principle is one of the driving principles behind the move to cloud storage and cloud processing systems.

Sensor side processing of data is one method for reducing the pressure on sensor network transmission, but a number of other strategies at the system level can help with data transfer loads. These include suitable file compression, standardised file and data formatting, and methods for sending data subset requests to the network. This is covered in more detail in D5.2 'System architecture and standards report'.

Though the WISP-M sends raw data directly to an online database where local processing and quality control is performed, an API allows users to extract either L2 reflectance data or L1B data. WI intends to publish a dedicated Python library to extract sets of L1B data and process the spectra to L2 reflectance using various processing schemes for subset selection (from the ensembles) and glint correction.

CSIC will send data from KdUSTICK to a processing and storage server using Internet of Things technologies. Two different ways of communication are being considered: LoRa (by The Things Network) and Sigfox. Each method has associated limitations on the volume of data that can be transmitted as illustrated in Table 1.

Table 1 Differences between Sigfox and TTN

	Sigfox	TTN
Type of Technology	Private ($1 per device / yr)	Open and Free
Back-end	1 year	7 days
Uplink	140 messages @ 12bytes	30 seconds / day
Downlink	4 messages @ 8 bytes	10 messages / day
Bytes per day (Uplink)	1680 bytes	>6000 bytes
Max message size	12 bytes	5 seconds on air or 222 bytes

During the Lake Balaton field campaign, Sigfox technology will be tested as there is currently Sigfox coverage in most of the relevant area. Once data are received by the data server, CSIC will pre-process these and forward them using the Sensor Observation Service protocol to the MONOCLE back-end.

5. Enhanced information retrieval

All the approaches discussed below are designed to increase the quality, and utility of the retrieved data. Within the MONOCLE programme we are testing the approaches relating to two main goals; automated processing in order to flag anomalous data and, automated feature detection that might trigger changes to the sampling schemes etc.

It is worth noting that some sensors or sensor platforms perform automated checks on recorded data in order to filter out ‘bad’ data (such as negative reflectance values). This sort of filtering or correction should not be confused with the ‘anomaly detection’ discussed below, which is intended to inspect the data considered ‘good’ for unexpected anomalies in comparison to the rest of the datasets, rather than compared to some fixed thresholds. An example of this automated checking utilised within MONOCLE is the Simis and Olsson (2013) algorithm i.e. a fingerprint technique which attempts removal of atmospheric influence from the Rrs spectrum measured by sensors such as TriOS. Filters are also applied to eliminate spectra contaminated by sun glare, poor signal-to-noise, or non-water reflectance.

Anomaly detection

A number of statistical tests exist for the identification of outliers; in the context of MONOCLE we will focus on those that are most applicable to time series datasets and image data.

Statistical tests such as Grubbs test, Dixon's test, Hampel's test and quartile methods are candidates for identifying outliers in a given dataset but these methods are not necessarily suitable for use with time series data. Dixon's test for example can become hard to implement with a very large dataset and requires lookup tables. A Grubbs test requires the data to conform to a normal distribution, a fair assumption for some, but not all, datasets from natural waters. Additionally, Grubbs test requires the expected number of outliers to be predefined. The Hampel's test is theoretically more resistant to factors such as the volume of the dataset and the magnitude of an outlier. The robustness of the Hampel test is due to using the median to estimate the valid data 'centre' and median absolute deviation (MAD) to estimate the data standard deviation. Recent attempts to improve the capability of the Hampel test include the addition of Jacknife methodology (Lasisi and Shangodouin, 2014).

