RECOMMENDATIONS FOR ENHANCEMENT OF O2 QC METHODS: GUIDE TO DRIFT CORRECTION PROCEDURE FOR OXYGEN OPTODES ON ARGO FLOATS

Ref.: D4.6_V1.0

Date: 14/12/2021

Euro-Argo Research Infrastructure Sustainability and Enhancement Project (EA RISE Project) - 824131

Under EC review

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 824131.
Call INFRADEV-03-2018-2019: Individual support to ESFRI and other world-class research infrastructures

Disclaimer:

This Deliverable reflects only the author's views and the European Commission is not responsible for any use that may be made of the information contained therein.

Document Reference

Project	**Euro-Argo RISE - 824131**
Deliverable number	**D4.6**
Deliverable title	**Recommendations for Enhancement of O2 QC Methods: Guide to Drift Correction Procedure for Oxygen Optodes on Argo Floats**
Description	**Recommendations for enhancement of O2 QC Methods**
Work Package number	**4**
Work Package title	**Biogeochemical Extension**
Lead Institute	**GEOMAR**
Lead authors	**Arne Estelmann, Henry Bittig, Arne Körtzinger**
Contributors	**Violetta Paba, Matt Donnelly (BODC)**
Submission date	**14. 12. 2021**
Due date	**31. 12. 2021**
Comments	
Accepted by	**F. D'Ortenzio**

Document History

Version	Issue Date	Author	Comments
Ver. 1.0	14. 12. 21	A. Estelmann	

EXECUTIVE SUMMARY

Dissolved oxygen arguably is the most mature biogeochemical variable that can be routinely observed from Argo floats. Large and successful efforts have been made to characterize the most commonly used optode-based oxygen sensor, which allowed for the development of robust best practice recommendations in data quality control. A milestone in this context is the implementation of an in-air-measurement routine, which forms the crucial basis for the delayed mode data control and is now a firm recommendation for all BGC-Argo floats. What was missing so far, is a consistent, evidence-based procedure for treatment of the four identified components that need to be addressed in the drift correction (initial gain change, carry-over effect from surface water on air measurement, inadequate characterization of temperature response, in-situ drift). Here we present a guide for the delayed mode correction of dissolved oxygen data that is based on an in-depth analysis of 178 floats and proposes a decision path which takes the availability of in-air measurements and the float's duration time into account and follows a suite of rules. In the final step, information criteria are employed to choose and evaluate the optimal correction for a given time.

TABLE OF CONTENTS

1 Background

1.1 Need for Oxygen Sensor Drift Compensation

Oxygen sensors on Argo floats exhibit a drift, both while being stored prior to deployment as well as during their deployment (Bittig, Körtzinger et al. 2018 [3]). While careful calibration, conducted immediately before the deployment, can prevent the former, the latter can only be compensated by *in situ* drift correction during and/or after the float's lifetime. Since oxygen sensors used in Argo floats have been demonstrated to be capable of in-air measurements, Körtzinger et al. 2005 [8] proposed such in-air records for drift control. In-air measurements can be conducted during surfacing for satellite communication, as Argo floats need to ascend to the surface of the water column for data transmittance.

The principle of in-air measurements is described by Bittig, Körtzinger et al. 2018 [3]. Following this concept, a slope m is defined that is used to correct any observed partial pressure of oxygen $pO_2^{obs.}$ (1).

$$pO_2 = m \cdot pO_2^{obs.} \tag{1}$$

m is the ratio of the the actual atmospheric partial pressure of oxygen $pO_{2,air}$ to the partial pressure of oxygen $pO_{2,infl}^{obs.}$ observed by the float surfacing in fully inflated mode (2). To obtain $pO_{2,air}$, a uniform mixing ratio of oxygen in the atmosphere of $\chi_{O_2} = 20.946$ % is assumed. Then, oxygen partial pressure pO_{2air} is calculated by use of surface air pressure p_{air} and surface atmospheric partial pressure of water vapor $pH_2O_{surf.}$, derived from surface temperature and salinity (3). For surface air pressure, NCEP/NCAR Reanalysis-1 data are used (Kalnay et al. 1996 [7]). As the optode height and thus the elevation over the water surface varies between different Argo types, a scaling factor x is added (see Bittig and Körtzinger 2015 [2] for further details).

$$m = \frac{pO_{2air}}{pO_{2infl.}^{obs.}} \tag{2}$$

$$pO_{2air} = \chi_{O_2} \cdot \left(p_{air} - x \cdot pH_2O_{surf.}\left(T_{surf.}, S_{surf.}\right) \right) \tag{3}$$

1.2 Components of Oxygen Sensor Drift Compensation

Four different drift effects that contribute to m can be identified: change in sensor's O_2-gain, a carry-over effect of any surface water on the sensor's optical window during air measurement, an inaccurate characterization of the sensor's temperature dependence, and an *in situ* drift of the sensor response. The so-called O_2-gain is the basic, underlying correction of the optode signal that is free of any further influencing parameter such as temperature or deployment duration, hence it is applied to all data in equal measure (4).

$$m = 1 + \frac{b}{100} \tag{4}$$

However, the sensor might be submerged or wetted by sea spray during the in-air measurement, particularly in rough seas. Hence, a bias towards the surface water pO_2, referred to as carry-over effect, can occur. This can be corrected by a slope c, considering surface water measurements of the deflated float $pO_{2,defl}^{obs.}$ that are routinely performed immediately prior to $pO_{2,infl}^{obs.}$. Thus, the carry-over correction

is not a drift correction of the Argo float data but a corrective variable for all air measurements made as reference (5).

