PROFILE CLASSIFICATION MODELS

BY
G. MAZE[1], H. MERCIER[2], C. CABANES[3]

INTRODUCTION

Ocean dynamics and the induced 3-dimensional structure and variability are so complex that it is very difficult to develop objective and efficient diagnostics of horizontally and vertically coherent oceanic patterns. However, identifying such patterns is crucial to the understanding of interior mechanisms as, for instance, the integrand giving rise to Global Ocean Indicators (e.g. heat content and sea level rise). We believe that, by using state of the art machine learning algorithms and by building on the increasing availability of ever-larger in situ and numerical model datasets, we can address this challenge in a way that was simply not possible a few years ago. This letter aims to present the principles and first results of an approach introduced by Maze et al (2017) based on what we coined a «Profile Classification Model» or PCM that focuses on vertically coherent patterns and their spatial distribution.

The goal of a PCM is to automatically extract out of a collection of profiles a synthetic statistical description, i.e. a model, of typical profiles present in the collection. Once a PCM is built, i.e. trained, one can use this model to determine, with probabilities, the typical class any new profile most resembles. Therefore, it becomes possible to assign to a given typical class of profiles appropriate parameters for a specific diagnostic (e.g.: a finely tuned density threshold for mixed layer depth computation, a depth range for a pycnocline or mode water identification), or simply to use the PCM distributions to analyze the climatology or variability of the coherent patterns in space and/or time.

Hereafter, we present the PCM method, its first results and four possible applications for a variety of ocean analysis problems.

[1] **gmaze@ifremer.fr, Ifremer, Univ. Brest, CNRS, IRD, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, F-29280,** Plouzané, France.

[2] **Herle.Mercier@ifremer.fr, Ifremer, Univ. Brest, CNRS, IRD, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, F-29280** Plouzané, France.

[3] **Cecile.Cabanes@ifremer.fr.fr, Ifremer, Univ. Brest, CNRS, IRD, Laboratoire d'Océanographie Physique et Spatiale (LOPS), IUEM, F-29280** Plouzané, France

METHOD

The obvious difficulty is in the construction of the PCM. The goal is to automatically determine typical profiles. This can be achieved from one simple idea: if a profile is typical, then it will be redundant (although with small variations) in a sufficiently large and heterogeneous collection of profiles. Hence, a typical profile will have a high probability of occurrence and creates a peak in the probability density function (PDF) of the collection of profiles. We thus can determine typical profiles by creating a model for the peaks of the collection PDF. To do so, we used Gaussian Mixture Models that belong to the class of unsupervised classification methods (Bishop et al 2006). It determines the most likely decomposition of a PDF into a finite sum of Gaussian modes. Each Gaussian mode property provides a model, i.e. a description, for a typical profile, including a mean profile (the center of the mode) and a spread/pattern (the squared covariance matrix of the mode). One should note that a PCM based on a Gaussian Mixture Model identifies vertically coherent patterns because the multi-dimensional Gaussian mode covariance matrix has no reason to be diagonal, which allows for complex vertical relations.

RESULTS

We trained a PCM, based on a Gaussian mixture model, with about 100,000 Argo temperature profiles located in the North Atlantic Ocean. Using a series of subset of the collection with uniform space/time distribution, we determined both objectively and through trial/error that 8 typical profiles characterize the interior large scale temperature structure of the North Atlantic between the surface and 1400m. These typical profiles, together with their spread, are shown in Figure 1. To follow the data mining vocabulary, we may also refer to typical profiles as class of profiles.

