Automated quality control of biogeochemical profiling float data using change-point detection

Tatjana E. Ellis, California Polytechnic University - San Luis Obispo

Mentors: Ken Johnson, Tanya Maurer, and Josh Plant

Summer 2018

Keywords: change-point detection, biogeochemical data, quality control, data corrections, Argo, SOCCOM

ABSTRACT

To better understand the biogeochemical processes of the ocean, many Argo profiling floats equipped with biogeochemical (BGC) sensors have been deployed. This includes numerous floats in the Southern Ocean, an area that is extremely vital to our ocean ecosystem overall, yet has been massively under-sampled. Many of these floats are part of the SOCCOM project, a project that aims to better understand the Southern Ocean. All float data from the active SOCCOM fleet is currently managed at MBARI, and must be processed and corrected as it comes in.

The number of active BGC floats is growing, however the sensors on these floats are still under development and any offsets and drifts apparent within the data due to sensor fouling must be corrected prior to scientific use. Until now, the data has been manually corrected by visually assessing it against deep reference fields to find change points from which to correct any sensor drifts and/or offsets. However, this is extremely time consuming, especially as the array of floats is growing, and includes a level of subjectivity, leading to an increased likelihood of over-correcting the data. Because of this, an autonomous method to find the optimal change-points has been implemented, which is being discussed in detail throughout this paper.

INTRODUCTION

To gather biogeochemical (BGC) data, many profiling floats within the Argo program are equipped with BGC sensors (Argo, 2018). The Argo program is a global array of free-drifting floats that profile up to 2000m depth every 10 days, and all gathered data is made publically available. They drift at a depth of 1000m for nine days, then descend to 2000m and finally collect a depth profile for all parameters as they ascend back to surface from depth. All Argo floats are equipped with Temperature, Salinity, and Pressure sensors, while the BGC Argo floats are equipped with additional pH, Nitrate, Oxygen, and bio-optics sensors.

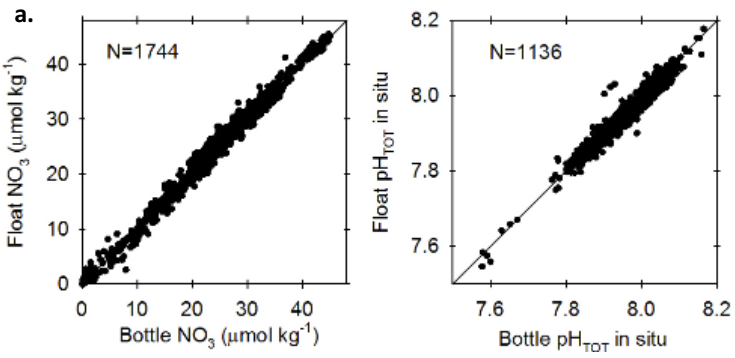
This gathered BGC data can help describe the biogeochemical processes occurring in the ocean, such as phytoplankton metabolism, carbon fluxes, and ocean acidification. The Southern Ocean is of particular interest in this regard, as it has a massive influence on the nutrient composition and distribution of the entire ocean ecosystem, and the effects of climate change are projected to be strongest here (SOCCOM, 2017).

However, due to extreme weather conditions, as well as other factors such as cost and location, the Southern Ocean has been massively under sampled. To combat this issue, the SOCCOM, or Southern Ocean Carbon and Climate Observations and Modelling project was born. This project aims to have 200 active BGC Argo floats out in the Southern Ocean. Currently 108 BGC Argo floats are actively sampling the Southern Ocean as part of the SOCCOM project. All floats within the SOCCOM project perform the nominal Argo mission described above, and all data are managed at MBARI.

Profiling floats significantly reduced the costs of data collection. Floats can stay out in the ocean for around 5-10 years taking samples, without anyone having to go out themselves. However, since the sensors are not usually recovered or re-calibrated for the period the floats are active at sea, the sensor response may drift with time, and this drift needs to be corrected.