Alternatively, Seasonal and Trend decomposition using Loess (STL) is a robust method for decomposing time series through non-linear relationships. The STL method was developed by Cleveland *et al.* (1990). STL fundamentally decomposes a time series into a trend, a cyclical component and a residual as shown in

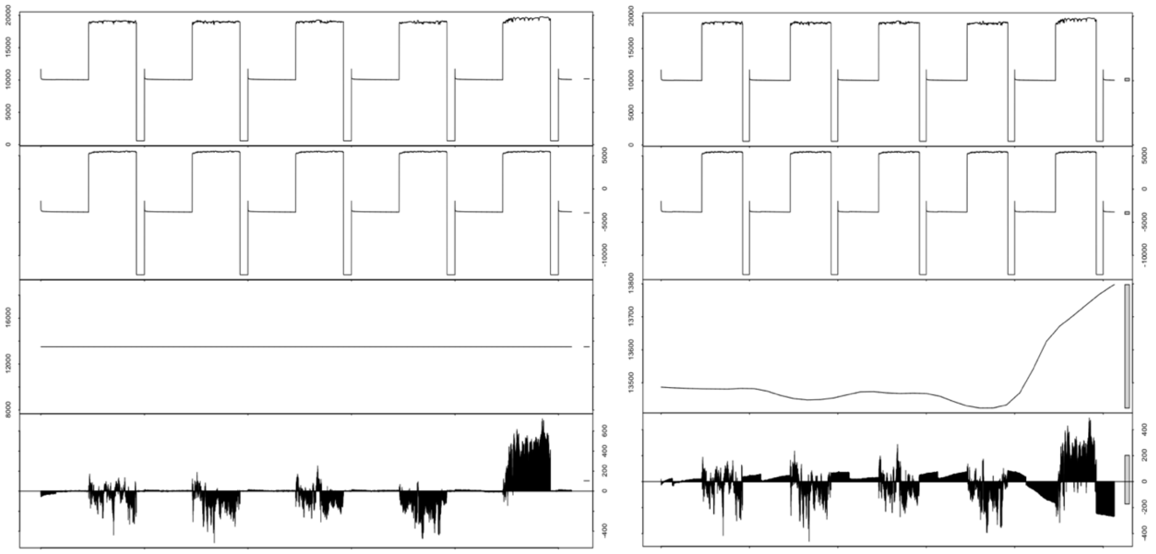


Figure 10. Both the trend and residuals can then be analysed for anomalies or significant trends that are worthy of note.

The Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) builds upon the Generalized ESD test for detecting anomalies. S-H-ESD has been shown to detect both global and local anomalies, where global anomalies are generally more apparent in the raw data streams while local anomalies tend to be visible only in the residual. The performance of S-H-ESD is achieved by employing time series decomposition and using robust statistical metrics, such as median (as in the Hampel test), together with ESD. For extended time series (e.g. 250,000 measurements), the algorithm can be implemented with a piecewise approximation. The piecewise approximation is used as a long time series can contain sufficient anomalies to interfere with the initial trend detection (Vallis et al 2014). An example of this approach is shown in Figure 11.

These techniques will be applied to data streams from the Kduino, CLAM, PAR and Radiometer sensors.

