

Coupling Oceanic Observation Systems to Study Mesoscale Ocean Dynamics

Gautier Cosne, Guillaume Maze, Pierre Tandeo

▶ **To cite this version:**

Gautier Cosne, Guillaume Maze, Pierre Tandeo. Coupling Oceanic Observation Systems to Study Mesoscale Ocean Dynamics. NeurIPS 2019 : 33rd Conference on Neural Information Processing Systems, Dec 2019, Vancouver, Canada. hal-02928416

HAL Id: hal-02928416

https://imt-atlantique.hal.science/hal-02928416

Submitted on 2 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1910.08573v1 [physics.ao-ph] 18 Oct 2019

Coupling Oceanic Observation Systems to Study Mesoscale Ocean Dynamics

Gautier Cosne [1] **Guillaume Maze** [*2] **Pierre Tandeo** [*3]

Abstract

Understanding local currents in the North Atlantic region of the ocean is a key part of modelling heat transfer and global climate patterns. Satellites provide a surface signature of the temperature of the ocean with a high horizontal resolution while in situ autonomous probes supply high vertical resolution, but horizontally sparse, knowledge of the ocean interior thermal structure. The objective of this paper is to develop a methodology to combine these complementary ocean observing systems measurements to obtain a three-dimensional time series of ocean temperatures with high horizontal and vertical resolution. Within an observation-driven framework, we investigate the extent to which mesoscale ocean dynamics in the North Atlantic region may be decomposed into a mixture of dynamical modes, characterized by different local regressions between Sea Surface Temperature (SST), Sea Level Anomalies (SLA) and Vertical Temperature fields. Ultimately we propose a Latent-class regression method to improve prediction of vertical ocean temperature.

1. Introduction

The ocean is a turbulent and chaotic dynamic system, very complex to observe, understand and to predict. This is especially true off the northeast coast of the United States where the Gulf Stream shrinks to intense fronts, meanders, and numerous eddies. For instance, Zhang et al. (2014) shows that oceanic transports of heat of mesoscale eddies, which covers scales from 50 to 200 km^2, is comparable in magnitude to that of the large-scale wind and thermohaline-driven circulation. Understanding their thermal structure is key to follow the diffusion and transport of heat in the ocean.

However, none of the existing ocean observing systems are able to provide an estimate of 3D thermal field with enough resolution to study individual eddies of mesoscale. Satellite observations have a high local frequency and high horizontal resolution, but they only capture the surface signature of the ocean dynamic and mesoscale eddies. On the other hand, *in situ* measurements collected by autonomous probes such as the ARGO floats (Riser et al., 2016) are optimally providing one vertical profile of temperature (and salinity) measurements over $300km^2$ per month, which is way too sparse to sample mesoscale eddies; but uniquely provides information about the ocean interior. Other observation systems exist, such as gliders, but they are not providing measurements systematically and in real time such as the other two mentioned above. This makes our understanding of these regions and their behavior under the influence of global climate change further limited by our ability to diagnose their 3D thermal structure and temporal evolution.

Latent class regression methods aiming to predict the ocean temperature profile based on several local linear relationship with satellite data (Sea Surface Temperature, SST, and Sea Level Anomaly, SLA) are used to take into account the presence of several dynamic modes observed in turbulent regions (Tandeo et al., 2014), each mode being characterized by local regression. Maze et al. (2017) introduced an unsupervised classification of oceanic profiles that provides a direct probabilistic link between the vertical oceanic structure of a measurement and the dynamic modes. In this paper, we develop several strategies to combine satellites and *in situ* ocean measurements to produce a 3D time series of temperature with high horizontal resolution. By leveraging knowledge from the former unsupervised classification of oceanic profiles, we demonstrate that latent class regression methods improve the prediction of vertical ocean temperature fields.

2. Data

In a real-case scenario we would be using remote sensing (SST and SLA) and *in-situ* data taken from the ARGO database. ARGO is a real-time global ocean observation network consisting of about 3000 autonomous profiling probes randomly distributed in all oceans and measuring

[*]Equal contribution [1]IMT Atlantique, Technopôle Brest-Iroise Brest, France
[2]Ifremer, University of Brest, CNRS, IRD, Laboratoire d'Oceanographie Physique et Spatiale (LOPS), IUEM, Brest, France [3]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France. Correspondence to: Gautier Cosne <gautier.cosne@imt-atlantique.net>.