Until now, all the corrections for pH and nitrate data have been done manually for each individual float, meaning someone visually assesses the data against high-quality reference datasets to derive any necessary adjustments. This has been working very well, as can be seen in **figure 1a**. When comparing the manually corrected float data to bottle measurements taken at the time of deployment, the relationship is very linear with a slope of one, showing the values align to each other. Additionally, the accuracy and precision for each parameter (based on float minus bottle difference statistics) for the fleet in **figure 1b** support that the manual corrections have been reliable. Accuracy here indicates the proximity of the manually corrected data to the true data, so the small values show that the two data sets are in very close proximity to each other. The small precision present shows that the corrected data does not deviate by much compared to the true data, as it represents the standard deviation between both data sets. This further confirms that the manual corrections have been accurate and reliable.

a.



b.

	Fleet average accuracy	Fleet average precision (SD)
NO_3	0.1 µmol/kg	0.5 µmol/kg
pH	0.005	0.008

Figure 1. Manually corrected data versus bottle collected data for pH and Nitrate parameters. **a.** Manually corrected float data versus bottle collected data for both parameters. **b.** Fleet average accuracy and precision (SD) for both parameters.

However, correcting manually does pose some issues: it includes an inherent level of subjectivity, which increases the potential for over-correction, meaning real data signals could be smoothed out, and most importantly it is extremely time consuming, especially as the number of active floats is growing. Correcting the entire SOCCOM fleet could take multiple days or weeks, depending on how many trained operators are present and active. For these reasons an automated approach was created, in order to objectively correct incoming data at a much faster rate.

METHODS

CORRECTION PROCESS

To correct data in any manner, whether manually or autonomously, there is a correction process that is being followed. First, the raw float data is extracted at depth between 1400 – 1500m. This data is then cleaned, meaning all bad data is removed as to not interfere with the correction. Since the data has not been corrected yet, the gathered data has not yet been categorized as good or bad. Therefore, bad data here refers to unavailable data that is being signified by a value of -1e10, by which it can easily be found and removed from the data set. Getting the data at depth is important since parameters such as nitrate or pH are more stable in this region. This is because there are less processes that penetrate deeper into the ocean, while those that do are usually slower and much easier to predict, resulting in much more predictable and stable parameter data. In contrast, those occurring near the surface, such as seasonal cycles, atmosphere exchanges, photosynthesis, and more can occur very fast and lead to more rapid and unpredictable changes in these parameters. Therefore, by using data at depth, we can better assess where the drift is occurring with a lesser chance of removing real data. Also, since we cannot readily re-calibrate the sensors, it is important to have a reliable reference to compare and correct this data to. For this we can compute deep reference fields using Locally Interpolated Regressions (LIRs) for pH (LIPHR) and nitrate (LINR; Carter et al., 2017).

These algorithms were fit to high-quality shipboard data (GLODAP v2 REFXXX; Olsen et al., 2016) and produce deep-sea reference fields based on available Temperature, Oxygen, and Salinity data for a certain depth and location specified. These have been shown to be very accurate, especially at depth where parameters are more stable and processes are better known, with error estimates of 0.006 pH units for pH and 0.47 μmol/kg for nitrate (Carter et al., 2017). Therefore, using LIR data allows the raw float data to be compared and corrected to a reliable standard.

Also, as seen in the analysis when comparing raw float data to LIR data in **figure 2**, they are often extremely similar, besides there being an offset or drift of some kind present in the raw data. This drift is what must be corrected for through this process.