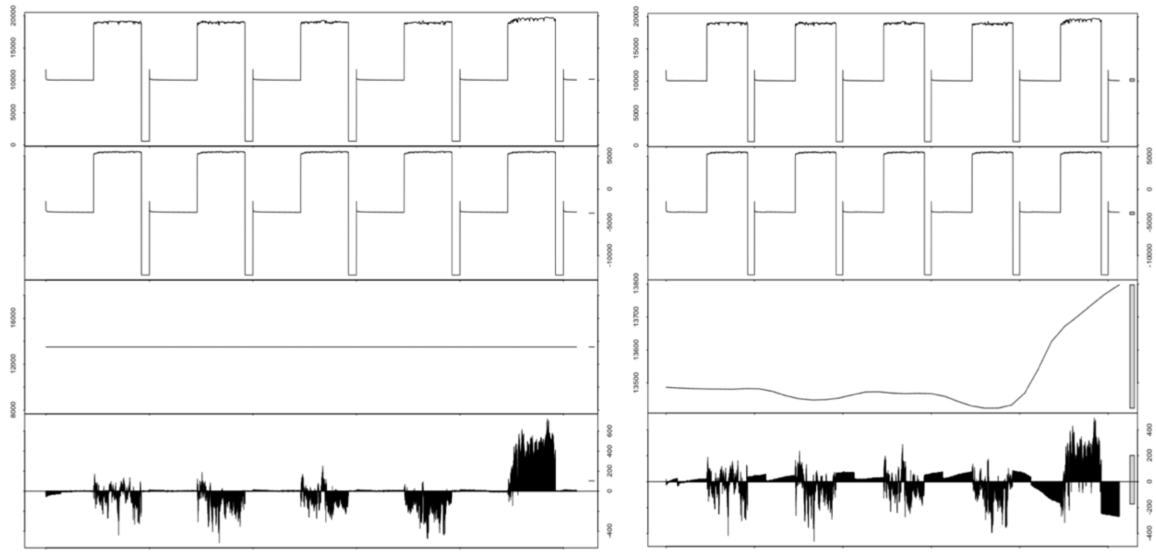


Figure 10: examples of STL decomposition for strongly cyclical data from a MONOCLE sensor. Left plots show decomposition with no trend permitted; right plots include a trend component. Top rows show input data, 2nd rows show cyclical component, 3rd row is trend and 4th is residual.

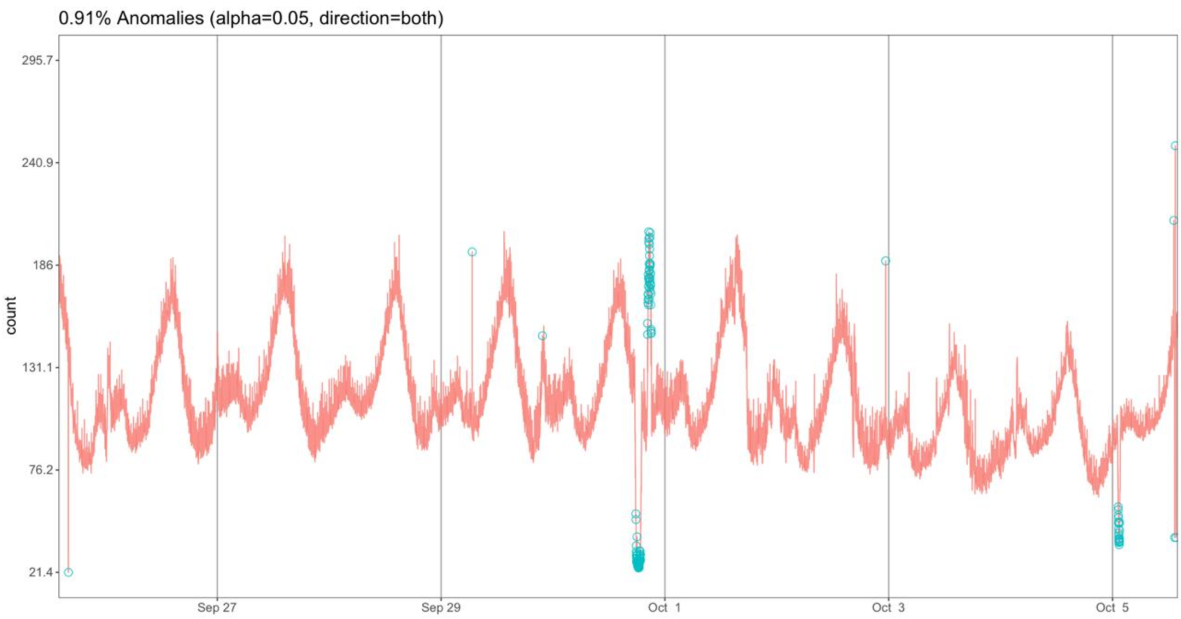


Figure 11: Example of anomaly detection in a time series using S-H-ESD approach.

Separating sensor degradation, drift and fouling

A key procedure to assure the quality of observations from reference sensors is to subject them to periodic calibration. This, however, is a prominent operational cost which causes significant disturbance to operational data gathering, particularly in remote locations.