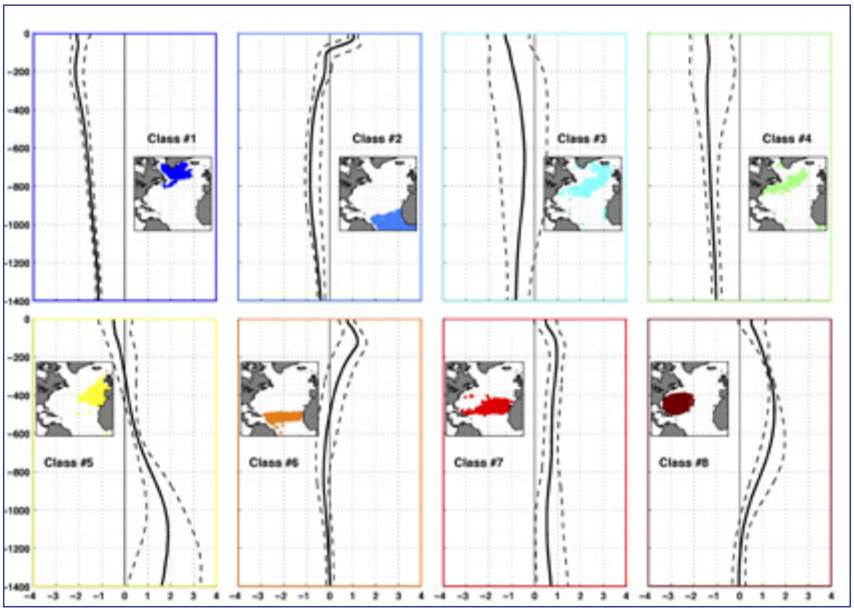

FIGURE 1

The 8 typical temperature profiles of the North Atlantic. Temperature profiles are centered/standardized at each depth level and black dashed lines indicate the 5%-95% spread of the class. Map insets show the location of profiles attributed to each class.

Two classes (#1/#4) show cold anomalies throughout the water column with amplitude decreasing with depth. One class (#3) has nearly zero anomalies, and a large spread throughout the water column. One class (#2) has warm anomalies near the surface (50m) and cold ones below 200m. The remaining four classes have warm anomalies throughout the water column, one without depth dependence (#7), the other three (#5/#6/#8) with clear maxima at different depths (1000m, 100m and 400m respectively).

Another key point of the Maze et al (2017) study is that the 8 typical profiles were identified without using the information of latitudes, longitudes and times of profile samplings. So, we furthermore investigated the locations in time and space of profile classes (note that to classify a profile, we compute the probabilities it has to be similar to each of the typical profiles and select the class maximizing these probabilities). On the one hand we found no correlation between the time of samplings and the classes (not shown). This means that whatever the season (the largest source of temporal variance in the dataset) the same collection of typical profiles characterizes the dataset. On the other hand, we found a key result in locating in space the class of profiles. Figure 1 insets show the location map of profiles attributed to each class. One can see that each class delineates a specific and physically coherent region of the ocean. This is a truly remarkable result because it demonstrates objectively that a given region corresponds to a unique vertical temperature pattern. In other word, the vertical stack of water masses and thermoclines is specific to a region of the North Atlantic Ocean.

A more detailed description of typical profiles, how they relates to known water masses and thermoclines and a sensitivity analysis can be found in Maze et al (2017).

APPLICATIONS

The PCM results briefly presented above pave the way for many possible applications in data analysis and physical studies. Below we briefly review four of these promising applications.

Study of a region with natural boundaries

Let's take the class #1 that delineates the North Atlantic subpolar gyre. We can apply a PCM to a gridded interpolation of Argo data and naturally delineate the subpolar region without using rectangular boxes, complex polygons or surface data from another, possibly incoherent, source. Here we used the Argo-based PCM to classify the ISAS13 time series of optimally interpolated Argo temperature data (Gaillard et al, 2016). Figure 2-A and B show the 2002-2015 grid point average and monthly variance of the local temperature profile probabilities classified in class #1. Map A clearly shows the natural contouring of the subpolar gyre, while map B indicates that the gyre variability is mostly located along its boundaries with a narrow band in the North West Corner Region and a wider band in the Iceland Basin in the North Atlantic Current region. Furthermore, using the PCM property that the sum of profile probabilities to belong to each of the 8 classes, namely), goes to one, we can decompose the local water column heat content into 8 fractions attributed to each class **c**:

Eq.(1)

$$OHC_z(\boldsymbol{c},\theta) = \iint_{x,y}\Big(p(c|x,y,t)\int_{z=0}^{z}\rho_0 C_p\theta(x,y,z,t)dz\Big)dxdy$$

without losing heat because

$\sum_{c=1}^{8} p(c|x,y,t)=1$, hence $\sum_{c=1}^{8} OHC_z(\boldsymbol{c},\theta) = 1.$