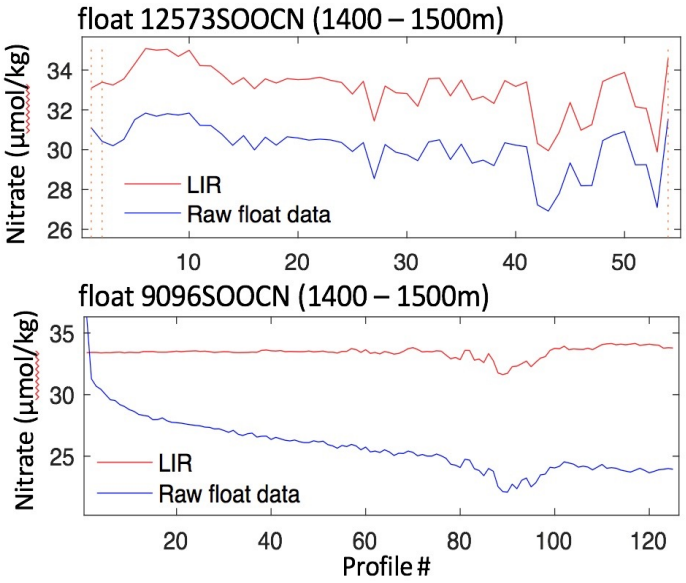


Figure 2. Raw nitrate float data plotted with deep-sea reference LINR data for floats 12573SOOCN and 9096SOOCN, at depth 1400 to 1500m.

Once the reference data has been generated, the residuals are computed by subtracting the reference from the raw float data for that specified depth. From these residuals, the change points are then computed by some method. Change points are defined as abrupt shifts in the parameters of a distribution or in the coefficients of a regression model (Beaulieu, Chen, Sarmiento, 2012). For this process, the change points are found using the residuals which show how much the raw data is offset from the reference data. If the offset is changing over time, then a drift is occurring, which the change points capture. This step has been done manually up until this point, meaning the best change points are visually assessed by trained operators, but has now been automated, as described in the following section.

After the change points have been found by either method, the offsets and drifts between these change points are calculated using least squares regression by the SAGE GUI, which stands for SOCCOM Assessment and Graphical Evaluation. This is software that was recently developed at MBARI and is continually being refined. SAGE can also store the generated correction values.

Lastly, this correction is applied to all depths for the float being corrected, since it is assumed that if a drift or offset is present at depth for a sensor, this same drift or offset will be present as the sensor rises to the surface.

AUTOMATED CHANGE-POINT DETECTION

Since manual change-point detection presents many disadvantages, as mentioned beforehand, this process of change-point detection has been automated. The automation has been primarily written in MATLAB, and it is planned to be integrated into the SAGE GUI for permanent use.

To find the best change-point locations statistically in the residual data set, MATLAB's built-in *ischange* function was used. As described on the MathWorks website, the *ischange* function finds points of distinct change in a series by iteratively minimizing the sum of the cost functions of each segment between potential change-points (MathWorks, 2018).

In order to find the optimal number of change points for a particular float's record, the *ischange* function is iteratively called on a range of values for the number of change points it should find, ranging from 1 until a specified cutoff maximum. The determination of the cutoff is pulled from Jones 1995, which states that the number of fitting parameters, k, must be smaller than the sample size of the data set, n, divided by 4 minus 1 (Jones, 1995). Using the value (n/4 – 1) as the cutoff maximum accounts for this qualification, ensuring that the sample size of the data set is large enough for a correction to be run. If it is not met, it means that not enough data is currently available for a change point correction scheme to be imposed on that float, and it will be excluded from the process until sufficient data is added.

Once a set of change points has been computed for each iteration, the quality of the resulting model generated from each respective set of change points was assessed using the Bayesian Information Criteria (BIC).

BIC equation

$$BIC = \log\left(\frac{SSR}{n} + \alpha^2\right) + \frac{K \log n}{n}$$

n = number of data points (n residuals)
SSR = sum of squared residuals
K = number of model parameters ((# change points + 1) * 2 + 2)
* One change point divides 2 regions each with a slope and intercept, so the number must be doubled
α = threshold cap on mean anomaly (representative of sensor accuracy)
= 0.004 for pH, 0.3 for NO3

The BIC is a statistical index that penalizes for too many change points, giving a good balance between goodness of fit and number of fitting parameters. This means that as the number of change points increases, even if the fit improves and the residuals get smaller, there may be too many change points, making it non-optimal. By using BIC as a determinant for the optimal number of change points, we are accounting for this issue. The associated change points of the iteration giving the lowest BIC are then stored and used to correct the data later-on, since a lower BIC corresponds to a more optimal model.

