Supporting Information for "Water Mass and Biogeochemical Variability in the Kerguelen Sector of the Southern Ocean: A Machine Learning Approach for a Mixing Hotspot"

Isabella Rosso[1], Matthew R. Mazloff[1], Lynne D. Talley[1], Sarah G. Purkey[1], Natalie M. Freeman[1], and Guillaume Maze[2]

Contents of this file

1. Text S1
2. Text S2
3. Text S3
4. Text S4

Corresponding author: I. Rosso, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA (irosso@ucsd.edu)

[1]Scripps Institution of Oceanography,
University of California San Diego, La Jolla,
CA, USA.

[2]Ifremer, University of Brest, CNRS,
IRD, Laboratoire d'Océanographie Physique
et Spatiale, IUEM, 29280, Plouzané, France

5. Table S1

6. Figures S2 to S4

Introduction This supplementary material contains the table discussed in Section 2.1 and the figures presented in Section 2.2.

Text S1. List of biogeochemical Argo floats. Table S1 lists the biogeochemical-(BGC-) Argo floats used in the analysis, together with the research cruise they have been deployed from. Cruise data sets are available at the Carbon and Climate Hydrographic Data Office (CCHDO) website (`http://cchdo.ucsd.edu`), and can be found by their expocode indicated in table.

Text S2. Dimensionality reduction.

We applied the Principal Component Analysis (PCA) to the 300-900 m core and BGC Argo θ and SA profiles. The first two components of the PCA, computed for the training set (i.e., 89% of the dataset), are shown in Fig. S2. We use these first two components only, because they explain $\sim$99% of the property variance for both temperature and salinity. The Gaussian Mixed Model algorithm is then applied to the reduced profiles, as discussed in Section 2.2.1.

Text S3. Model selection. The Bayesian Information Criterion [BIC, *Schwarz et al.*, 1978; *Konishi and Kitagawa*, 2008] is a common method used to evaluate the model selection and the optimal number of clusters [e.g., *Maze et al.*, 2017; *Jones et al.*, 2018], and is based on the model's posterior probability (Fig. S3; see *Maze et al.* [2017] for details). The criterion is computed for 10 different sets of independent profiles, each set assuring the information to be non-redundant. Each set is obtained by selecting 1 random profile in every $2.5° \times 2.5°$ box, finding a total of 2166 temperature and salinity profiles (i.e.,

$\sim$2.2% of the dataset). This number is comparable to the number of independent profiles found by using the decorrelation scales from *Ninove et al.* [2016]: e.g., the decorrelation scale is about 150×200 km^2 for salinity and 150×150 km^2 for temperature; assuming an approximated value of 100 km per degree of latitude and longitude, we can calculate the total area of the South Indian Ocean between 0°–180° and 70°S–30°S, as 180° × 100 km×40° × 100 km= 72×10^6 km^2. Thus, we find 3200 independent temperature profiles and 2400 salinity profiles, which are comparable to the number of profiles we randomly extract using 2.5° × 2.5° boxes (i.e., 2166).

In general, the minimum number of BIC indicates the best model [*Schwarz et al.*, 1978; *Konishi and Kitagawa*, 2008]. However, we do not find a clear minimum, but, as in *Jones et al.* [2018], a range of optimal values of clusters k (between 9 and 15), which have all been tested. This is due to the fact that the problem we are analysing is a continuous manifold, and cannot be separated into different regimes (Fig. S3). We select $k = 9$ among all the statistically equivalent possibilities, as this allowed for a meaningful separation of the profiles into the desired Southern Ocean regimes.

Text S4. Principal Components Analysis.

A way to visualize the information of the Gaussian distributions of (1) is to look at the covariance error ellipses (Fig. S4). The diagram in figure shows standardized temperature (ordinate axis) and salinity (abscissa) of the 2 PCAs, and 9 covariance error ellipses, centered in the cluster's centroid (i.e., μ_k of equation 1), and with axis oriented in the directions in which the variance vary the most (defined by the eigenvectors of the covariance matrix $Sigma_k$). The axis of the ellipses are scaled by a factor s which is

computed as the quantiles of the χ^2 distribution (figure shows the 95% confidence level, thus $s = 5.991$ for 2 degrees of freedom).

References

Jones, D. C., H. J. Holt, A. J. Meijers, and E. F. Shuckburgh (2018), Unsupervised clustering of Southern Ocean Argo float temperature profiles, *Journal of Geophysical Research: Oceans*.

Konishi, S., and G. Kitagawa (2008), *Bayesian Information Criteria*, pp. 211–237, Springer New York, New York, NY.

Maze, G., H. Mercier, R. Fablet, P. Tandeo, M. L. Radcenco, P. Lenca, C. Feucher, and C. Le Goff (2017), Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean, *Progress in Oceanography*, *151*, 275–292.

