Towards Chl-a Bloom Understanding by EM-based Unsupervised Event Detection

Alain Lefebvre, Émilie Poisson Caillault

Towards Chl-a Bloom Understanding by EM-based Unsupervised Event Detection

Émilie POISSON CAILLAULT*†
* Univ. Littoral Côte d'Opale, EA 4491-LISIC
F-62228 Calais, France

Alain LEFEBVRE †
† IFREMER, LER BL
F-62321 Boulogne-sur-mer, France

***Abstract*—Marine water quality monitoring and subsequent management require to know when a specific event like harmful algae bloom may occur and which environmental conditions and pressures lead to this event. So, event detection and its dynamic understanding are crucial to adapt strategy. An algorithm is proposed to identify curves mixture and their dynamics features - initiation, duration, peaks and ends of the event. The approach is fully unsupervised, it requires no tuning parameters and is based on Expectation Maximization process to estimate the most robust mixture according to fixed criteria. A complete framework is proposed to deal with a univariate time series with missing data. The approach is applied on Chlorophyll-*a* series collected weekly since 1989. Chlorophyll-*a* is a proxy of the phytoplankton biomass. The results are promising according to the phytoplankton composition knowledge, collected at lower frequency, and allowing to discuss about the annual variability of phytoplankton dynamics.**

***Keywords*—Time series; Event detection; Expectation-Maximisation; Phenology; Chlorophyll-a; Phaeocystis.**

I. Introduction

In numerical ecology, classification methods are of paramount importance to synthesize information, to understand data structure and then, to extract the maximum amount of information. They are used for improving knowledge and issuing practical recommendations for environmental management. In the context of better assessing the quality of marine waters and the Ecological (MSFD [1]) or Eutrophication Status (OSPAR [2]), knowledge about ecosystem dynamics at fine timescale is often incomplete and requires an unsupervised approach [3].

Let's consider the dynamic phenomenon of phytoplankton abundance and the biomass dynamics of specific harmful algae. Species composition and relative abundance of algal groups are fundamental determinants of aquatic ecosystem structure and function. But their behavior change over the years and within one year: event dates, duration, peak dates and amplitude.

Several approaches has been tested in order to assume a trend and/or a shift in phytoplankton biomass, and a modification of the bloom phenology responding to varying environmental condition under human pressures. FULCRUM approach [4] based on the cumsum approach and breakpoints both does not permit to identify the beginning/maximum or end of the bloom and to detect general patterns or specific patterns. Seasonal, trend analysis [5] and Empirical Mode Decomposition (EMD [6]) are not adapted for non monotonic trend series with mode under half-year frequency.

So the problem we investigate is to identify curves mixture by Expectation-Maximisation process [7] in Chlorophyll-*a* time series. For this first study, we assume that, within time series, each event could be associated to a Gauss curve, and an event cannot be fully included in another. Only small overlap between events is admissible.

This paper is organized as follows. Section 2 introduces materials and methods, including: Data presentation; data completion process and the algorithm of event detection and time characterization in yearly time series. Next, in Section 3 experimental results are presented and discussed. Conclusion and future works are drawn in Section 4.

II. Materials and Methods

A. Data presentation and general scheme

Chlorophyll-*a* (*Chla* in $\mu g.l^{-1}$) time series have been acquired weekly since 1989 and monthly since 1978 in coastal waters in the southern bight of the North Sea (Ifremer IGA Gravelines monitoring program [8]). This time series is one of the longest one in France. In the Phytoplankton field, this series are building our overall understanding of Phytoplankton bloom dynamics in the context of the massive proliferation of the Prymnesiophycea *Phaeocystis globosa* [9]. *Phaeocystis globosa* is classified as a High-Biomass HAB species (Harmful Algal blooms - HAB) and its proliferation leads to nuisance and impacts benthic and pelagic habitats.

The objectives to define various typology/phenology of algae blooms from Chlorophyll-*a* data first require an explorative analysis and method of missing data completion before the event detection and characterization. Our methodology consists of 3 steps described as follows.

First step corresponds to data preprocessing : data validation, time shifting, missing data completion. The second and main level concerns the identification of recurrent or rare events that look like a bell-shaped curve in an subsequence. Here the algorithm to estimate one mixture of Gauss curves is proposed to model seasonal *Chla* series or to model each yearly *Chla* series. This method automatically provides information up to week scale. This method can be extended to any other series and at other scale frequency. The series should be cut in subsequences and a mixture model may be extracted from each subsequence by the algorithm.