Unofficial Proceedings of NeurIPS 2019 Workshop - Tackling Climate Change with Machine Learning

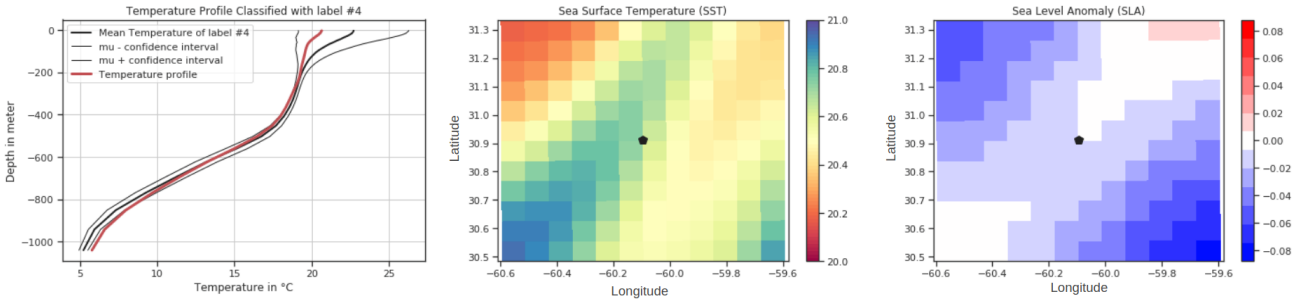


Figure 1. Overview of the data: On the left a vertical temperature profile between 0 and 1000 meters belonging to the mode #4. In the middle a co-localised 13×13 pixels patch of SST and on the right of SLA. One pixel is about $70km^2$.

pressure, temperature and salinity. The ARGO database is a collection of temperature and salinity vertical profiles going from the surface down to 2000 m, evenly distributed throughout the seasonal cycle and with approximately one profile per month per 3^ox3^o cell between 2000 and 2014. Routinely used in physical oceanography, these data are key to the observation of the ocean climate (Riser et al., 2016).

However, in order to keep our demonstration away from technical difficulties associated with real-world observations, in this study we will use synthetic observations, i.e. data generated with an ocean general circulation model (Treguier, 2008) that provide all necessary data on a 3D grid representing the North Atlantic Ocean, i.e. full data every 8-km. This is a state of the art synthetic set of observations that has been shown to reproduce the ocean structure and dynamic faithfully, especially the turbulent Gulf Stream (Maze et al., 2013). More precisely we will sub-sampled the synthetic set of observations to mimic the real sampling of ARGO sparse probes.

Given Sea Surface Temperature (SST) and Sea Level Anomalies (SLA) patches of size 13×13 pixels covering approximately 70km^2, we develop a model to predict the associated Vertical Temperature Profile at the center of the patches. A temperature profile represents the temperature measured at 36 different depths distributed between 0 and 1040 meters below the surface. We note X the concatenation of (SST, SLA) and Y a profile of temperature. The dataset of study used for training and evaluation is composed of 700 pairs of vectors X and Y sampled at different locations in space and time inside the region of study, with 66 % of the data kept to fit the model and the rest used for evaluation. In this document we will denote by *UC-label* the four labels from the unsupervised classification of profiles applied to the this domain and following Maze et al. (2017). We can clearly see the different water masses with cold and warm temperature at the surface in Figure 2-left, and matching regions with specific *UC-label*. *UC-label* 1 and 2 denote the cold and warm flanks of the Gulf Stream, while the *UC-label* 3 and 4 denote the frontal regions of the Gulf Stream and continental shelf zones, respectively.

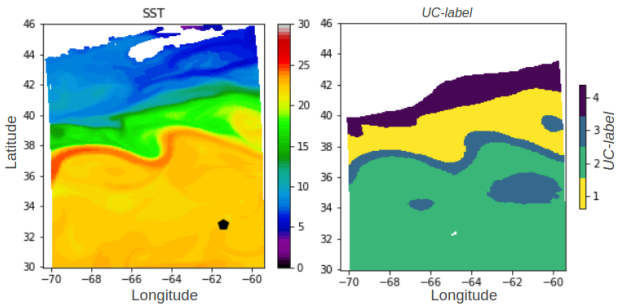


Figure 2. On the left the SST field covering the zone of study. On the right the *UC-label* distribution. Both fields vary with time.