$$pO_{2,\text{infl.}} - pO_{2,\text{air}} = c \cdot \left(pO_{2,\text{defl.}} - pO_{2,\text{air}}\right) \tag{5}$$

As the optode sensing principle is significantly temperature-dependent, temperature effects must be considered when sensing foils are calibrated. The excited-state lifetime of the luminescent dye incorporated in the membrane decreases at a higher temperature, while oxygen quenching efficiency increases. The sensor's temperature-dependence is carefully characterized by the initial multi-point calibration. In some cases, however, the temperature dependence may not be adequately captured by the sensor calibration, e.g., for older optodes that used a batch calibration (Bittig, Körtzinger et al. 2018 [3]). Hence, a temperature correction term a can be introduced, which leads to equation (6). Therefore, the in situ temperature in °C is termed as $\vartheta_{\text{Optode}}$.

$$m = 1 + \frac{a \cdot \vartheta_{\text{Optode}}}{100} \tag{6}$$

The fourth correction parameter $b_{\text{in situ}}$ accounts for any drift of the optode during the deployment time $t_{\text{deployment}}$, since sensing foils tend to exhibit a small but significant sensitivity loss over multiple years (7).

$$m = 1 + \frac{b_{\text{in situ}} \cdot t_{\text{deployment}}}{100} \tag{7}$$

1.3 Need for Drift Correction Guidance

Since floats undergo different oxygen sensor calibration procedures, are exposed to different environmental conditions and age individually, not all four correction parameters are significant and have to be considered for every float. As a general rule, the degrees of freedom of the fit function should be kept as small as possible to avoid over-fitting. Hence, an informed and objective decision has to be taken whether the inclusion of a given correction parameter is of advantage or instead may even lead to information loss. A decision tree is therefore proposed to guide the optimal complexity of the applied correction factor m.

In order to facilitate the discussion, a three-digit binary flag is introduced to indicate the composition of m. Digits of the binary flag are, from right to left, carry-over 1, temperature gain 10 and *in situ* drift 100. The O_2 gain b is defined as being not facultative, as it is the fundamental base of the correction principle. According to this, a binary flag value of 3, i.e. 11, includes equations (4,5,6). A binary flag value of 7, i.e. 111, instead encompasses all four factors and leads to fit function (8). A binary flag value equal to zero only includes b and thus has only one degree of freedom.

$$pO_{2,\text{air}} = \left(1 + \frac{b + a \cdot \vartheta_{\text{Optode}} + b_{\text{in situ}} \cdot t_{\text{deployment}}}{100}\right) \cdot \frac{pO_{2,\text{infl.}}^{\text{obs.}} - |c| \cdot pO_{2,\text{defl.}}^{\text{obs.}}}{1 - |c|} \tag{8}$$

Our work was guided by the goal to establish a decision tree that suggests the best suited binary flag for a given float data series at a given time. This decision tree should be reasonably simple and practical, avoid over-fitting and frequent binary flag changes over time. Fit parameters should be as stable as possible so that a chosen binary flag and the respective fitting values would persist. Moreover, a

time frame should be provided for active floats to guide the timing of revisions of the chosen binary flag. Hence, statistical or empirical key criteria have to be identified that indicate both fit quality and stability.

Following this decision tree, drift correction could be integrated into the real-time quality correction performed at data acquisition centers.

2 Evaluation of Drift Correction Performance

A fleet of 49 Coriolis floats and 129 MBARI floats that all performed in-air oxygen measurements were used to investigate the stability and quality of possible drift corrections. This represents the full subset of floats for which in-air data were available as of Nov. 2020 and which possessed a sufficiently long time series of at least 2 years to be included in our analysis. Floats that showed clear signs of a defect, such as oxygen depletion for in-air measurements or sudden but significant, persistent increase in measurement uncertainty or oxygen concentrations, were not used after such defects arose. Few singular outliers in the reciprocal time series of m were removed as well.

For the entire fleet of 178 floats, robust fits of binary flags 1, 3, 5 and 7 were calculated every 90 days, starting at 180 days of their deployment. Drift correction could thus be calculated over time frames of different length. The results were visualized for each float to evaluate the temporal stability of fitting parameters (see figure 1 for an example). The subsequent, future drift development could then be compared with the fitting parameters at a given time.

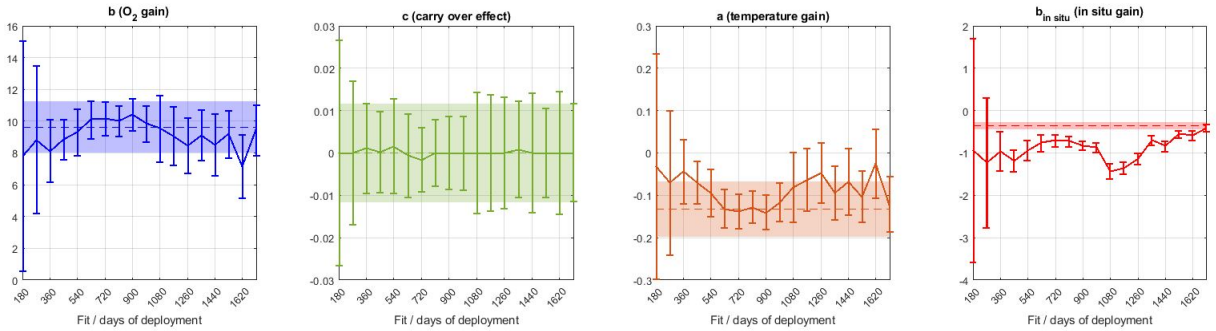

Figure 1: Slopes b (blue), c (green), a (orange) and $b_{\text{in situ}}$ (red) of float 3901084 at binary flag 7, calculated for every 90 days since day 180 of its deployment. A fit over the entire deployment time gave the dotted line and error bars indicated by the colored area.