The last process is the correspondence analysis between obtained Gauss curves and the associated phytoplankton community structure, abundance of taxa, up to the species level in this area.

B. Data preprocessing

Extraction and time alignment. Figure 1 illustrates the *Chla* time series from 1978 onwards. A monthly resampling from 1978 to 1989 will lead to high information loss both in terms of phenology and bloom dynamics. And it is difficult to validate data values before 1989. That is the reason why only the period 1989-2014 are retained for the study. During this period corresponding to 1,352 weeks, 146 samples are missing. But only small missing subsequences (1 week) occur during the critical phytoplankton productive period and gaps from one to seven weeks for the non productive period, so filling this gap by an adapted moving average is appropriate.

Imputation part. The filling process consists in two steps. First, single missing data (1-length) $x(t)$ are completed by the mean of the precedent value $x(t-1)$ and the future value $x(t+1)$. Then, missing subsequences are examined by human expert who validates the completion process according to the dynamics knowledge of the considered process. Here, the holes for non productive periods are always imputed. A hole during the productive period will be completed only if the size is less than the process change that we want to detect. For phytoplankton dynamics, the completion of hole is based on the timescale graph of phytoplankton phenomena described by Dickey et al. [10].
A hole is imputed by a moving average weighted using a membership function mixing confidence score of the data value (acquired data or inserted data), and degree of spacing in the considered temporal window. Table I details the mechanism of dynamic completion, step by step, for a subsequence of three missing values noted NA.

$x(t)$	5.9	5.9	11.2	NA	NA	NA	3.7	5.1	6.0
weights	3	4	50	0	0	0	3	2	1
$x(t+1)$	5.9	5.9	11.2	9.97	NA	NA	3.7	5.1	6.0
weights	0	2	3	1	0	0	3	2	1
$x(t+2)$	5.9	5.9	11.2	9.97	6.88	NA	3.7	5.1	6.0
weights	0	0	1	1	1	0	30	2	1
$x(t+3)$	5.9	5.9	11.2	9.97	6.88	4.31	3.7	5.1	6.0

TABLE I
IMPUTATION WITH WEIGHTED AVERAGE FOR A SERIES WITH 3 CONSECUTIVE MISSING VALUES (NOTED NA).

The weighting is built such as Dirac peak: directly consecutive data ($t-1$ and $t+1$) are assigned of a high weight, and the others of a decreasing weight coefficient. Completed data are inserted for the computation of the next point, but with a low confidence weight (1-value). To be more robust, a first imputation is made in clockwise direction and another in the anticlockwise direction; the retained completed values are the mean of the two obtained series. Figure 2 is a zoom of the *Chla* series where the longest hole is filled.

Exploratory analysis. Autocorrelation (ACF) highlights an annual periodicity with peaks repeated 52 weeks, but the correlation coefficient is lower than 0.4. This seasonal cycle can be extracted from an additive decomposition (R decompose function, package stats) given in Figure 3. Figure 4 is the overlay of each year *Chla* subsequence: an important variability of the dynamics exists. The level and the number of peaks are different from one year to another, with clear trend and/or shift which may be linked to varying environmental conditions .

C. EM Gauss curves extraction and criteria

Our automatic event detection is based on normalmixEM algorithm from the R-package mixtools [11]. In this study, event is assumed to be a bell-shaped curve and follows this equation: $g(t) = lambda * exp(-(t-mu)^2/(2*sigma^2))$ with (mu, sigma, lambda) its parameters to estimate. The search of the number of Gauss curves noted Ng are incremental, and stopped according to a rebuilding criterion. Expectation-Maximization methods (EM) are not exact, they find a local optimization of parameters and strongly depend on the initialization step. Our EM process are repeated T times. Then, a model selection step is applied to retain the most stable model (repeatable), and best fulfills our criterion. T is in correlation with k the number of clusters to determine the dominant model. From experiments, ($T = 20$, $k = 3$) is relevant and enough to extract a robust model.

The framework of the algorithm is detailed with the following pseudocode in Algorithm 1 (the R code is available upon mail request to the authors).