By combining the use of *ischange*, which statistically finds the best location of change points for a given number of change points, and BIC, which finds the best balance between number of change points and goodness of fit, we can objectively find the optimal number and location of change points for that specific data set, while reducing the possibility of over-correction.

RESULTS

EXAMPLE CORRECTION

Figure 3 shows an example nitrate correction for float 9313SOOCN, one of the SOCCOM floats that has been sampling the Southern Ocean for almost four years.

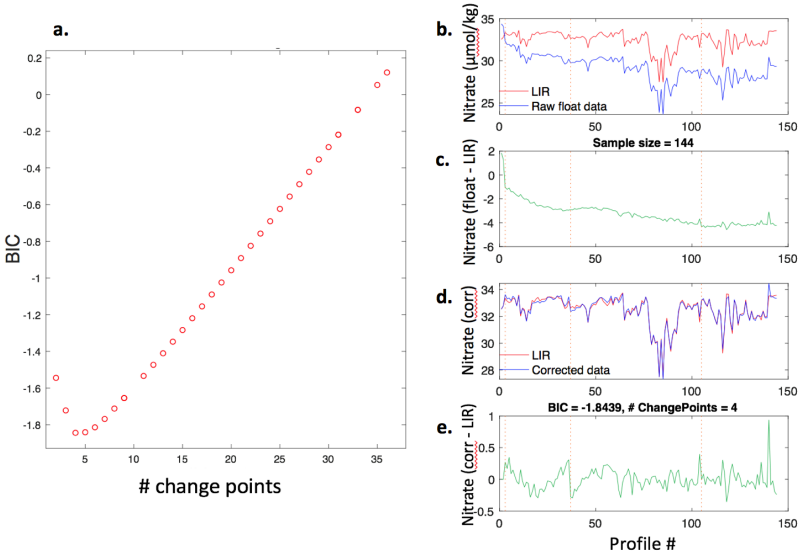


Figure 3. Example nitrate correction for float 9313SOOCN. The dotted orange vertical lines in **b.-e.** indicate change point locations. **a.** BIC values versus number of change points. **b.** Raw float data plotted with LINR data. **c.** Residuals of raw data minus LINR data**. d.** Corrected data plotted with LINR data. **e.** Residuals of corrected data minus LINR data.

Figure 3a shows BIC values versus the number of change points. It becomes visible that as the number of change points increases past an optimal value chosen, in this case four, the BIC starts increasing in value, penalizing for too many change points. This implies an optimal balance using the chosen value of four change points. Using these optimal

change points, whose locations are indicated by the spotted orange vertical lines in **figure 3b-e**, the corrected data is computed by finding the slopes and intercepts between them in the raw float data set, and by these removing the drift from the raw data set.

When looking at the plot in **figure 3b** showing the raw data versus the LINR data that was computed at depth, one can see how similar they are in structure, besides being separated by an initial offset of -1 at profile 3 which increases with time, meaning the sensor is drifting. **Figure 3c** shows these offsets, where the increasing drift of the raw data from the LINR becomes more noticeable. From this data set the change points are detected. The following plot in **figure 3d** shows the corrected data versus the LINR, which now match up very well. This becomes more apparent through the final plot in **figure 3e** showing the residuals computed by subtracting the LINR data from the newly corrected data. The residuals are within a range of plus or minus 0.5 µmol/kg on average, showing how closely the corrected data matches the reference, implying a good correction.

METHOD COMPARISON

When comparing the manual method to the autonomous (auto) method in **table 1** for the same float 9313SOOCN used above, an improvement in the autonomous method is visible. The BIC value for the autonomous method is -1.8, which is slightly lower than that of the manual of -1.6. This difference of 0.2 is notable, since the autonomous method distinguished the best BIC value even to such a small scale (**figure 3a**). This shows that the autonomous method always finds the lowest BIC value possible for a given data set.