Figure 2-C shows in blue the detrended interannual time series of . We won't explain here the structure of the events shown in this time series but rather focus on its decomposition into the variability arising from local temperature variations vs. gyre horizontal extent. We can indeed approximate Eq.(1) for class #1 by the sum of two terms where either the class extent or the temperature are being set to their time average: Eq(2)

$$OHC_z(\bar{\boldsymbol{c}}\boldsymbol{=1},\theta) = \iint_{x,y}\Big(\overline{p(\boldsymbol{c=1}|x,y,t)}\int_{z=0}^{z}\rho_0 C_p\theta(x,y,z,t)dz\Big)dxdy$$

Eq(3)

$$OHC_z(\boldsymbol{c=1},\bar{\theta}) = \iint_{x,y}\Big(p(\boldsymbol{c=1}|x,y,t)\int_{z=0}^{z}\rho_0 C_p\overline{\theta(x,y,z,t)}dz\Big)dxdy$$

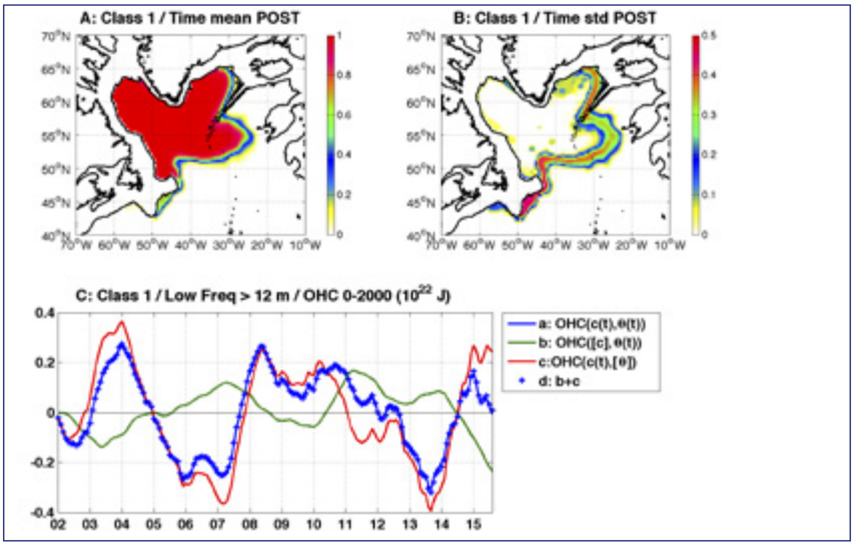

FIGURE 2

Plot A: 2002-2015 time mean probabilities of the subpolar class (#1 from Fig. 1). Plot B: 2002-2015 monthly variance of the subpolar class probabilities. Plot C: Detrended low-frequency variability of the 0-2000m OHC for the subpolar region with blue: total (Eq.1), red: due to class contour variations (Eq.2), green: due to class temperature variations (Eq.3).

Figure 2-C shows the subpolar gyre heat content variability driven by local temperature variations in green, Eq.(2), and by the gyre horizontal extent variations in red, Eq.(3). One can clearly see how the total heat content of the gyre, blue curve - Eq.(1), is driven by this later term while temperature variations appear to be anti-correlated. This means that when the gyre shrinks it also gets warmer and vice versa. We will show elsewhere how this is consistent with the diabatic and adiabatic atmospheric forcing patterns. One could furthermore show that the gyre extent defined through the class contour is consistent with the one diagnosed using the surface embedded within a fixed Sea Surface Height contour (not shown), demonstrating the relevance of the method presented here to delineate the gyre contours.

Structure of frontal regions

The PCM used in Maze et al (2017) is probabilistic, meaning that the transition between a class and another is not a step function but rather fuzzy, allowing for ambiguous profiles to be taken into account. A metric can be derived to interpret how robust is the classification of a profile. When mapped in space, robust classifications are found for profiles located in the core of the region they define (see their Fig.12). But one striking result is that highly robust classifications also appear to be located along frontal regions. This simply means that a PCM easily differentiates profiles from both sides of a front.