Ninove, F., P.-Y. L. Traon, E. Remy, and S. Guinehut (2016), Spatial scales of temperature and salinity variability estimated from Argo observations, *Ocean Science*, *12*(1), 1–7.

Schwarz, G., et al. (1978), Estimating the dimension of a model, *The annals of statistics*, *6*(2), 461–464.

Deployment Cruise (expocode)	Cruise Affiliation	Float UW/WMO ID	Float Type
A12 (06AQ20141202)	Global Ocean Ship-based Hydrographic Investigations Program (US GO–SHIP)		
		9313/5904474	Apex
		0508/5904476	Navis
		9260/5904473	Apex
		9096/5904469	Apex
I08S (33RR20160208)	GO–SHIP		
		0510/5904686	Navis
		9602/5904684	Apex
		9637/5904682	Apex
		9650/5904683	Apex
		9600/5904688	Apex
SR03 (096U20180111)	GO–SHIP		
		0688/5904846	Navis
		12736/5905376	Apex
		12779/5905371	Apex
		12769/5905373	Apex
		12782/5905374	Apex
		12709/5905375	Apex
		12702/5905379	Apex
		12741/5905377	Apex
		12748/5905378	Apex
		12370/5905103	Apex
Antarctic Circumnavigation Expedition (ACE; RUB320161220)	Swiss Polar Institute		
		0691/5905072	Navis
		0692/5905073	Navis
		12558/5905069	Apex
		12537/5905070	Apex
K-Axis (09AR20160111)	Australian Antarctic Division and Antarctic Climate and Ecosystems Cooperative Research Centre		
		0506/5904670	Navis
		0507/5904671	Navis

Heard Earth–Ocean–Biosphere Interactions (HEOBI; 096U20160108)	Institute for Marine and Antarctic Studies, University of Tasmania (IMAS, UTas)	9749/5904675	Apex
		9645/5904676	Apex
		9757/5904679	Apex
Southern Ocean Expedition 10 (SOE10)	National Centre for Antarctic and Ocean Research	0690/5905071	Navis
		12734/5905370	Apex
		12755/5905367	Apex
		12730/5905369	Apex
		12781/5905368	Apex
		12757/5905366	Apex

Table 1: List of floats used in this study, including deployment cruise and cruise affiliation information. The deployment cruises are listed together with their expocode, the identifier used in the Carbon and Climate Hydrographic Data Office (CCHDO) website (`http://cchdo.ucsd.edu`). The floats are identified by their University of Washington (UW) and WMO IDs.

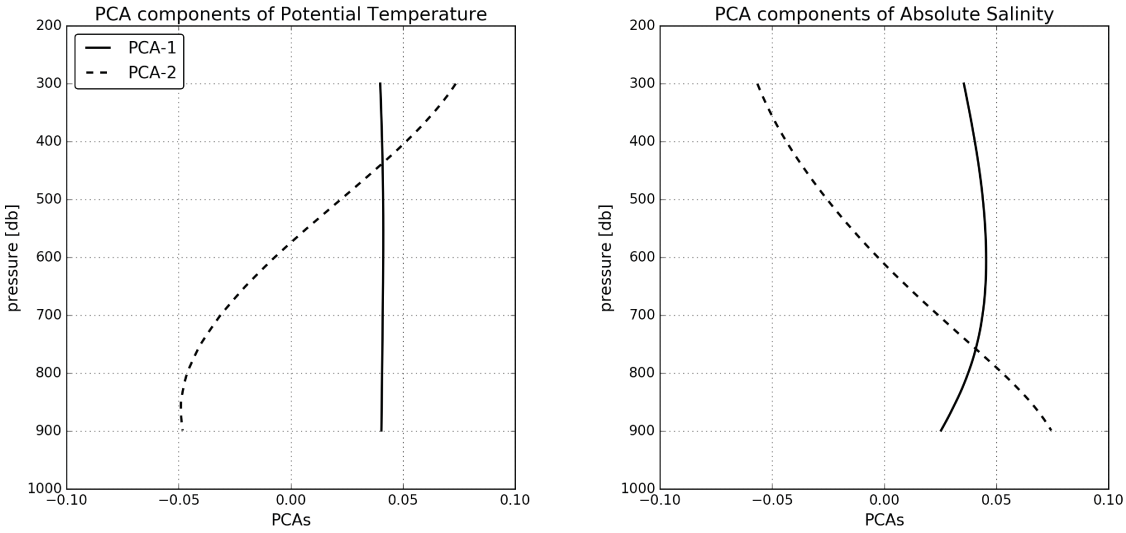

Figure S2. First two principal components for (left) potential temperature and (right) absolute salinity, using the training set of the Argo vertical profiles between 300 m and 900 m. Solid (dashed) lines are the first (second) PCAs.