Figure 1 presents a sample of the data set; the black dot in the center of SST and SLA patches is the location of the vertical temperature profile that is used to produce *UC-label* in 2-right. The training and evaluation set are voluntarily small to stay true to the amount of ARGO profiles we would have access on a short period of time. The test set is however composed of 1000 randomly chosen profiles and 27 different time-step equally spaced in time over a year period.

3. Latent Class Regression Model

Motivated by the fact that one SST/SLA relationship can arise from different vertical temperature profiles but that one vertical temperature profile correspond to a single SST/SLA relationship, our objective is to identify K modes (latent variable $\mathcal{Z}$) from a joint set of SST and SLA patches (p-dimensional vector X) adapted to predict the temperature profiles. Since (Maze et al., 2017) have suggested that a given dynamic region (hence caracterised by a single SST/SLA relashionship) is associated with a unique vertical profile class, we will assume here that *UC-label* are an appropriate choice to learn local linear relationship between SST, SLA and the temperature profile. We assume that the conditional probability of Y given X and a dynamical mode

$\mathcal{Z} = k$ is given by equation 1.

$$p(Y|X, \mathcal{Z} = k) \sim \mathcal{N}_k(Y; X\beta_k, \Sigma_k) \quad (1)$$

Where $\mathcal{N}_k$ represents a multivariate Gaussian probability density function evaluated in Y with mean $X\beta_k$ and covariance Σ_k. Hence, the conditional probability of $Y|X$ resorts to a mixture of normal distributions, where λ_k is the prior probability of mode k. :

$$p(Y|X, \Theta) = \sum_{k=1}^{K} \lambda_k \mathcal{N}_k(Y; X\beta_k, \Sigma_k) \quad (2)$$

To simplify the notations, we store the overall parameters of equation 2 in $\Theta = (\lambda_1, .., \lambda_K, \beta_1, .., \beta_K, \Sigma_1, .., \Sigma_K)$. In the literature this model is referred to as a "latent class regression". The maximum likelihood estimation procedure for model parameters Θ is given in DeSarbo & Cron (1988). A key aspect of the latent class regression is the choice of K the number of latent mode. A popular method to estimate K is the Bayesian Information Criteria (Schwarz, 1978). The BIC is an empirical approach of the model probability computed as:

$$\mathrm{BIC}(K) = -2\mathcal{L}(K) + \mathrm{N}_f(K)\log(n) \quad (3)$$

where $\mathcal{L}(K)$ is the log likelihood of the trained model with K classes. $\mathrm{N}_f(K)$ is the number of independent parameters to be estimated, in our case the parameters to estimate are the one of Σ_k the covariance matrix and β_k the local linear coefficient. For a given number of mode K, we suggest different evaluations of the Expectation Maximization (EM) algorithm (Dempster et al., 1977). The idea is to test different initialization of $\hat{\pi}_k$ (probability of belonging to mode k) for each sample for the EM procedure and then select parameter estimates corresponding to the greatest likelihood.

The main difficulty in training a latent-class regression method with oceanic profiles is that the dimensions along the patches and the vertical axis are large compared to the number of samples: we have to deal with a vector Y of size 36 and images of 13×13 of SST and SLA. This large number of dimensions translates into a large number of parameters to estimate in the covariance matrices Σ_k and in the local linear coefficient β_k. To reduce the number of dimensions, we use Principal Component Analysis. We chose to keep 99.9 % of variance for both vectors X and Y. Once we have projected X and Y in the PCA space, we can train our model with K varying and compute the BICs to set the best probabilistic K. From a BIC analysis we learn that the best K is 3. In the following section we compare K=3 (best BIC) to K=4 motivated by the fact that there are four *UC-label*.

Model\Metrics	MAE train	MAE Eval	MAE Test
PCA variance kept = 99.9 %			
Linear Model	1,96	2,09	**1,72**
KMEANS K=4	**1,54**	**1,86**	1,81
KMEANS K=3	1,65	1,96	1,98
UC-label	**1,47**	**1,87**	**1,69**
No PCA			
CNN	**0,87**	**1,27**	2,02

Table 1. Results of the different latent class regression model against the baseline. K is the number of modes..