This is illustrated in Figure 3. We trained a 3-class PCM from temperature profiles of an eddy-resolving model simulation at 1/12° resolution in the Gulf Stream region (the DRAKKAR simulation referenced as NATL12-BAMT20, used by Maze et al, 2013 to study subtropical mode water formation). Figure 3-A shows the median and spread of class profiles. One can see how the PCM distinguishes the cold northern flank waters (blue profiles, class #a, without a clear vertical structure but the surface spread due to the seasonal cycle) from warm southern flank waters (orange profiles, class #c with almost no spread at 300m indicating the depth of the homogeneous Eighteen Degree Mode Water located above the permanent pycnocline with a larger spread). The remaining class (#b, in green) has a large spread almost throughout the water column. When mapped in space for the 5-days period centered on May 20th, 2003 (Figure 3-B), the class distribution is coherently revealing the horizontal distribution of the vertical structures identified by the PCM, i.e. the northern flank (class #a), the core of the front (class #b) and the southern flank (class #c) of the Gulf Stream. Figure 3-C and D with interior temperature and surface relative vorticity furthermore illustrates the accurate distinction being made by the PCM between these regions.

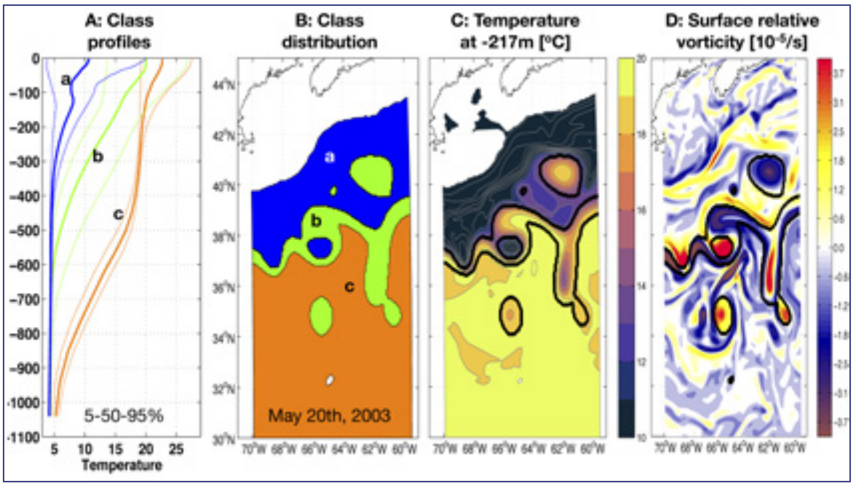

FIGURE 3

Demonstration with a 5-days averaged model output (1/12° resolution) that a PCM can distinguish the Gulf Stream front (class #b) from its flank water masses (classes #a and #c) only with ocean interior data. Plot A: 5-50-95% percentile class profiles. Plot B: Class attributed to profiles for May 20th, 2003. Plots C-D: interior temperature and surface relative vorticity superimposed with class contours (black) for the same date.

It is also of high interest to note that geographical incursions of a given class into the other coincide with meso-scale eddies. In fact, if one increases the number of classes, one could even distinguish cyclonic from anticyclonic eddies into separate classes (not shown). This PCM property will be exploited in the LEFE GMMC/IMAGO project «SOMOVAR» over the next 3 years.

Profile selection for QC in frontal regions

One can also make use of the frontal region PCM performance to improve the selection of reference profiles for Quality Control. This is illustrated in Figure 4. Again, let's take the Gulf Stream Extension region as an example. We trained a 3-class PCM from temperature profiles of the Argo reference database[3]. This is the reference database used in standard

[3] http://www.argodatamgt.org/Reference-data-base/Latest-Argo-Reference-DB

QC methods, such as OW (Cabanes et al, 2016). Figure 4-A shows the median and spread of class profiles and Figure 4-B shows the regional distribution of the class attributed to each profiles. Like in the previous case with the eddy resolving simulation, we again can distinguish the front from its flank waters. On the map we indicated the climatological location of the Gulf Stream core (black dashed line, determined from AVISO Sea Surface Height data as the latitude of the maximum zonal geostrophic velocity) to emphasize the appropriate results of the PCM.