2.1 Manual Overview

In a first step, the quality of the drift corrections was evaluated manually for the 49 Coriolis floats. Therefore, the quality of the fit was observed visually and quantitatively by use of the reciprocal of slope m (9) and a post-correction version m_{corr}^{-1} (11).

According to equation (8), m^{-1} of a binary flag is equal to:

$$m^{-1} = \frac{pO_{2,\text{infl.}}^{\text{obs.}} - |c| \cdot pO_{2,\text{defl.}}^{\text{obs.}}}{(1 - |c|) \cdot pO_{2,\text{air}}} \tag{9}$$

Replacing $pO_{2,\text{infl.}}^{\text{obs.}}$ and $pO_{2,\text{defl.}}^{\text{obs.}}$ by their drift-corrected counterpart, m_{corr}^{-1} is obtained. For a float with ideal drift correction, m_{corr}^{-1} should reach unity.

$$m_{\text{corr}}^{-1} = \frac{pO_{2,\text{infl.}} - |c| \cdot pO_{2,\text{defl.}}}{(1 - |c|) \cdot pO_{2,\text{air}}} \tag{10}$$

$$m_{\text{corr}}^{-1} \longrightarrow 1 \tag{11}$$

The more m_{corr}^{-1} of a single cast deviates from unity, the less suitable is the applied fit for the respective cast. Calculated for multiple, consecutive casts, $m_{\text{corr}}^{-1}(t)$ can be plotted to give a powerful visualization and quantification tool for the fit and binary flag quality over the deployment time (figure 2).

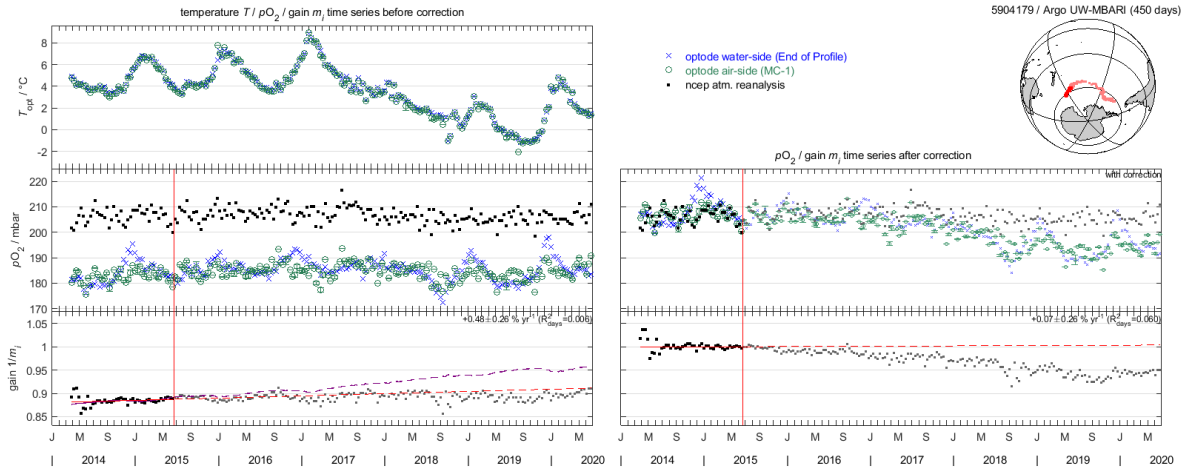

Figure 2: Binary flag 7 fit for float 5904179 at day 450 (red vertical line). In the upper right, the drift trajectory of the float is shown. In the upper left, the temperature during in-air (green circles) and surface water measurements (blue crosses) is shown. Below, $pO_{2,\text{infl.}}^{\text{obs.}}$ (green circles) and $pO_{2,\text{defl.}}^{\text{obs.}}$ (blue crosses) as well as pO_2 data (black dots) are shown. On the right, $pO_{2,\text{infl.}}$, $pO_{2,\text{defl.}}$ and pO_2 of the fitted deployment time (normal colors) and the future deployment time (faded colors) are given. Below, m^{-1} (left, with the dotted violet drift correction fit) and $m_{\text{corr.}}^{-1}$ (right) of the fitted deployment time (black dots) and the future deployment time (grey dots) are given. While for the first two to three years of the deployment, the fit at day 450 compensates for the difference between optode measurements and pO_2, the fit overestimates the *in situ* drift. Hence, future m^{-1} shows an offset from unity growing with deployment time.

For every moment t_i since the deployment start of a float t_0, multiple parameters were considered to asses fit quality: the absolute changes, the relative changes of the fitting parameters and the ratio of the 95 % confidence interval to the absolute value of each parameter were evaluated. The root mean square error, the median and the median absolute deviation of $m_{\text{corr}}^{-1}(t_0, t_i)$ were investigated. The performances of the fit parameters obtained for [t_0;t_i] and those obtained for [t_i;t_n], with t_n being the

latest available in-air measurement date of a float, were compared by calculating the root mean square error, the median and the median absolute deviation of $m_{\text{corr}}^{-1}(t_0, t_i)$, $m_{\text{corr}}^{-1}(t_i, t_n)$ and $m_{\text{corr}}^{-1}(t_0, t_n)$ for every parameter set. Possible temperature dependency of fitted data was detected by the slope of a linear correlation of m_{corr}^{-1} and the respective air-side optode temperature as well as reviewing the temperature range the float has been exposed to.

