Supplementary Information: Quantifying observational errors in Biogeochemical-Argo oxygen, nitrate and chlorophyll *a* concentrations

A. Mignot, F. D'Ortenzio, V. Taillandier , G. Cossarini and S. Salon

1. Triple Collocation Analysis

In this section, we review the triple collocation analysis that estimates the RMSE, gain and offset (multiplicative and additive biases) of three collocated data sets from the covariance between the datasets. The full description of the method as well as a detailed discussion of the underlying assumptions can be found in Gruber et al. (2016).

The TC analysis assumes an affine error model relating three collocated data sets to the same unknown true value. The affine error model has the following form:

$$d_i = \alpha_i + \beta_i t + \varepsilon_i, \quad (1)$$

where d_i ($i \in [x, y, z]$) are the three collocated data sets, t is the unknown true value, β_i and α_i are the gain and offset and ε_i is the random error. The covariances between the data sets are given by

$$\sigma_{ij} = \beta_i \beta_j \sigma_t^2 + \beta_j \sigma_{t\varepsilon_i} + \beta_i \sigma_{t\varepsilon_j} + \sigma_{\varepsilon_i \varepsilon_j}. \tag{2}$$

Assuming that the random errors are uncorrelated with t ($\sigma_{t\varepsilon_i} = 0$), and which each other ($\sigma_{\varepsilon_i \varepsilon_j} = 0, for\ i \neq j$), Eq. (2) reduces to

$$\sigma_{ij} = \begin{cases} \sigma_i^2 = \beta_i^2 \sigma_t^2 + \sigma_{\varepsilon_i}^2, & for\ i = j \\ \sigma_{ij} = \beta_i \beta_j \sigma_t^2, & for\ i \neq j \end{cases}. \tag{3}$$

Under these assumptions, the random error variance of each data set, $\sigma_{\varepsilon_i}^2$, is equal to the variance of the data set, σ_i^2, minus the term $\beta_i^2 \sigma_t^2$. An expression for $\beta_i^2 \sigma_t^2$ can be obtained by combining the covariance between the data sets:

$$\begin{aligned} \beta_x^2 \sigma_t^2 &= \frac{\sigma_{xy} \sigma_{xz}}{\sigma_{yz}} \\ \beta_y^2 \sigma_t^2 &= \frac{\sigma_{yx} \sigma_{yz}}{\sigma_{xz}} \\ \beta_z^2 \sigma_t^2 &= \frac{\sigma_{zx} \sigma_{zy}}{\sigma_{xy}} \end{aligned} \tag{4}$$

The random error standard deviation (or equivalently the root-mean-squared random error, hereinafter denoted the root-mean-squared error, RMSE) , σ_{ε_i} ($i \in [x, y, z]$), for each data set can now be obtained by substituting Eq. (4) into Eq. (3) and taking the square-root:

$$\begin{aligned} \sigma_{\varepsilon_x} &= \sqrt{\sigma_x^2 - \frac{\sigma_{xy} \sigma_{xz}}{\sigma_{yz}}} \\ \sigma_{\varepsilon_y} &= \sqrt{\sigma_y^2 - \frac{\sigma_{yx} \sigma_{yz}}{\sigma_{xz}}} \\ \sigma_{\varepsilon_z} &= \sqrt{\sigma_z^2 - \frac{\sigma_{zx} \sigma_{zy}}{\sigma_{xy}}} \end{aligned} \tag{5}$$

The TC analysis also provides the gain and offset (Stoffelen, 1998; Yilmaz & Crow, 2013), on the assumption that one data set is perfectly calibrated. Assuming that x is the reference data set, so that $\alpha_x = 0$ and $\beta_x = 1$, we obtain two expressions for β_y and β_z:

$$\beta_y = \frac{\sigma_{yz}}{\sigma_{xz}} \qquad (6)$$
$$\beta_z = \frac{\sigma_{zy}}{\sigma_{xy}}$$

Taking the sample mean of Eq. (1) yields two expressions for α_y and α_z:

$$\alpha_y = \overline{d_y} - \frac{\sigma_{yz}}{\sigma_{xz}}\overline{d_x} \qquad (7)$$
$$\alpha_z = \overline{d_z} - \frac{\sigma_{zy}}{\sigma_{xy}}\overline{d_x}$$

Note that the RMSE estimates are calculated in their own data space, which means that they cannot be compared with each other. However, The RMSEs can be rescaled within the reference data space using the gain determined in Eqs. (6) following:

$$\sigma_{\varepsilon_y^x} = \frac{\sigma_{\varepsilon_y}}{\beta_y} \qquad (8)$$
$$\sigma_{\varepsilon_z^x} = \frac{\sigma_{\varepsilon_z}}{\beta_z}$$

This rescaling allows direct comparison between the three RMSEs.

2. Quality-control procedures for BGC-Argo O_2, NO_3^- and Chl*a*

2.1. QC procedure for O_2

Vertical profiles of O_2 acquired by BGC-Argo floats are affected by two major flaws. First, the deepest 50 m of the profiles are systematically depressed and show a characteristic "hook" at the base of the profiles. The cause of these hooks is still being investigated, with optode response time being widely suspected. Consequently, the deepest 50 m of the profiles were discarded. Second, estimates O_2 need to be gain-corrected due to imperfect factory calibration (Johnson et al., 2015; Körtzinger et al., 2005). Following Takeshita (2013) and (Johnson et al., 2015), for each float profiles, gain values were estimated by comparing float surface

percent oxygen saturation with monthly climatological values from the World Ocean Atlas Climatology 2013 (Garcia et al., 2014). Then a single gain, taking as the median of all gain values, was used to correct the oxygen data for each float.