Now imagine that we'd like to quality control a new Argo float set of profiles. A classic approach would be to take the reference collection and to compare float data with statistics from the reference. This is illustrated Figure 4-C and F for profiles #78 and #80 of the Argo float with WMO 4900136 for which the locations, and trajectory, are shown in Figure 4-E and H. From the reference database, we computed at each depth the distance weighted mean and standard deviation of temperature (for the same season as the profile to validate) using 300km and 150km decorrelation length scales in the zonal and meridional directions (results are qualitatively similar if one reduces these scales by a factor of 2). This standard reference envelop is shown in blue in Figure 4-C and F, while the float profiles are shown in red. For profile #78, data are out of the standard range from -100m to -600m depth. For profile #80, data are out of the standard range from -100m down to the bottom of the profile. Thus, using the standard reference envelop, these profiles would look suspicious and would create false alarm from automatic QC procedures.

The PCM method can, in this case, provides useful information to compare float data to the appropriate reference statistics. But first, let's examine the dynamical context of these profiles. In Figure 4-E and H we show the AVISO absolute dynamic ocean surface topography (based on all-satellites in delayed-time) for the same days as the profiles. On the

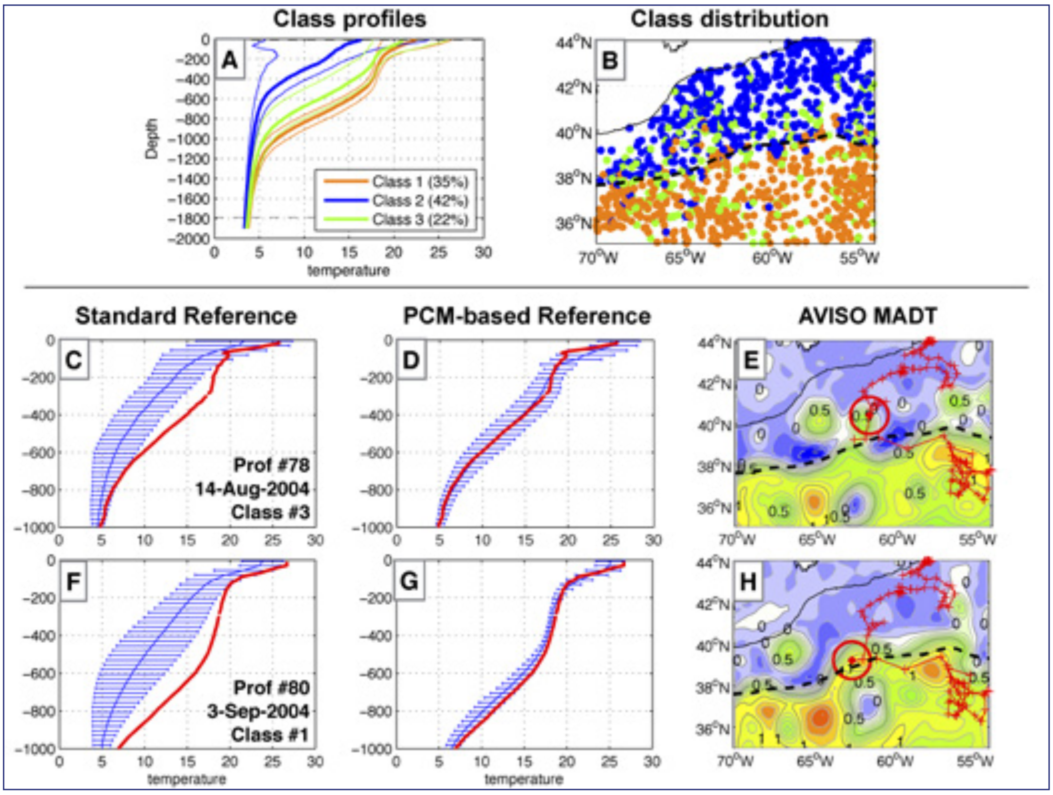

FIGURE 4

Argo-based PCM with 3 classes in the Gulf Stream Extension region. Plot A: 5-50-95% percentiles of the class temperature profiles. Plot B: Location of the profiles attributed to each class superimposed with the climatological Gulf Stream core position determined from AVISO altimetry data (black dashed). Plot C and F: Argo float 4900136 profiles #78 and #80 (in red) superimposed with the standard reference envelop (blue). Plot D and G: Argo float 4900136 profiles #78 and #80 (in red) superimposed with the PCM-based reference envelop (blue). Plot E and H: Argo float 4900136 trajectory in red with profiles #78 and #80 (red circle) superimposed with the AVISO map of Absolute Dynamic Topography height (contours every 0.1m) and climatological Gulf Stream position (black dashed). See text for details of reference envelops.

one hand, this surface dynamical context helps us to localize the profile #78 within the realm of a warm anti-cyclonic eddy located to the North of the Gulf Stream core. Using the PCM shown in Figure 4-A/B the float profile #78 was attributed to class #3, the frontal class, which is coherent with its dynamical context shown Figure 4-E. On the other hand, the float profile #80 is localized on the warm flank of the Gulf Stream that, for this particular date, is further north than its climatological position (black dashed line, Figure 4-H). Interestingly, despite the profile being localized close to the Gulf Stream core and flanked to the South by a cyclonic cold eddy, the PCM appropriately attributed it to the class #1 of southern warm waters.