The criterion is based on a strict fusion of similarity measures and admissible deformations. The details and rules for each criterion between x and the rebuilt signal r from the Gauss parameters

$$r(t) = \sum_{g}^{nbG} lambda_g \times exp(-(t-mu_g)/sigma_g^2)$$

are cited below:

1) Correlation Coefficient R2.cor $(x, r) \geq$ par.R2
2) Normalized Mean Square Error:
   NMSE$(x, r) = mean(x-r)/(mean(x) * mean(r)) <$ par.NMSE
3) Similarity by area
   simArea$(x, r) = sum(x-r)/sum(x) >$par.simArea
4) Percentage of outliers (predictions within a factor of two of the observed values).
   FA2$= length(0,5 \leq (x/r) \leq 2)/length(x)$
   FA2$(x, r) \geq$ par.FA2
5) Fractional Bias, rate of data under or overestimated according the mean noted m.
   $|FB(x, r)| == 2 * (m_x - m_r)/(m_x + m_r) <$par.FB
6) Fractional Standard deviation (sd)
   $|FS(x, r)| = 2 * (sd_x - sd_r)/(sd_x + sd_r) <$ par.FS

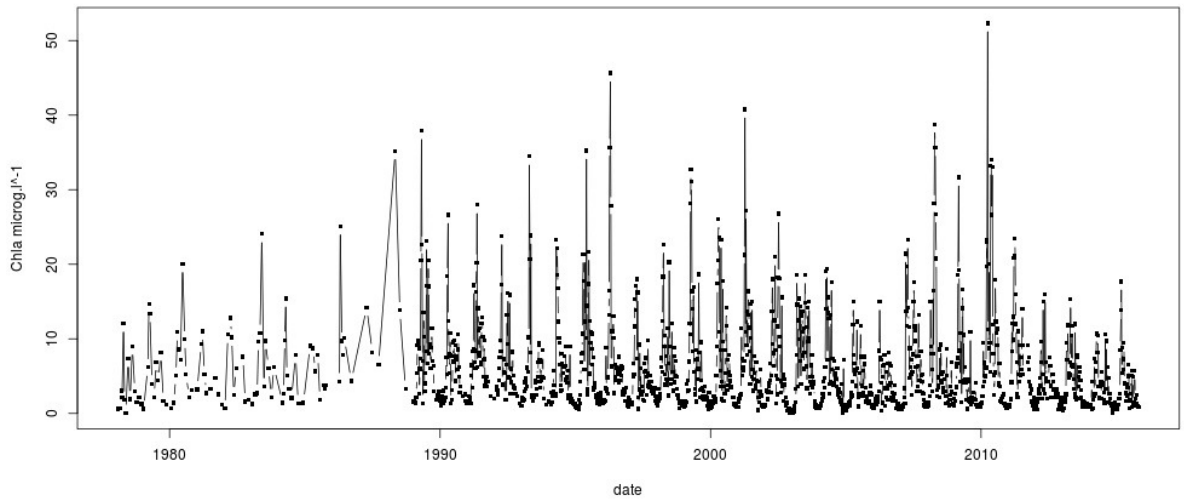

Fig. 1. Chlorophyll-*a* time series collected in the Southern bight of the North Sea (Lat. 51.015 N, Long. 2.092 E).

Algorithm 1 PseudoCode of curve mixture detection

Input: (t, x, acceptRate)
Output: 3 vectors (mu, sigma, lambda)
Initialisation : test $\leftarrow$ false;
Initialisation : mu, sigma, lambda $\leftarrow$ null;
$y \leftarrow x/\max(x)$;
maxG $\leftarrow$ compute the max accepted number of Gauss curves according to the number of peaks and valleys;
$yh \leftarrow$ transform y to obtain a density repartition
while test==false and $g \leq$maxG **do**
  # research of a robust model
  for $i = 1$ to 20 **do**
    mixmodel(i)$\leftarrow$normalmixEM(yh,k= g);
  end for
  # Research of the best representative model
  cluster$\leftarrow$kmeans(mixmodel, k=3)
  index$\leftarrow$select the dominant group according the criterion of 50% mu in the same group are close. (no singleton cluster).
  for i in index **do**
    $C_i \leftarrow$sumGaussCurve(mixmodel(i),t)
    $score_i \leftarrow$compute rebuilding criteria between C_i and x;
  end for
  $b \leftarrow argmax_i(score_i)$;
  test$\leftarrow$ true if all criteria respected between $score_b$ and acceptRate, false otherwise;
  $g \leftarrow g + 1$;
end while
return $lambda, mu, sigma$

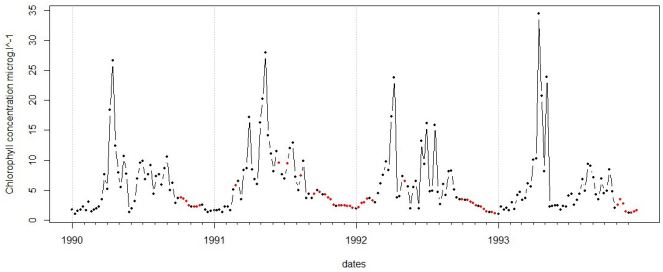

Fig. 2. Zoom on the *Chla* series for the period 1990-1993 with filled gap of length 3 and 6 in red color the completed points.