For radiometry solutions, MONOCLE aims to combine sensor data with models as well as data filters to establish the degree of calibration drift vs sensor degradation and fouling, based on the following principles:

- Radiometric measurements contain information about illumination conditions, even when sensor drift has occurred
- The nature of downwelling solar (ir)radiance is such that cyclical patterns (day/night) and variable light intensity (according to sun angles and cloud cover) will be observed over time.
- Models can predict for a given location and time what the idealized sky radiance will look like.
- Over time, radiometric measurements will increasingly deviate from idealized sky radiance. This deviation is associated with sensor drift and/or fouling.
- Fouling can occur suddenly (detected as an anomaly with lasting impact) or gradually.
- In-field calibration of sensors against characterized targets, or other sensors, can help determine the nature of gradual sensor drift.
- Satellite overpasses provide an independent measure of the spectral nature of the up and downwelling light fields, albeit with their own uncertainties

Within MONOCLE, we will explore whether a decision-tree approach to quantifying the elements of sensor drift is a feasible way to establish the most strategic calibration intervals. This work is due to start when autonomous sensors are deployed at mature test sites, expected from summer 2019.

Automated Feature detection

Feature detection is more complicated than anomaly detection as here one is looking for coherent structures within datasets rather than point anomalies or step/phase changes. Feature detection analyses can be performed on time series in order to determine phenological metrics. Fitting of Gaussian peaks to annual time-series for example can be used to determine the features of algal growth cycles (Platt et al. 2009, Racault et al. 2016).

Feature detection on image data is particularly interesting as this can be used to elucidate the dynamics of water masses in an automated manner. Feature tracking in water bodies using observations such as drone images is possible but difficult. Three common assumptions in feature tracking methods are ‘Brightness consistency’, ‘Spatial coherence’ and ‘Temporal persistence’, all of which could be broken in some form by simple advection of waters upwards from below, or downwards beneath, the optical depth visible to the sensor.

Cross system information

In this context, ‘cross system information’ refers to the utilisation of data from one sensor in order to validate, compliment or enhance data from another sensor. This includes the communication of in-field calibration results and the utilisation of in situ sensor data attachment to relevant drone images. There are number of cases of this principle that will be tested within the MONOCLE programme.

A prime example of cross system information for complimentary data streams is the synergy between the HSP-1 (pyranometer), SO-RAD R_{rs} (using TriOS RAMSES) observations and use of the iSPEX along ship transects. Among other aspects, the combination would allow the modelling component (Gregg and Carder 1990) of the 3C approach (Groetsch et al. 2017) to be replaced with observations. As a first test we will use in situ measurements taken by the HSP-1 to take a ratio of the hyperspectral direct vs diffuse irradiance. We will then compare these values with modelled values calculated using the Gregg and Carder (1990) approach. There is also the possibility that photovoltaic panels (for example, the charge level registered by the solar charging controller onboard the TriOS platform) can be used as a proxy for E_d.

An example of cross system validation is the co-deployment of high and low cost sensors. MONOCLE provides the opportunity to examine concurrent data from the KdUSTICK alongside high cost estimates of absorption and scattering from the Wetlabs AC-S instrument. We can compare the low cost estimates of K_d with those we would expect given the IOPs measured by the ACS package. Additionally, in situ measurements will be made of phytoplankton pigment concentrations that can be compared to estimates from the AC-S and from remote sensors.

The HSP-1 collected data can also be used for the processing of drone imagery, as the direct vs diffuse radiance is important in the modelling of sky glint effects. Ultimately, drone data can be used in support of other sensors: e.g. when the WISP of KdUINO detect a sudden change in the water properties, they can send a trigger to a drone and conduct a drone flight to monitor a larger extent and detects where the anomaly comes from. In situ measurements are used to validate the drone products.

We will also use in situ measurements taken using existing in situ Rrs AERONET-OC instrumentation (5 years of historical data) to study the effect of using diffuse sky radiance as opposed to diffuse irradiance in that ratio. Using sky radiance distribution models, the intensity offset for a given viewing geometry may be correctable for clear sky conditions.