Fits proved to be more stable when the degrees of freedom were reduced. The more complex binary flags (3,5,7) showed larger instability concerning parameter values in almost every case. Often, seasonality was misleadingly taken for a drastic long-term drift, particularly in short time series (i.e., $t_1 \leq 365$ days), leading to over- or underestimation of the actual trend. The only fitting parameter that stabilized quickly was the basic O_2 gain correction b. Usually, a and $b_{\text{in situ}}$ converged towards a final stable value, too. However, they required significantly more time (multiple years) for such stabilization. Especially a showed high variability and frequent changes. Thus, we conclude that using a fit with a binary flag ≥ 1 is not recommended during the first 365 days of deployment. We therefore propose to uniformly only apply the basic O_2-gain correction factor b to all oxygen floats during the first 365 days of their life.

Under the condition that the only allowed binary flag changes are 1 $\longrightarrow$ 3/5, 3/5 $\longrightarrow$ 5/3 and 3/5 $\longrightarrow$ 7 for this manual evaluation, dates for the change of the binary flag t_{change} were determined subjectively. By that, fit changes were deemed reasonable only if the more complex fit remained stable. Based on this, fits that included a temperature drift (binary flags 3,7) seemed to be appropriate only for an absolute temperature dependency of m_{corr}^{-1} of 0.1 °C^{-1} or higher. However, no further absolute or relative criteria that did not rely on the "future" development $m_{\text{corr}}^{-1}(t_i, t_n)$ could be found.

2.2 Introduction of Information Criteria

As the previous search for binary flag criteria remained unsatisfactory, information criteria were introduced. Information criteria estimate the quality of a model by considering not only its performance but also its complexity, as the most popular information criteria consist of a part considering the likelihood function (quantifying the goodness of the fit) and a second part of simple penalties for the degrees of freedom to avoid over-fitting (Dziak et al. 2020 [5]). Information criteria have been used successfully for conductivity drift correction of Argo floats (Owens and Wong 2009 [10]). After completion of the work shown here, Maurer et al. 2021 [9] published an oxygen sensor drift correction for BGC-Argo floats that also relies on an information criterion.

The three information criteria used here are the *Akaike* Information criterion (AIC, see Akaike 1974 [1]), a bias-corrected AIC (AIC_C, see Hurvich and Tsai 1989 [6]) for small data sets and the Bayesian Information Criterion (BIC, see Schwarz 1978 [11]). All three consist of the negative natural logarithm of the maximum value of the likelihood function of the model $\hat{\mathcal{L}}$ and a penalty term that is multiplied with the degrees of freedom k (12) that is, in fact, the number of fitting parameters (b, c, a, $b_{\text{in situ}}$) increased by one due to the natural variability of the fit (Dziak et al. 2020 [5]).

$$\text{IC} = -2 \cdot \hat{\mathcal{L}} + k \cdot [\text{penalty}] \tag{12}$$

For a regression model with n data points that have independent, normally distributed errors, $\hat{\mathcal{L}}$

equals the residual sum of squares divided by n (Burnham and Anderson 2002 [4]). Thus, the information criteria can be calculated using the equations (13-15).

$$\mathrm{AIC} = n \cdot \ln\left(\frac{1}{n}\mathrm{SSR}\right) + 2k \tag{13}$$

$$\mathrm{AIC_C} = n \cdot \ln\left(\frac{1}{n}\mathrm{SSR}\right) + 2k \cdot \frac{n}{n-k-1} \tag{14}$$

$$\mathrm{BIC} = n \cdot \ln\left(\frac{1}{n}\mathrm{SSR}\right) + k \cdot \ln n \tag{15}$$

As the information criteria such as AIC, in particular, do not punish the number of parameters for very small data sets sufficiently, (e.g. $\frac{n}{k} \leq 40$), information criteria should be handled with care if a lot of fitting parameters were used for very small n (Burnham and Anderson 2002 [4]). Hence, all information criteria were manipulated to give back an empty value as soon as $4 \cdot k > n - 1$, as was done before by Owens and Wong 2019 [10] for Argo conductivity data.

Although AIC, $\mathrm{AIC_C}$ and BIC cannot be compared directly, the values of one information criterion can be compared with the same information criterion for all binary flags obtained for a date t_i. The smaller the information criterion value, the more suitable is a given fit for the oxygen drift correction. Thus, the different fits can be ranked according to their quality.

AIC, $\mathrm{AIC_C}$ and BIC of all fits for every float, made every 90 days since day 180 of the respective deployment, were calculated (figure 3). The ranking obtained and the suggested fit by minimum information criteria value were checked against the manually suggested best fit and compared with the possible criteria obtained in section 2.1.

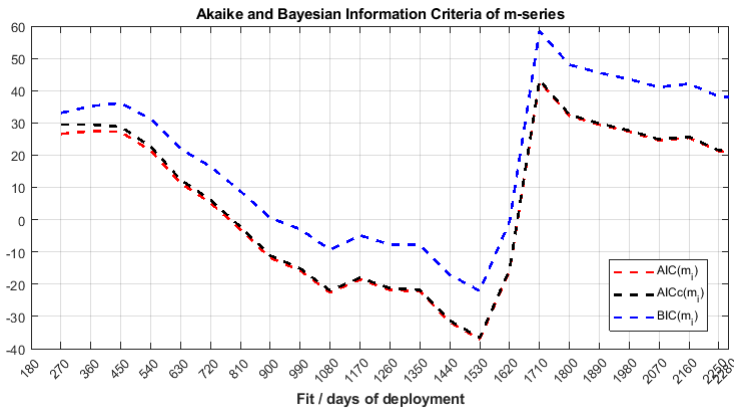

Figure 3: AIC (red dotted line), $\mathrm{AIC_C}$ (black dotted line) and BIC (blue line) of binary flag 7 of fit 5904179. Here, it is clearly visible that the $\mathrm{AIC_C}$ converges towards the AIC for longer deployment times.