7) Geometric Mean bias [12]
   MG$(x, r) = exp(mean(ln(x)) - mean(ln(r))$
   par.MGmin $\leq$ MG$(x, r) \leq$ par.MGmax
8) Geometric Mean Variance [12]
   VG$= exp(mean(ln(x) - ln(r))^2)$
   par.VGmin $\leq$ VG$(x, r) \leq$ par.VGmax

Variables beginning by "par." correspond to the element of the vector "acceptRate", input of the precedent algorithm. This vector of criterion could be adapted according to the problem. For this application, par.R2=0.7, par.FA2=0.8, par.FB=0.3, par.FS=0.05, par.NMSE=0.4, par.simArea=0.95, par.MGmin=par.VGMin=0.75, par.MGmax=par.VGmax=1.25. The signal reconstruction constraints are hard.

The algorithm is applied to the seasonal cycle, then to each year (composed of 52 *Chla* values per week).

III. Results and Discussion

The proposed approach detects a mixture of 7 Gauss curves within a seasonal pattern of the General *Chlorophyll-*

# Curve number	G1	G2	G3	G4	G5	G6	G7
Peak date	2.6	11.8	15.4	20.2	27.1	36.5	48.8
Range dates	1-7	1-23	11-20	15-25	15-39	21-52	43-52
Sigma-dates	1-4	8-16	13-17	18-22	23-31	30-42	40-51
Shannon index	0.6;2.6;3.8;	0.5;2.1;3.7	0.5;2.1;3.7	0.6;2.6;3.9	0.5;3.0;4.3	0.3;2.8;4.0	0.1;2.0;4.5
#m Taxon	22	36	30	27	38	44	26
#m 95% cell/L	12	8	1.5	4	14	17.5	14.5
Dominant species	Melosira P.S	Phaeocystis	Phaeocystis	Phaeocystis	Rhizolenia imb.	Leptocylindrus	Melosira P.S.
# years/25	17	18	21	16	14	4	13

TABLE II
CHARACTERISTICS OF EACH EVENT, GAUSS CURVE NAMED G_i.
SHANNON INDEX ARE MIN/MEDIAN/MAX VALUES AND $\#m$ REPRESENTS THE NUMBER OF TAXA.

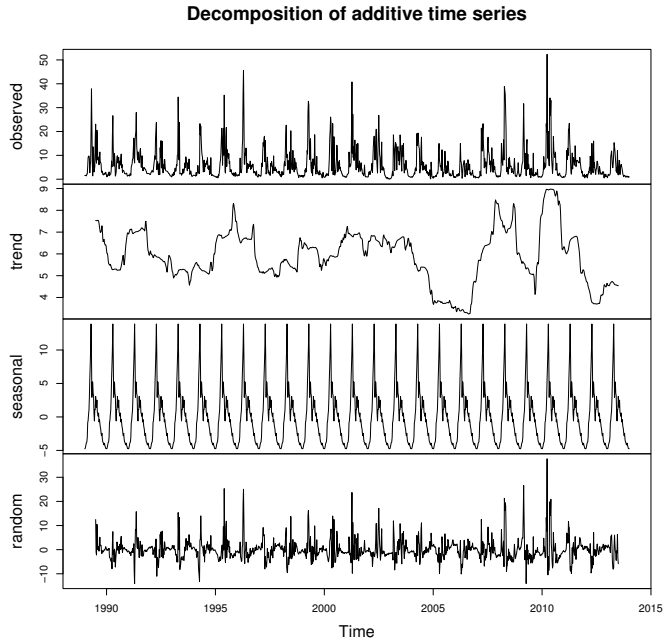

Fig. 3. Additive decomposition in trend and seasonal cycle (R-decompose method) of the completed *Chla* time series in $\mu g.l^{-1}$