Statistic	Manual	Auto
BIC	-1.6	-1.8
# change pts	8	4
change pts	1, 2, 3, 16, 33, 55, 115, 130	1, 3, 37, 105
Sum squared residuals	3.13	3.74

Table 1. Statistics for float 9313SOOCN nitrate corrections comparing manual method to auto method.

What is more noticeable is that the number of change points for the autonomous method in comparison to the manual method was halved, from 8 change points to 4 change points. This is a big difference, since it shows that manually the data may have been over-corrected. When looking at the change point locations, the optimal change

points are very similar to four of those chosen manually, within a maximum range of 10 change points apart. The other four, however, were not chosen to be necessary or optimal by the autonomous method, showing that although they are usable, they are not necessary to perform a good correction.

The average residuals are slightly better for the manual method, since more change points have been chosen, allowing for a tighter fit. However, when taking the fit as well as the number of change points into account, the autonomous method chose only four. And as shown in **figure 3e,** the calculated residuals after correction using these four change points are within +/- 0.5 µmol/kg which is considered to be within the accuracy of the sensor.

OVERALL METHOD COMPARISON

a.

Param	Statistic (Avg.)	Manual	Auto
pH	BIC	-8.98	-9.92
	diff. in BIC	0.95	
	# chPts	4.88	2.98
	diff in # chPts	1.9	
NO3	BIC	-0.81	-1.69
	diff. in BIC	0.87	
	# chPts	4.82	2.68
	diff in # chPts	2.14	

b.

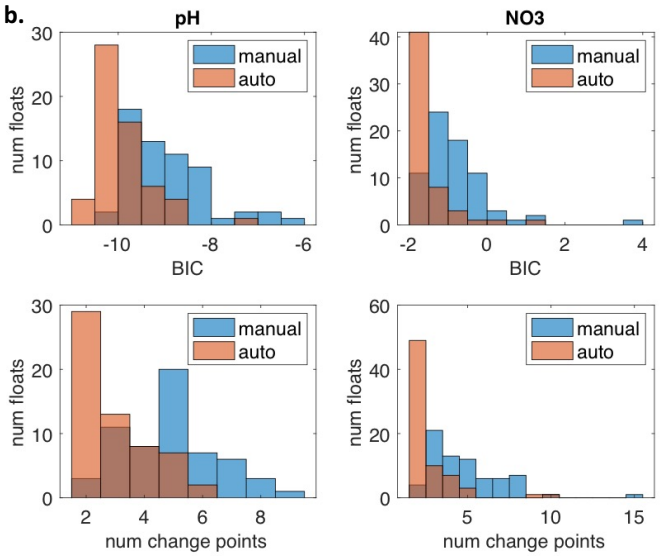


Figure 4. Comparison statistics between manual and auto correction methods for all available SOCCOM floats. **a.** Statistic table, showing each average value for both pH and Nitrate data. **b.** Histograms for both pH and Nitrate data statistics: number of floats versus BIC in the top two charts, number of floats versus number of change points in the bottom two charts.

This newly implemented approach to change point detection has been tested and run on approximately 60 floats for pH, and approximately 80 floats for nitrate within the SOCCOM fleet. Comparing both methods over all these floats for each parameter separately, as seen in the statistics table in **figure 4a**, the average BIC value for both pH

and nitrate is lower for the auto versus the manual method. Also, as seen by the charts in **figure 4b**, the spread of the BIC values is smaller in the auto method than the manual. This further confirms that the autonomous method is finding more optimal values, and does this more consistently by computing them statistically.

Another noticeable difference is that, on average, the auto method almost halves the number of change points chosen in the manual method. Such a result may imply that manually the data has been over-corrected, and that the sensors are better than thought when visually assessing the data. However, some of this discrepancy can be attributed to the manual assessment imposing multiple offsets during the first part of the record instead of a single offset with a drift.