In theory, AIC and $\mathrm{AIC_C}$ may slightly tend to over-fitting while BIC is usually referred to as a more conservative, rather under-fitting model (Dziak et al. 2020 [5]). However, the differences between AIC and BIC were relatively small. Instead, the second-order punishment incorporated in $\mathrm{AIC_C}$ proved to be necessary, as $\mathrm{AIC_C}$ and AIC differed significantly for the first 450 to 630 days. The choice of the binary flag

by use of the minimum AIC_C gave a valuable fit decision that could be considered rather conservative, especially for the first two years of the deployment.

2.3 Handling of Large Data Gaps

The MBARI fleet contained several floats that exhibited large data gaps, e.g. due to in-air measurements being rendered impossible by the presence of sea ice in the Southern Ocean. Thus, larger data gaps and incomplete seasonal cycles had to be accommodated for possible fit decisions. Especially the temperature drift was overestimated if in-air measurements were available only for one, short season (figure 4).

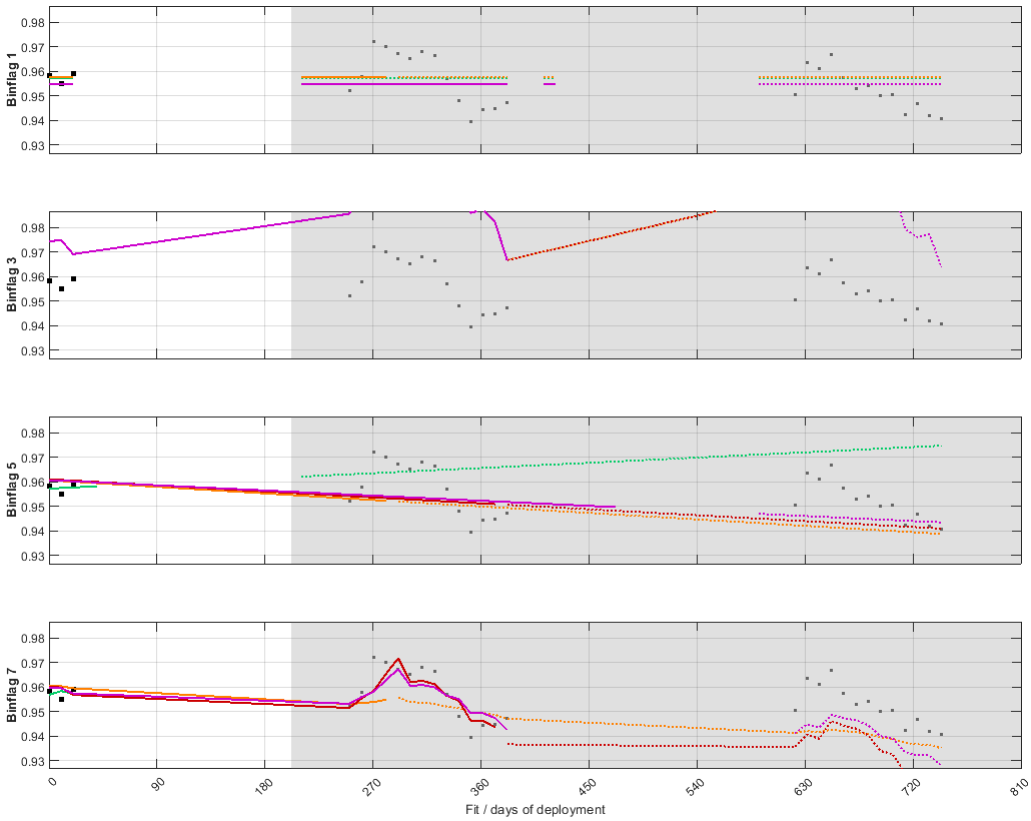

Figure 4: m of float 5905639 for binary flag 1 (upper graph), 3 (second graph), 5 (third graph) and 7 (at bottom). Drift correction was done for the first 201 days (white area, green fit), 291 days (orange fit), 381 days (red fit) and 561 days (violet fit). Binary flag 3 is not suited at all, as it overestimates the temperature drift.

The existing rules (a binary flag minor or equal to 1 in the first year, use of the fit which gives the minimum AIC_C) were expanded by rules that prohibited the use of a, $b_{\text{in situ}}$ or both for data gaps of different lengths. These rules were tested for 10 to 15 randomly chosen floats each and eventually refined to be tested again. After several testing cycles, the rule of prohibiting consideration of temperature correction in the first 450 days was applied for floats that showed data gaps of more than 120 days. Although

most cases were successfully identified by the use of AIC_C, over-fitting of a could not be prevented in every case. In contrast, $b_{\text{in situ}}$ was less sensitive to data gaps when the deployment time exceeded one year. Hence, it was excluded from the rule.

2.4 Avoidance of Frequent Fit Complexity Changes

In case two or more fits give similar AIC_C values over multiple time slices, the AIC_C criteria suggested frequent binary flag changes. These changes (referred to as "fit zapping") did not significantly influence the fit quality nor the numeric values of the corrected data as the fits performed at almost equal goodness. However, it contradicted the wanted fit stability: changes in the recommended binary flag should occur only at a significant increase in fit quality, as the less complex fits are generally less sensitive to future changes. Also, some floats deployed for more than three years showed a decrease in complexity after two to three years. Hence, it can be assumed that the AIC_C criterion leads to slight over-fitting during that time. A rule for fit changes therefore had to be introduced.