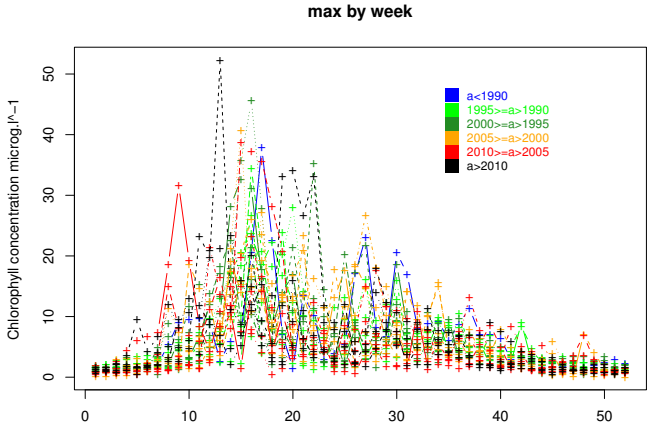

Fig. 4. Overlay of each *Chla* time series per year, weekly sampled.

a concentration dynamics during 1989-2014 period. Figure 5 illustrates 7 events, black color represents the seasonal cycle (by multiplicative decomposition), and pink color with dotted line the reconstructed signal.

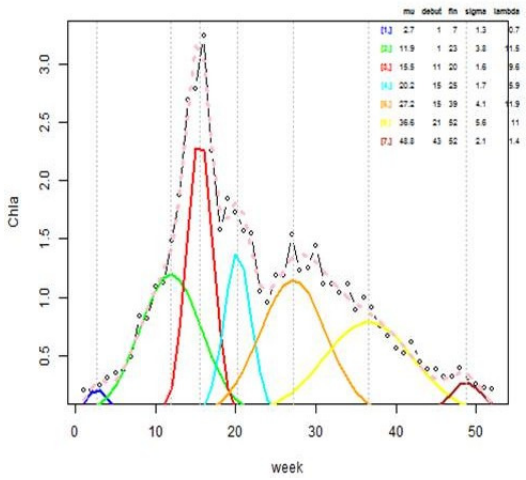

Fig. 5. 7 Estimated Gauss Curves, representative of a typical phytoplankton biomass dynamics (1989-2014) in the southern bight of the North Sea. *Chla* unit is in $\mu g.l^{-1}$.

Event characteristics are then compared to abundance information of 79 labels (Taxonomic Units) derived from an expert (manual) classification from the IFREMER Quadrige2 database.

Our system is also able to propose an automated identification of diversity and richness indexes of each event, and an automated identification of the most contributed taxonomic units for each period/state. Table II presents these temporal characteristics per event. These results permit to define a climatology 1989-2014: the mean yearly cycle of the phytoplankton development corresponds to the succession of 7 phytoplankton communities. For instance, G1 event occurs from week 1 to week 7 and presents a concentration peak in *Chl-a* between the second and third week of the year. Shannon diversity index during this period ranges from 0.6 to 3.8. This event is characterized by a co-dominance of 22 taxa out of which twelve represent 95 percent of the total abundance. Three events are characterized by one to three taxa (>95 % of abundance), with very low Shannon index and a dominance of *Phaeocystis Globosa*.

A comparison between obtained model from the seasonal decomposition and obtained models per year shows an important variability of the number of events, of their shapes and thus of the Gauss parameters: dates of initiation, ends of the bloom, duration and value range. Yearly Chlorophyll-*a* concentration dynamics is explained by 5 to 14 Gauss curves. Figure 6 show 2006 year explains by 8 events, clearly separated in opposite of the mean pattern.

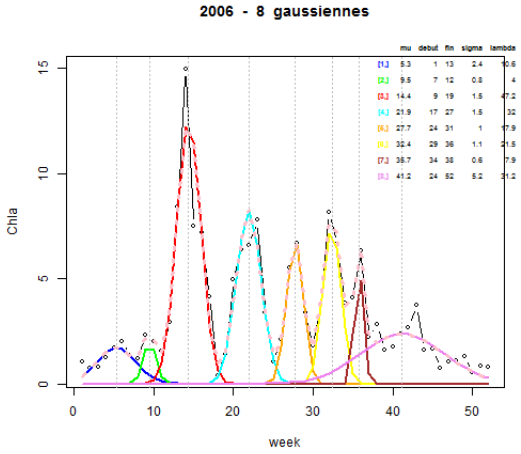

Fig. 6. 8 Estimated Gauss Curves from the *Chla* 2006-subsequence in the southern bight of the North Sea. *Chla* unit is in $\mu g.l^{-1}$.

Next, it will be important to consider this variability and to try to understand the 5 dominant yearly models computed from an unsupervised spectral clustering [13] that have no evident succession.