2.2. QC procedure for NO_3^-

NO_3^- data estimated from the Satlantic SUNA sensor suffer from an additive bias which is constant over the entire profile. The quality-control procedure adjusts the deepest 50 m of a vertical profile of NO_3^- to a climatological value estimated with a neural network. This procedure is based on the premise that deep nitrate concentrations have little spatial and temporal variability. The neural network (Sauzède et al., 2017) uses in situ temperature, salinity, pressure, the gain-corrected O_2, date, longitude and latitude to predict a climatological nitrate concentration.

2.3. QC procedure for Chl*a*

The adjustment procedure for the raw Chl*a* determined from fluorescence corrects from imperfect initial calibrations coefficients and a photo-physiological process that systematically bias the surface fluorescence values during daytime. First, each fluorescence profile is set to 0 at the bottom by subtracting the median of the 5 deepest points from the profile to account that deep fluorescence values are non-null in the absence of algal fluorescence. Second, Roesler et al. (2017) recently found that Chl*a* acquired by WET Labs ECO-series chlorophyll *a* fluorometers have a global mean systematic multiplicative bias of 2. Consistently, the raw Chl*a* were divided by a factor of 2. Finally, non-photochemical quenching (NPQ) at high irradiance results in a decrease of the fluorescence to chlorophyll ratio when the phytoplankton cells are exposed to intense solar illumination in the upper layers of the ocean during daytime. As a result, surface Chl*a* collected with fluorometers implemented on BGC-Argo floats are expected to be systematically biased (depressed) when floats surfaces at noon. NPQ was corrected using the method of Xing et al. (2012).

2.4. A note on negative values in BGC-Argo O_2, NO_3^- and Chl*a* data sets

Following Johnson et al. (2017), negative values were retained in the float data sets to avoid impairing the triple collocation biases and RMSE estimates. Removing negative values

or setting them to zero will eventually impact the random error standard deviation and will create artificial systematic error in the data.

3. Features of the MedBFM model system

The MedBFM model system is part of the Mediterranean Sea Monitoring and Forecasting Centre (Med-MFC) of the Copernicus Marine Environment Monitoring Services (CMEMS), and weekly produces analysis and forecasts of the Mediterranean Sea biogeochemistry. The MedBFM model system is based on the Biogeochemical Flux Model (BFM, Cossarini et al., 2015; Lazzari et al., 2012, 2016)**.** The BFM biogeochemical model includes four phytoplankton functional groups (diatom, flagellates, pico-phytoplankton and dinoflagellates) and five heterotrophic functional groups (carnivorous and omnivorous meso-zooplankton, bacteria, heterotrophic nanoflagellates and micro-zooplankton). The chlorophyll and carbon dynamics are based on the parameterization of chlorophyll synthesis proposed by Geider et al. (1997). The model resolves the cycling of dissolved oxygen in the water dissolved phase as well as carbon, phosphorus, nitrogen, silicon through dissolved inorganic, living organic, and non-living organic phases. Nitrate and ammonia are considered for the dissolved inorganic nitrogen. The MedBFM model system includes a variational 3DVAR-BIO scheme that assimilate remote sensed surface chlorophyll (Teruzzi et al., 2014).

The biogeochemical and biological tracers are transported and mixed with the transport OGSTM model based on the OPA system (Foujols et al., 2000). The physical dynamics that drive the transport biogeochemical tracers and their dynamics are pre-computed by NEMO Ocean General Circulation Model in its implementation of the Mediterranean Sea (Oddo et al., 2009), which is also included in the Med-MFC. The Mediterranean Sea NEMO implementation assimilates through the OceanVar variational scheme (Dobricic & Pinardi, 2008; Storto et al., 2014) satellite-observed Sea Level Anomaly, and in-situ vertical temperature and salinity profiles from XBTs, Argo floats and Gliders. The off-line coupled hydrodynamics and biogeochemical model is implemented for the Mediterranean Sea domain based on a 1/16 ˚ horizontal resolution and 72 vertical z-levels unevenly spaced (about 1.5 meter at the first layer and about 8 m at 100 m depth).

The units of the model O_2 and NO_3^- data sets were transformed from µmol l^{-1} into µmol kg^{-1} with the 3D fields of temperature and salinity that were used to force the biogeochemical

model. This transformation guarantees units coherency between the float, ship and model data sets of O_2 and NO_3^-.

Supplementary Table 1. Minimum, maximum and range of O_2, NO_3^- and Chl*a* values observed by the BGC-Argo floats.

variable	minimum	maximum	range
O_2	173.3 µmol kg^{-1}	257.4 µmol kg^{-1}	84.1 µmol kg^{-1}
NO_3^-	-0.47 µmol kg^{-1}	7.42 µmol kg^{-1}	7.89 µmol kg^{-1}
Chl*a*	-0.01 mg m^{-3}	0.55 mg m^{-3}	0.56 mg m^{-3}