To estimate the relative plausibility of the different drift corrections j to the respective correction with the minimum information criterion value, the relative likelihood was calculated (16) (Burnham and Anderson 2002 [4]).

$$\mathcal{L}_{\text{rel.}}(t_i) \propto \exp\left(\frac{\text{AIC}_{\text{min}}(t_i) - \text{AIC}_j(t_i)}{2}\right) \tag{16}$$

Fit complexity was ranked in the order of the binary flags $1 < 5 < 3 < 7$, as a did not converge as fast as $b_{\text{in situ}}$ towards a final value (see section 2.1). Then, equation (16) was used to estimate the advantage of a binary flag change, as this does not only permit to rank but also to quantify the plausibility of the different binary flags (Burnham and Anderson 2002 [4]). For a probability p in %, indicating the possibility of a fit to be of superior goodness than the fit chosen by AIC_C, the required AIC_C difference could be calculated.

$$\Delta\text{AIC}_C(p) = 2 \cdot \ln\left(\frac{p}{100}\right) \tag{17}$$

Thus, for AIC_C of fits that did not differ by more than the chosen $\Delta\text{AIC}_C(p)$ of the minimum AIC_C value, the least complex fit was chosen. Different p were tested and this way, the number of overall increases and decreases in complexity of the entire fleet were compared. Also, for every p, 10 to 15 randomly chosen floats were investigated manually to see whether the rule did prohibit mandatory binary flag changes. As the fundamental idea is to choose a rather conservative than overly responsive fit guidance system, $p = 10\,\%$ proved to decrease the total number of complexity decreases after three years (see figure 5 for an example). Unfortunately, no efficient reduction in the absolute number of overall binary flag changes could be observed. However, the mean stability of the entire fleet could be increased, especially in the first years of deployment, thus reducing the frequency at which the fit decision would have to be reassessed. In summary, consideration of the relative likelihood thus increased the stability of the fit decision.

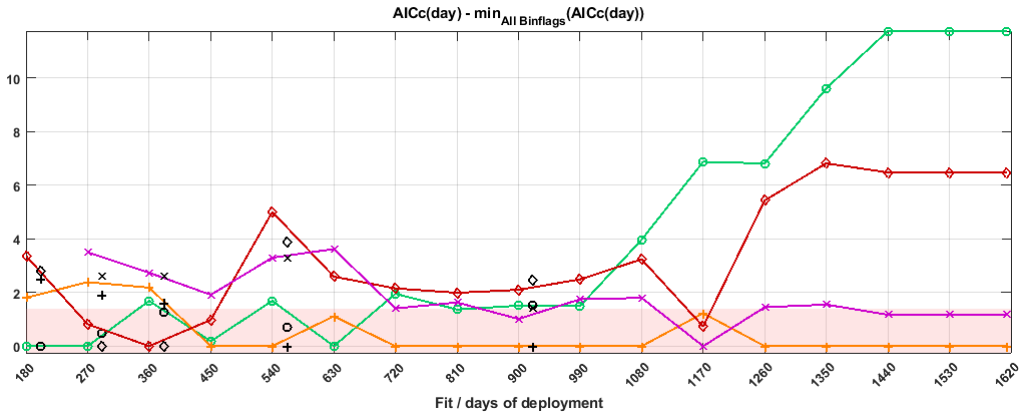

Figure 5: Relative difference of AIC_C of binary flags 1 (green, circle), 3 (orange, plus), 5 (red, diamond) and 7 (violet, cross) in 90 days steps over the deployment time. The interval of $p = 10$ % is indicated by the colored area.

2.5 Expanding Decision Guidance on Data Series $<$ 365 Days

As the previous analyses were conducted for time series of more than 365 days, an oxygen drift correction guidance for the first year was also required. For such time series, drift corrections are very sensible for even small deviations of $pO_{2,obs.}$ from $pO_{2,air}$, as the number of air measurements is small. Seasonality is rapidly translated into an overestimated temperature or temporal *in situ* drift. Hence, use of drifts of binary flags > 1 are not recommended for short deployments (see section 2.1). Fits of higher binary flags with more degrees of freedom mainly were not possible during the first 180 days, since the few in-air measurements available did not permit a complex fit.

The AIC_C rules defined in sections 2.2 and 2.4 were applied on all floats for the first 365 days. A major problem for AIC_C consideration was that during the first weeks of deployment, the number of fitting parameters was almost of the same magnitude as the number of in-air measurements, i.e. $4 \cdot k > n - 1$. In these cases, the respective AIC_C gave an empty value (see section 2.2). Hence, for such cases, the easiest case (binary flag $= 0$) has been chosen.

3 Final Oxygen Drift Correction Guidance

3.1 Final Decision Path

The proposed final decision tree can be divided into two steps: first, the possible fitting parameters are identified by investigation of the deployment duration, and the frequency of in-air measurements, i.e. the maximum length between two subsequent in-air measurements is determined (referred to as ”data gaps”). Therefore, the underlying rules are presented in figure 6. Then, the goodness of the possible fits for a float at a given time is evaluated by the following steps:

1. The fit parameters of all possible oxygen drift corrections are calculated.
2. The AIC_C values of these fits are calculated.

3. The minimum AIC_C is determined.
4. The least complex fit that lays within a ΔAIC_C of $p = 10\,\%$ according to (17) is chosen. Complexity of the fits in ascending order is 0 < 1 < 5 < 3 < 7.