This illustrates how appropriate is the PCM attribution of profiles to group with similar vertical structures, whatever their location in space. Thus, in Figure 4-D and G are shown the PCM-based reference envelop (in blue) for which the distance weighted mean and standard deviations were computed only with profiles of the reference collection attributed to the same class as the Argo float profiles to be validated. The difference with the standard reference envelops is striking. Both profiles #78 and #80 are now within the bounds of the reference statistics and would no longer raise false alarms.

One can note that these results are robust to the decorrelation length scales and to the number of profiles used to compute the standard deviation (also note that we used the same number of reference profiles to compute the standard and PCM-based statistics).

Model Evaluation

A PCM also represents an elegant method to synthesize the structural information of a profile collection. Then a PCM trained with observations can be used to evaluate numerical model realism by comparing the space/time distribution of the classes. We could also compare the two optimal PCMs trained with observations on the one hand and the numerical model on the other hand (e.g. Fig.3-A from a model compared to Fig.4-A from observations). But let's illustrate a simpler first case here. We used the Argo-based PCM of Maze et al (2017) for the North-Atlantic to evaluate a state of the art global ¼° resolution configuration of a NEMO simulation. Figure 4-A and B show the distribution of the classes attributed to model temperature profiles for the first year of the simulation (1958) and for the entire run period (1958-2015). This distribution should be compared with Fig.1 insets or Fig.11 in Maze et al (2017).

This evaluation indicates that, although the class distribution is correct at the beginning of the run (no spin-up was performed, so the first year remains close to the initial conditions based on observations), the model dynamics clearly modifies the stratification structure (not shown), which leads to a re-arrangement of the classes in space. With regard to this method, the model performs well in most of the North Atlantic Ocean, except over the Western subtropical region, south of the Gulf Stream, where class #3 (cyan) takes over class #8 (brown) that, in turn, considerably shrinks. This is due to the model dynamics in the Gulf Stream region that erodes the vertical stratification structure (the mode water and underlying permanent pycnocline, Feucher et al, 2016) and sustains a more vertically uniform structure, thus the new state is more like class #3 than class #8 (see Fig.1).

CONCLUSION

In this letter, we briefly presented a data mining statistical method recently proposed by Maze et al (2017) for physical oceanographic studies. This method is coined «Profile Classification Model» or PCM. It is based on a state of the art un-supervised classification method, a Gaussian Mixture Model, being applied to ocean profiles. They proposed several PCM applications, four are illustrated here.

This method is coined «Profile Classification Model» or PCM. It is based on a state of the art un-supervised classification method, a Gaussian Mixture Model, being applied to ocean profiles.