IV. Conclusion

Two approaches are detailed in this paper: a fuzzy moving average completion of univariate time series, and its unsupervised segmentation. This automatic extraction of mixture of Gauss curves in univariate time series is proposed in order to detect event and its temporal characteristics (beginning and end date, peak level). The framework is applied on Chlorophyll-*a* series weekly collected in the south bight of the North Sea, and permits to define automatically a typical long-term-based seasonal pattern of phytoplankton biomass. Moreover, for each subpattern we are able to define parameters of phenology diversity. In particular, three events correspond to one dominant harmful species, *Phaeocystis*, in the coastal area. The method has also been applied to each yearly sequence, and shows an important variability of the numbers of events. The future works will consist in understanding and labelling all these detected events in order to study their dynamics. Then, the assumption of bell shape has to be relaxed.

Acknowledgments

This work has been funded the French government and the region Hauts-de-France in the framework of the project CPER 2014-2020 MARCO. Thanks to the IFREMER LER/BL team in Boulogne-sur-Mer for their helps during the extraction, corrections and combination process of the phytoplankton database. The phytoplankton data were collected within the framework of the research programmes IGA (Impact des Grands Aménagements), conducted by IFREMER with financial support from EDF (Electricité de France).

References

[1] Directive 2008/56/EC of the European Parliament and of the council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy, "Msfd: Marine strategy framework directive. official journal of the european union," 2008.

[2] Oslo Paris Convention for the Protection of the North Sea OSPAR, "Common assessment criteria, their assessment levels and area classification within the comprehensive procedure of the common procedure. ospar commission for the protection of the marine environment of the north-east atlantic. http:// www.ospar.org/eng/html/welcome.html." 2002.

[3] J. G. Ferreira, J. H. Andersen, A. Borja, S. B. Bricker, J. Camp, M. C. da Silva, E. Garcés, A.-S. Heiskanen, C. Humborg, L. Ignatiades, C. Lancelot, A. Menesguen, P. Tett, N. Hoepffner, and U. Claussen, "Overview of eutrophication indicators to assess environmental status within the european marine strategy framework directive," *Estuarine, Coastal and Shelf Science*, vol. 93, no. 2, pp. 117 – 131, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0272771411001077

[4] J. C. Kromkamp and T. Van Engeland, "Changes in phytoplankton biomass in the western scheldt estuary during the period 1978–2006," *Estuaries and Coasts*, vol. 33, no. 2, pp. 270–285, 2010. [Online]. Available: http://dx.doi.org/10.1007/s12237-009-9215-3

[5] I. Alvarez, M. N. Lorenzo, and M. deCastro, "Analysis of chlorophyll a concentration along the galician coast: seasonal variability and trends," *ICES Journal of Marine Science: Journal du Conseil*, vol. 69, no. 5, pp. 728–738, 2012. [Online]. Available: http://icesjms.oxfordjournals.org/content/69/5/728.abstract

[6] J. YAN, Tongand WANG, "Multi-timescale analysis of chlorophyll and its related physical factors northwest of the luzon island based on hht," *Journal of Tropical Oceanography*, vol. 30, no. 5, p. 38, 2011. [Online]. Available: http://www.jto.ac.cn/EN/abstract/article_287.shtml

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[8] D. Woehrling, A. Lefebvre, and R. Le Fèvre-Lehoërff, G.and Delesmont, "Seasonal and longer term trends in sea temperature along the french north sea coast, 1975 to 2002." *Journal of the Marine Biological Association U.K., 85 (1)*, pp. 39–48, 2005.

[9] A. Lefebvre, N. Guiselin, F. Barbet, and L. F. Artigas, "Long-term hydrological and phytoplankton monitoring (1992-2007) of three potentially eutrophicated systems in the eastern english channel and the southern bight of the north sea." *ICES Journal of Marine Science, 68(10)*, pp. 2029–2043, 2011.

[10] J. Dickey, "Emerging ocean observations for interdisciplinary data assimilation systems," *Journal of Marine Systems*, vol. 40-41, pp. 5–48, 2003.

[11] T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young, "mixtools: An r package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009. [Online]. Available: http://www.jstatsoft.org/v32/i06/

[12] S. Ahuja and A. Kumar, "Evaluation of MESOPUFF-II SOx transport and deposition in the great lakes region," in *AWMA Speciality Conference on Atmospheric Deposition to the Great Lakes, VIP-72, pp. 283-299, Oct. 28-30*, 1996.

[13] Wacquet, G., É. Poisson-Caillault, and P. Hébert, *Semi-supervised K-Way Spectral Clustering with Determination of Number of Clusters*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 317–332. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-35638-4_21