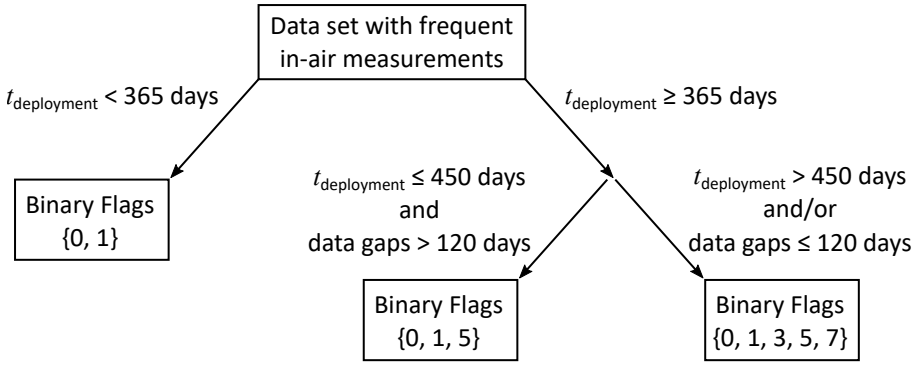

Figure 6: Proposed rules for the identification of the set of possible oxygen drift correction fits.

3.2 Fit Stability

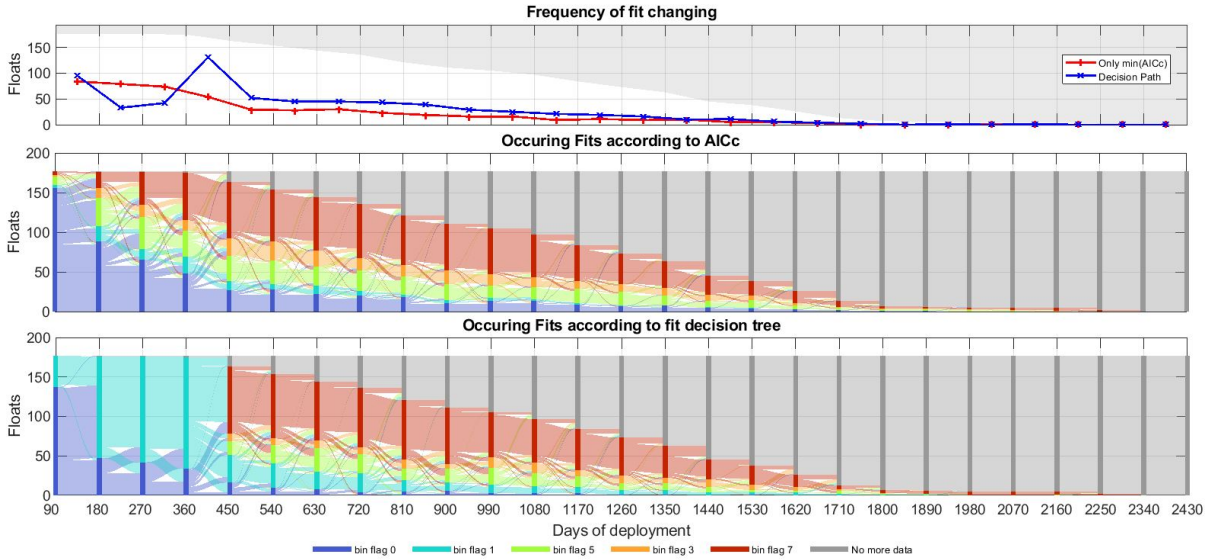

Figure 7: Binary flag changes of the entire fleet for the ”pure” AIC_C decision basing only on the minimum AIC_C value (centered plot, red line in upper plot) and fit rules (graph at bottom, blue line in upper plot) when the first 360 days are set as binary flag 1. Floats that keep their binary flag may still change their parameter values in the 90-day steps.

No key criteria could be identified to indicate either the frequency of re-evaluation of the chosen oxygen drift correction or the frequency of revising its parameters. All parameters that might give such

information were calculated by use of ”future” in-air measurements (time interval $[t_i;t_n]$, see section 2.1). However, in general, it can be said that the choice of a binary flag remained stable for more than 360 days on average. Also, binary flag changes became less frequent in the first 365 days (figure 7). Parameter changes decreased in ratio to the absolute value with increased deployment time. Thus, we recommend revising the oxygen drift correction at least once every year. By timing the reevaluation with the deployment time, a semi-annual or annual revision could be performed.