In Maze et al (2017), a PCM is derived from Argo temperature profiles in the North Atlantic Ocean. They showed that

In this letter, we briefly presented a data mining statistical method recently proposed by Maze et al (2017) for physical oceanographic studies. This method is coined «Profile Classification Model» or PCM. It is based on a state of the art un-supervised classification method, a Gaussian Mixture Model, being applied to ocean profiles. They proposed several PCM applications, four are illustrated here.

In Maze et al (2017), a PCM is derived from Argo temperature profiles in the North Atlantic Ocean. They showed that 8 classes of profiles capture the diversity of all possible vertical structures (Figure 1). It is worth noting that this vertical structure is not only defined by a vertical mean profile but also by a square covariance matrix that contains 8 classes of profiles capture the diversity of all possible vertical structures (Figure 1). It is worth noting that this information about mode waters (no spread, homogeneity), thermoclines - or any other vertical gradients of the tracer used (large spread) or lack of vertical coherence throughout the class (frontal regions).

Maze et al (2017) showed that, although no spatial information is used to train the PCM, each of the 8 classes is co-localized in space, defining regions of the ocean with a unique vertical structure. We thus illustrated a possible application of such natural region contouring for the North Atlantic subpolar gyre (Figure 2). We furthermore examined the decomposition of its integrated heat content variability into the component driven by local temperature variations and the component driven by the gyre expansion and contraction. We found that,

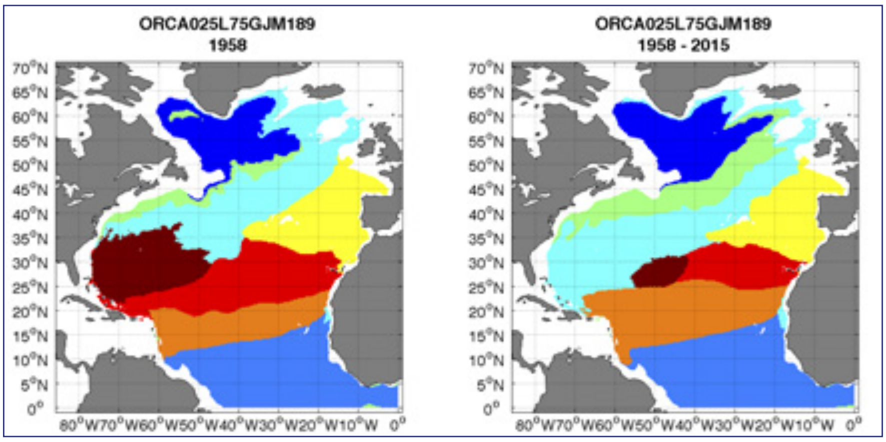

FIGURE 5

Distribution of the locally most frequent classes attributed to a GCM run, for the first year of the simulation, 1958 (left) and for the entire time series, 1958-2015 (right). The PCM used is the one trained with Argo data for which the structure and distribution are shown in Figure 1.

at the interannual time scale, the expansion component drives the gyre heat content while the temperature component is anti-correlated.

We also illustrated how a PCM can shed a new light on turbulent Western Boundary Current regions through the identification and grouping of the possible vertical structures. Using a 1/12° numerical model simulation from the DRAKKAR group (Barnier et al, 2014), we trained a PCM based on temperature data in the Gulf Stream Extension. We showed that, despite the strong seasonality of the profiles, a 3-class PCM is able to disentangle the horizontal complexity of the frontal region structure with a remarkable simplicity (Figure 3). This result will be further investigated and developed in the 2017-2019 INSU LEFE GMMC/IMAGO funded project «SOMOVAR» (LOPS, Telecom Bretagne, CERFACS). In particular, the project aims to use the PCM method for the detection of low-frequency variability and to develop a new product of in-situ gridded data in turbulent Western Boundary Current regions.