6. Exploitation and Dissemination

The information in this deliverable can be used in WP8 for dissemination and exploitation of the results from the field campaigns and to act as an initial reference point for testing and flagging of data streams from MONOCLE sensors. Protocols for use with the instruments following the field campaigns will be available, including sensor/platform details and documented test results from the field campaigns (e.g. complementary data streams and anomaly detection in sensor time series).

We will also provide test examples of the anomaly detection techniques alongside this data made available from WP5. This will be an ongoing process throughout sensor development after each planned field campaign.

7. Future activities/recommendations

Below is a summary table of the intended tests and required datasets to be acquired during MONOCLE programme in order to allow the investigation and refinement of the processing techniques discussed in the document above.

Use/test to perform	Required Dataset	Additional notes
'Representative' spectra generation from WISP-M measurements (averaging, deviant spectra exclusions and percentile filtering tests).	Database of WISP-M measurements at L1A.	
Performance estimates of WISP-M L1b to L2R processing options.	Independent in situ validation measurements taken alongside WISP-M L1B data in order to assess performance of each option.	
Demonstration of SNR improvement as a function of hovering time for drone imagery.	Repeated drone measurements over extended hovering period (perhaps up to 5 minutes) with reference panels and in-situ sampling location within field of view.	
Demonstration of increased KdUSTICK sampling rate at times of increased variability in the local environment (such as rainfall events).	KdUSTICK data at differing sampling rates during known 'event' (rainfall, introduction of sediment plume, etc)	It may be necessary to 'induce' an anomaly under lab conditions so that the exact timing and expected impact are known.
Resolving of resolve high-amplitude, short-duration light flashes caused by wave focusing using KdUSTICK measurements.	This requires sampling rates of up to 1 kHz, much higher required than for the routine estimation of K_d.	Ideally, these measurements would be taken under conditions of a regular and smooth swell of known period.
Demonstration of CLAM pathlength and integration time impact on SNR.	CLAM dataset in high and low turbidity waters with differing pathlengths and integration times.	Ideally this would be alongside independant estimates of Chl-a from either hplc, or extracted chl-a fluorometry, or using absorption standards.
Testing PCA based identification of water bodies from KdUINO/KdUSTICK data from environmental samples rather than modelling simulations.	KdUINO/KdUSTICK data across a known gradient (or transition) in water body proerties.	
Anomaly detection using multiple	Timeseries from KdUNIO,	Anomalies can be induced, e.g.

metrics for KdUNIO, CLAM, PAR and Radiometer sensor timeseries.	CLAM, PAR and Radiometer sensors with associated data on any known or induced 'anomaly' times.	producing bubbles, partially blocking sensor view, etc.
Trial feature detection and tracking from drone imagery.	Drone images of known transient feature such as river plume or drifting target.	The feature does not have to be naturally occurring as this is simply a test of tracking approaches, rather than tracking a particular class of feature.
Comparison of in situ data from (HSP-1, iSPEX, SO-RAD) and modelled irradiance values calculated using the Gregg and Carder (1990) approach in order to improve the 3C approach.	Co-incident measurements from suite of 3 sensors.	
Comparison of Kd estimates from KdUINO/KdSTICK with AC-S derived estimates.	Co-incident measurements from KdUINO and AC-S instruments.	

8. References

Berk A., Anderson G.P., Acharya P.K., Bernstein L.S. & Muratov L. (2006). MODTRAN5: 2006 update. Proceedings of SPIE 6233, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XII, 62331F.

Brando, V.E.; Lovell, J.L.; King, E.A.; Boadle, D.; Scott, R.; Schroeder, T. The Potential of Autonomous Ship-Borne Hyperspectral Radiometers for the Validation of Ocean Color Radiometry Data. *Remote Sens.* 2016, *8*, 150.