DISCUSSION

Both the manual and autonomous change point detection methods work very well, giving us accurate and usable results. However, given the disadvantages of the manual method, implementing and using an autonomous method is much more sustainable in the long run.

As shown in the results, the autonomous method on average halved the number of change points used, and has a narrower and lower range of BIC values. The autonomous method calculates the statistics very consistently over all floats, allowing for this narrower spread, whereas the manual method may vary depending on which operator is assessing the data, as well as many other conditions that may influence the operator. When correcting autonomously, the corrections are also much more objective, since they are entirely computed based on optimal statistics, thus removing the inherent level of subjectivity present when manually correcting data.

The biggest advantage of automating the correction process is time consumption, since even if manually correcting data results in accurate, scientific quality-data, it takes a lot of time to produce. Analyzing all incoming float data, where there are currently over 100 active floats and the number is growing, and continuously re-visiting floats to account for newly added profiles, takes a lot of time. Correcting all the floats can take

hours, days, weeks, or even months, depending on how many operators are trained, present, and available to work on this task. In contrast, the autonomous method can run its corrections and find the optimal values for all the floats available for both pH and nitrate in under 20 minutes, and no operator is required to do anything in that time. The autonomous method not only optimizes the corrected data, but saves hours and weeks' worth of work that can now be used freely to accomplish other research or tasks.

Due to its advantages, this automated process will be integrated into the SAGE GUI at MBARI in order to support the selection of change points while the data is being corrected. However, before full automation is implemented, its performance will be closely monitored for a period of time, and improvements and optimizations will be added as needed. If desired, this process can also be applied to many other devices that take depth profiles, as well as other parameters, such as oxygen data.

CONCLUSION

In conclusion, automating the correction process for incoming float data has shown to be more reliable and consistent, and it is much more effective in time consumption. It is a program that can be easily modified and applied to different devices or parameters if needed, proving to be very versatile and easily applicable.

With this, processing and correcting incoming data can be much more efficient, allowing for the active fleet to grow in number, without having to spend additional hours to maintain the corrected data.

ACKNOWLEDGEMENTS

I would like to acknowledge my advisors, Tanya Maurer, Josh Plant, and Ken Johnson, for allowing me to work for them this summer, introducing me to how data is collected and processed, and helping me accomplish the project goal. Additionally, I am very grateful to the MBARI staff that helped plan and organize the internship, as well as my fellow interns. All these people were helpful in allowing me to complete this research project, and allowed me to learn a lot of vital information to help me accomplish my task.

I also acknowledge the David and Lucile Packard foundation, for allowing the MBARI Summer Internship Program to take place, as well as all the National Science Foundation for funding the SOCCOM project, a great project allowing us to better understand and study the ocean.

References:

- Argo (2018). http://www.argo.ucsd.edu
- SOCCOM (2017). https://soccom.princeton.edu
- Carter, B.R., R.A. Feely, N.L. Williams, A.G. Dickson, M.B. Fong, and Y. Takeshita **(**2017). Updated Methods for Global Locally Interpolated Estimation of Alkalinity, pH, and Nitrate. *Limnology and Oceanography: Methods*, 16: 119-131. https://doi.org/10.1002/lom3.10232
- Olsen, A., Key, R.M., van Heuven, S., Lauvset, S.K., Velo, A., Lin, X., Schirnick, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishii, M., Pérez, F. F., and Suzuki, T. (2016). The Global Ocean Data Analysis Project version 2 (GLODAPv2) – an internally consistent data product for the world ocean. *Earth System Science Data*, 8: 297-323. doi.org/10.5194/essd-8-297-2016
- MathWorks (2018). https://www.mathworks.com/help/matlab/ref/ischange.html
- Jones, R.H., Dey, I. (1995). Determining one or more change points. *Chemistry and Physics of Lipids:* 76: 1-6. https://doi.org/10.1016/0009-3084(94)02422-2
- Beaulieu, C., Chen, J., Sarmiento, J.L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society*: 370: 1228-1249. doi:10.1098/rsta.2011.0383