4. Results and Discussion

To evaluate our prediction, we observe the Mean Absolute Error (MAE) on the temperature profile. In order to justify every increase in complexity in our approach, we first need to set up a baseline. We keep two different values for the number of modes used for the latent-class regression: $K \in \{3,4\}$. We compare the latent-class regression against a Linear model, a K-means initialization of EM algorithm, and a small Convolutional Neural Network. For the CNN approach, we use a small 8-layer architecture with two convolutional layer (32 filters of size 7×7) and a max pooling layer (2x2 filter and stride of 2), followed by two convolutional layer (64 filters of size 3×3), followed by a the same max pooling layer then by a flatten and a dense layer.

Table 1 summarizes the different results in terms of MAE. The linear model with variance kept above 99.9 % has good performance but it does not take into account difference between mass of warm water and cold ones (encoded by the modes). Although the CNN does not generalize well to the testing set, we notice from a seasonnal study that its predictions throughout the year are robust to seasonal variations compared to the basic linear model that would present seasonal errors. An entry point to improve CNN performance would be to use wider patches of SST and SLA. The latent-class regression has the best overall MAE although it contains misclassification errors (attributing a wrong label to a given X before applying the regression results in an error of about 3% of the MAE).

To fully exploit the potential of latent-regression, we plan to include the temporal aspect of ocean currents in the study: for instance, the profile classes can be tracked in space by applying a Kalman filter procedure with a simple evolution model to identify the class of a profile and an adequate linear coefficient. In that case of filtering method, we could keep the fuzzy value of belonging (probability density score) and work with fuzzy state of each profile.

5. Conclusion

In this study, we used a Latent Class Regression model that is an unsupervised classification - supervised regression

method on synthetic ARGO temperature profiles. Our goal was to explore to what extent this method would lead to a prediction of profile temperature sufficiently precise vertically and with a correct resolution horizontally to follow the eddies. We hence have developed several models to build the foundations of 3D ocean temperature field prediction from coupled surface and in-situ data. Although our results are not immediately applicable (since the filtering step is missing), they are a crucial step in evaluating latent regression methods for studying vertical ocean temperature fields. Finally, our approach is one more step to follow the diffusion and distribution of the heat in the ocean and thus to better understand global climate change.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

DeSarbo, W. S. and Cron, W. L. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5:249–282, 1988.

Maze, G., Deshayes, J., Marshall, J., Tréguier, A.-M., Chronis, A., and Vollmer, L. 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, 2013.

Maze, G., Mercier, H., Fablet, R., Tandeo, P., Radcenco, M. L., Lenca, P., Feucher, C., and Goff, C. L. Coherent heat patterns revealed by unsupervised classification of argo temperature profiles in the north atlantic ocean. *Progress in Oceanography*, 151:275 – 292, 2017. ISSN 0079-6611. doi: https://doi.org/10.1016/j.pocean.2016.12.008.

Riser, S., J. Freeland, H., Roemmich, D., Wijffels, S., Troisi, A., Belboch, M., Gilbert, D., Xu, J., Pouliquen, S., Thresher, A., Traon, P.-Y., Maze, G., Klein, B., Ravichandran, M., Grant, F., Poulain, P.-M., Suga, T., Lim, B., Sterl, A., and R. Jayne, S. Fifteen years of ocean observations with the global argo array. *Nature Clim. Change*, 6:145–153, 02 2016. doi: 10.1038/nclimate2872.

Schwarz, G. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136.

Tandeo, P., Chapron, B., Ba, S., Autret, E., and Fablet, R. Segmentation of mesoscale ocean surface dynamics using satellite sst and ssh observations. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7):4227–4235, July 2014. ISSN 0196-2892. doi: 10.1109/TGRS.2013.2280494.

Treguier, A.-M. Drakkar natl12 model configuration for a 27-year simulation of the north atlantic ocean. 01 2008.

Zhang, Z., Wang, Y., and Qiu, B. Oceanic mass transport by mesoscale eddies. *Science (New York, N.Y.)*, 345:322–4, 07 2014. doi: 10.1126/science.1252418.