Bricaud, A.; Claustre, H.; Ras, J.; Oubelkheir, K. (2004) Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations. J. Geophys. Res. Oceans,109.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3–33.

Curtis D. Mobley, "Polarized reflectance and transmittance properties of windblown sea surfaces," Appl. Opt. 54, 4828-4849 (2015)

Darecki, M., D. Stramski, and M. Sokólski. 2011. Measurements of high-frequency light fluctuations induced by sea surface waves with an Underwater Porcupine Radiometer System. Journal of Geophysical Research, 116, C00H09. http://dx.doi.org/10.1029/2011jc007338

De Keukelaere L., Sterckx S., Adriaensen S., Knaeps E., Reusen I., Giardino C., Bresciani M., Hunter P., Neil C., Van der Zande D & Vaiciute D. (2018). Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. European Journal of Remote Sensing, Vol. 51, No. 1, 525-542.

Gregg, Watson W., Carder, K. L., (1990), A simple spectral solar irradiance model for cloudless maritime atmospheres, Limnology and Oceanography, 35, doi: 10.4319/lo.1990.35.8.1657.

Groetsch, P. M. M., Gege, P., Simis, S. G. H., Eleveld, M. A., & Peters, S. W. M. (2017). Validation of a spectral correction procedure for sun and sky reflections in above-water reflectance measurements. *Optics Express*, *25*(16), A742-A761. https://doi.org/10.1364/OE.25.00A742

Khan, Nawsher & Alsaqer, Mohammed & Shah, Habib & Badsha, Gran & Ahmad Abbasi, Aftab & Salehian, Solmaz. (2018). The 10 Vs, Issues and Challenges of Big Data. 52-56. 10.1145/206157.3206166.

Koch, David G. et al (2010) Kepler Mission Design, Realized Photometric Performance, and Early Science ApJL 713 L79.

Laney, D. (2001) 3D data management: Controlling data volume, velocity and variety, Application Delivery Strategies (META group).

Lasisi, T.A. and Shangodouin, D.K. (2014) Robust Test for detecting Outliers in Periodic Processes using Modified Hampel's Statistic. Theoretical Mathematics & Applications, vol.4, no.3, 2014, 25-42 ISSN: 1792- 9687 (print), 1792-9709 (online)

Platt T, White III, Zhai L, Sathyendranath S, Roy S (2009) The phenology of phytoplankton blooms: ecosystem indicators from remote sensing. Ecol Model 220:3057–3069

Racault, M-FLP; Sathyendranath, S; Menon, NR; Platt, T. (2016) Phenological Responses to ENSO in the Global Oceans. Surveys in Geophysics. https://doi.org/10.1007/s10712-016-9391-1

Raymaekers D., De Keukelaere L., Knaeps E., Strackx G., Decrop B. & Bollen M. (2017) Integrated approach to monitor water dynamics with drones. CEDA Dredging Days 2017 – Sustainable Dredging – Continued Benefits.

Ruddick, Kevin G., De Cauwer, Vera, Park, Young-Je, Moore, Gerald, (2006), Seaborne measurements of near infrared water-leaving reflectance: The similarity spectrum for turbid waters, *Limnology and Oceanography*, 51, doi: 10.4319/lo.2006.51.2.1167.

Simis, S.G.H.; Olsson, J. Unattended processing of shipborne hyperspectral reflectance measurements. *Remote Sens. Environ.* **2013**, *135*, 202–212.

Ur Rehman, M.H., Liew, C.S., Abbas, A. et al. Data Sci. Eng. (2016) 1: 265. https://doi.org/10.1007/s41019-016-0022-0

Vallis, O., Hochenbaum, J. and Kejariwal, A., (2014) "A Novel Technique for Long-Term Anomaly Detection in the Cloud", 6th USENIX Workshop on Hot Topics in Cloud Computing, Philadelphia, PA.

Westoby M.J., Brasington J., Glasser N.F., Hambrey M. J. & Reynolds J. M. (2012). 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, Vol. 179, 300–314.