For observation data center, this latter result has a powerful direct application: a more appropriate selection of reference data for quality control procedures. This was illustrated in the Gulf Stream region for the hypothetic validation of two Argo float profiles located near the front (Figure 4). When statistics were computed (using a trivial approach of distance weighting and seasonal colocation) from a reference database to evaluate the Argo float profiles, we found the profiles to be outside of the standard reference envelop, hence raising a false alarm. But when the statistics were computed using only reference profiles with the same PCM class of the Argo float to be validated, then we found the profiles to be within the PCM-based reference envelop. The dynamical context of the profiles was provided by the AVISO altimetry data. It showed that profiles were too close to the front or within an eddy for the standard approach to be able not to bias low the reference envelop. Obviously, this can be avoided if the QC operator uses altimetry for context or more complex method, such as OW, working along isopycnal surfaces. But the PCM method, simple, automatic and in the depth/pressure space can surely help QC procedures.

Last, another application of the PCM method for model evaluation was illustrated. Indeed, a PCM is a reduced representation of the statistical properties of a collection of profiles. It thus provides the opportunity to compare two collections or to assess one with regard to the properties of the other. First case could be achieved for instance by comparing Fig.3-A from an eddy resolving simulation with Fig.4-A from Argo data. Here, we simply illustrated the later scenario in Figure 5. Using the Argo-based PCM of temperature profiles from Maze et al (2007), we classified a 1958-2015 time series of a global circulation model experiment at ¼° resolution (ORCA025, 75 vertical levels, referenced by the DRAKKAR group as GJM189). We found the model initial state to be close to observations (as expected, because the model had no spin-up) but to drift away from a realistic stratification in the Southern recirculation region of the Gulf Stream, while the other regions were stable and remained realistic. This synthetic metric provides an elegant way to assess the realism of the model state.

Our group is currently working on the PCM applications illustrated here. But, obviously, this approach can be used with other data (e.g. salinity, both temperature and salinity, density, stratification...), in other frontal regions (e.g. the ACC) and with other datasets (e.g. CORA4.2, high resolution model outputs). To foster such possible applications we made available online the Argo-based PCM (http://doi.org/10.17882/47106) and a toolbox to easily train a PCM and classify new data (https://github.com/obidam/pcm) from a collection of profiles or gridded datasets.

REFERENCES

Bishop, C. M., 2006: Pattern recognition and machine learning, Springer (738p).

Maze, G., 2017: A Profile Classification Model from North-Atlantic Argo temperature data, Seanoe, doi: 10.17882/47106.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Lopez Radcenco, M., Lenca, P., Feucher, C., and Le Goff, C., 2017: Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean. Progress in Oceanography, doi: http://dx.doi.org/10.1016/j.pocean.2016.12.008.

Feucher, C., Maze, G., and Mercier, H., 2016: Mean structure of the North Atlantic subtropical permanent pycnocline from in-situ observations. Journal of Atmospheric and Oceanic Technology, doi: 10.1175/JTECH-D-15-0192.1.

Cabanes, C., Thierry, V., and Lagadec, C., 2016: Improvement of bias detection in Argo float conductivity sensors and its application in the North Atlantic. Deep Sea Research Part I: Oceanographic Research Papers, doi: http://dx.doi.org/10.1016/j.dsr.2016.05.007.

Gaillard, F., Reynaud, T., Thierry, V., Kolodziejczyk, N., and von Schuckmann, K., 2015: In Situ–Based Reanalysis of the Global Ocean Temperature and Salinity with ISAS: Variability of the Heat Content and Steric Height. Journal of Climate, 29 (4), 1305--1323, doi: 10.1175/JCLI-D-15-0028.1.

Barnier, B., Blaker, A. T., Biastoch, A., Böning, C. W., Coward, A., Deshayes, J., Duchez, A., Hirschi, J., Sommer, J. L., Madec, G., Maze, G., Molines, J.-M., New, A., Penduff, T., Scheinert, M., Talandier, C., and Treguier, A.-M., 2014: DRAKKAR: developing high resolution ocean components for European Earth system models. CLIVAR Exchanges, 65, 18-21.

Maze, G., Deshayes, J., Marshall, J., Tréguier, A.-M., Chronis, A., and Vollmer, L., 2013: Surface vertical PV fluxes and subtropical mode water formation in an eddy-resolving numerical simulation. Deep Sea Research Part II: Topical Studies in Oceanography, 91 (0), 128--138, doi: 10.1016/j.dsr2.2013.02.026.