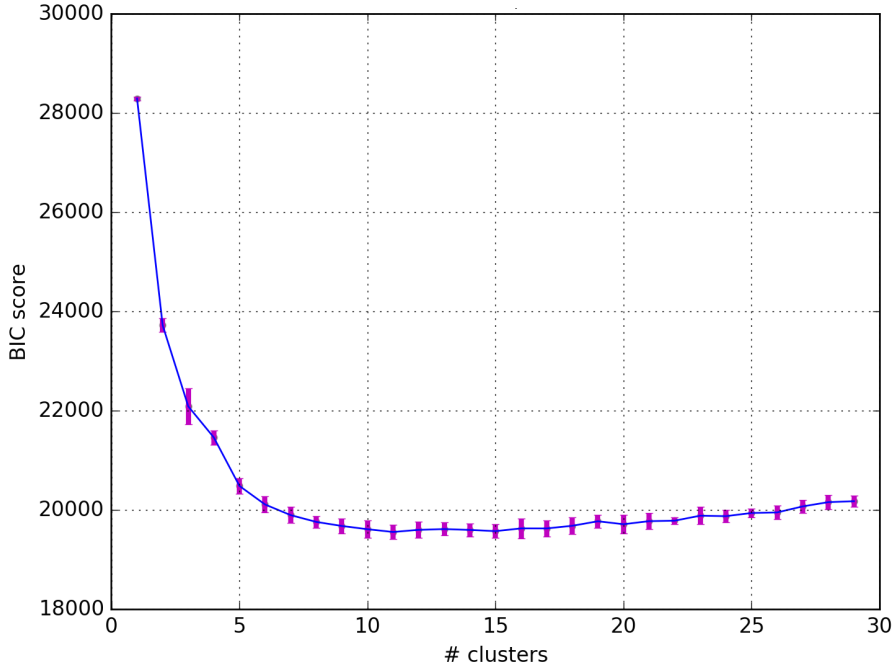

Figure S3. Mean of the Bayesian Information Criterion computed for 10 sets of randomly selected profiles of Argo floats' temperature and salinity, ranging the clusters number from 1 to 30. The error bars are standard deviation computed over the 10 sets.

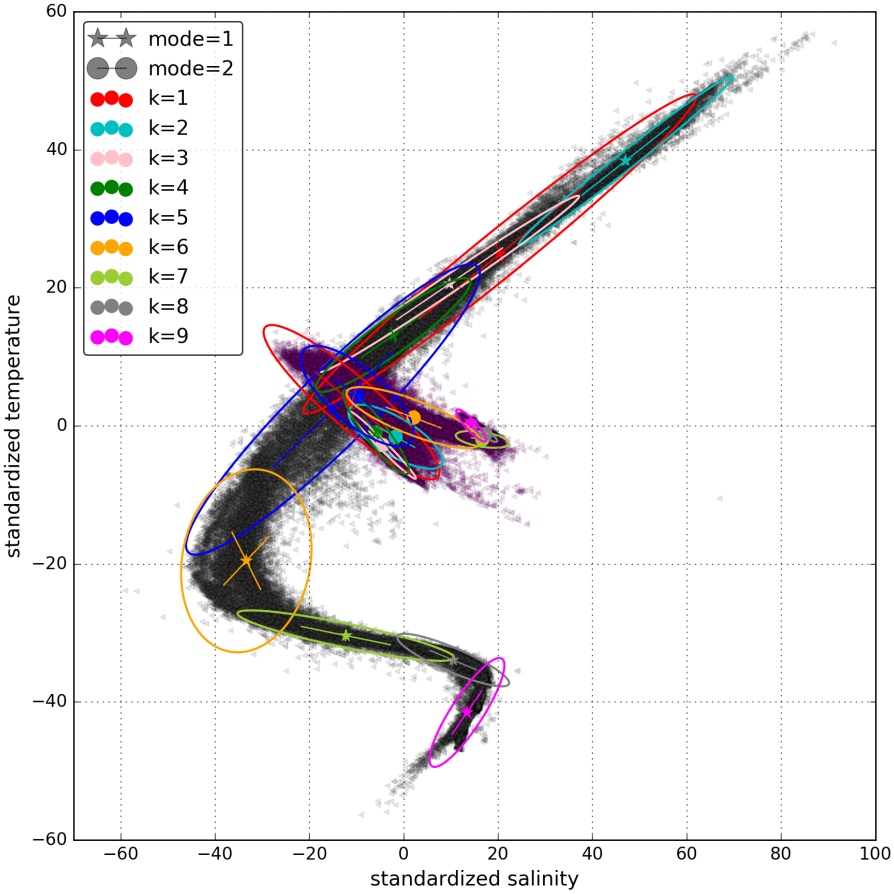

Figure S4. Diagram of the standardized temperature and salinity, decomposed into the first (black triangles) and second (magenta) principal component analysis modes, for the Argo floats training set. The covariance error ellipses are centered in the cluster's centroids, with stars (circular) markers for the first (second) principal component analysis mode.