3.3 Influence of Batch-Calibration

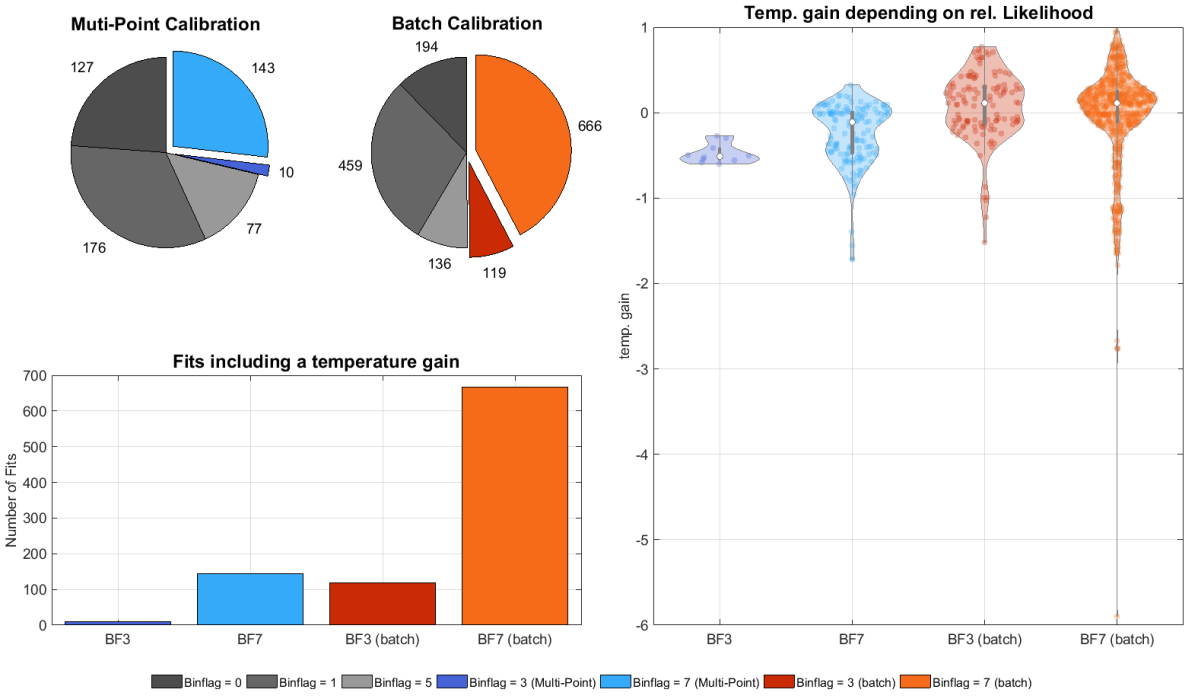

Figure 8: Comparison of the drift correction of multi-point calibrated floats (blue) and batch-calibrated floats (orange/red) for binary flags 3 (respective darker color) and 7 (respective lighter color), since these binary flags consider a. On the left, the absolute and relative number of such fits is given and on the right, the range of a is shown.

Oxygen sensors that had received only foil batch calibration generally exhibited a more prominent temperature dependency as those that had received the full recommended individual multi-point calibration (figure 8). Of all fits of batch-calibrated floats performed every 90 days since the beginning of the respective deployment, 8 % required a simple temperature drift correction (binary flag 3) and 42 % required a temperature and *in situ* drift correction (binary flag 7). These numbers were significantly smaller for the 74 multi-point calibrated floats (2 % and 27 %, respectively). Also, while the mean temperature drift of the individually calibrated floats was slightly negative (-0.50 °C^{-1} for binary flag 3 and -0.11 °C^{-1} for binary flag 7), the average temperature drift of batch-calibrated floats was slightly positive (0.11 °C^{-1} for binary flag 3, 0.11 °C^{-1} for binary flag 7). The absolute range of recommended temperature drift was significantly larger for batch calibrated floats than for multi-point calibrated floats.

3.4 Conclusions

The initial gain change correction is always required and was therefore defined as mandatory in all correction options. The other three corrections do not occur significantly in all floats and also depend on deployment duration. As general rule, we tried to keep the number of corrections as small as possible to avoid over-fitting and assure optimal fit stability. This was guided by a decision tree supported by information criteria. In the analyzed fleet of 178 floats, the carry-over correction was found necessary in quite a few cases starting at 180 days after deployment. The other two corrections were found to be often unstable and affected by seasonality during the first year and were therefore only allowed after 365 days. The need for a temperature correction was significant prevalent for oxygen optode that had only received a foil batch calibration instead of the recommended individual sensor calibration. We also found that data gaps in the in-air measurements of more than 120 days tended to effect the temperature correction negatively, which is why allowed this correction only after 450 days in these cases. Beyond 450 days of float lifetime all corrections we allowed. By considering fit stability over time we defined a preferred sequence of increasing fit complexity as guided by the chosen information criterion. It is recommended to repeat the complete delayed mode correction at least once every year during the float's lifetime.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, dec 1974.

[2] H. C. Bittig and A. Körtzinger. Tackling oxygen optode drift: Near-surface and in-air oxygen optode measurements on a float provide an accurate in situ reference. *J. Atmos. Ocean. Technol.*, 32(8):1536–1543, 2015.

[3] H. C. Bittig, A. Körtzinger, C. Neill, E. van Ooijen, J. N. Plant, J. Hahn, K. S. Johnson, B. Yang, and S. R. Emerson. Oxygen optode sensors: Principle, characterization, calibration, and application in the ocean. *Front. Mar. Sci.*, 4(JAN):1–25, 2018.

[4] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Interference - A Practial Informaiton-Theoretic Approach*. New York, USA, second ed. edition, 2002.

[5] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin. Sensitivity and specificity of information criteria. *Brief. Bioinform.*, 21(2):553–565, 2020.

[6] C. M. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[7] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, Roy Jenne, and D. Joseph. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.*, 77(3):437–471, mar 1996.

[8] A. Körtzinger, J. Schimanski, and U. Send. High Quality Oxygen Measurements from Profiling Floats: A Promising New Technique. *J. Atmos. Ocean. Technol.*, 22(3):302–308, 2005.

[9] T. L. Maurer, J. N. Plant, and K. S. Johnson. Delayed-Mode Quality Control of Oxygen, Nitrate, and pH Data on SOCCOM Biogeochemical Profiling Floats. *Front. Mar. Sci.*, 8(August):1–20, aug 2021.

[10] W. B. Owens and A. P. S. Wong. An improved calibration method for the drift of the conductivity sensor on autonomous CTD profiling floats by T–S climatology. *Deep Sea Res., Part I*, 56(3):450–457, 2009.

[11] G. Schwarz. Estimating the Dimension of a Model. *Ann. Stat.*, 6(2):461–464, mar 1978.