

## Mining an Ocean of Data: Application of modern statistical methods for addressing biological oceanography questions

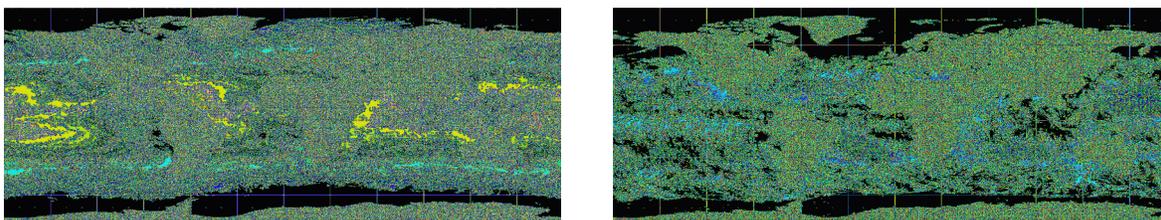
*Joaquim Goes<sup>1</sup> and Rahul Mazumder<sup>2</sup>*

*Lamont Doherty Earth Observatory, Palisades, New York, 10964*

*Department of Statistics, Columbia University, New York, 10027*

**Summary:** The explosive growth of oceanographic databases in recent years resulting from the advent of satellites and autonomous floats capable of measuring ocean properties beyond those possible by traditional research ships, demands that we find better methods of optimizing data analysis and interpretation for maximizing data use. As part of this proposal, we have identified two contemporary oceanographic research topics that could benefit immensely from the cross-fertilization of techniques and methods from natural and statistical sciences. One of these involves development of non-parametric statistical methods to merge and integrate in-situ depth profile oceanographic data from shipboard and ARGO floats, with sea surface data from satellites. The second involves the development and application of innovative statistical clustering and bioinformatic techniques to satellite derived geospatial ocean color and environmental data to identify phytoplankton functional types from space. The two topics are of immense relevance for ocean research and for understanding how ocean ecosystems and their biodiversity are being impacted by global warming and climate change.

**Background:** The oceans cover more than 70% of our planet and are critical to our environmental system because of their potential to store heat, sequester atmospheric carbon dioxide and influence major atmospheric weather events such as droughts, hurricanes and tornados. They are also essential to supporting a food chain that is important for world fisheries (Levitus et al., 2000) and high quality protein foed especially in the world's poorest countries (FAO, 2010). Until recently oceanographers relied on data collected during ship expeditions, many of limited duration. Although data poor, these expeditionary research approaches resulted in major advances in our understanding of global ocean circulation, heat and energy transport associated with mesoscale circulation, global ocean productivity, climate-ocean coupling, etc. While these and many other successes expanded our understanding of ocean processes, they also demonstrated the need for sustained sampling over a range of temporal and spatial scales essential to capture highly complex processes and necessary for unraveling how the oceans operate physically, chemically, and biologically as an integrated system. Within the field of ecosystem dynamics in particular, the coupling of biological processes to large-scale physical processes occurring at different spatial and time scales (e.g., mixing, mesoscale eddies, water



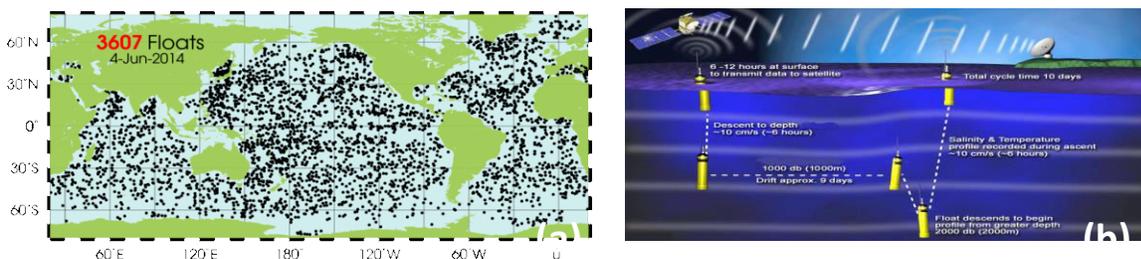
**Fig. 1** Weekly composites of satellite derived a) Sea Surface Temperature (SST) and b) Chlorophyll a for June 2013

mass transports, upwelling, ice dynamics, etc.) was not properly understood until the advent of satellites (Dickey & Bidigare, 2005). With the introduction of satellite remote sensing of the oceans in the 1970s, traditional oceanographers were provided with a new means to collect synoptic observations of surface global ocean conditions at unprecedented time and space scales

(Munk, 2000; Robinson, 2010). Over the years, with rapid advancements in the development of novel sensor technologies, improvements in data processing, communication and information transfer methods, etc., oceanographers are now in a position to obtain near real-time, global, high spatial and temporal resolution data of key variables such as sea surface temperature (SST), Chlorophyll a (a proxy for phytoplankton biomass and productivity), sea surface height (SSH), sea surface salinity, winds and ocean circulation, etc. almost continuously from space (examples in Fig. 1). As part of an ongoing NASA project, we are developing methods for estimating nutrients from space (Goes et al., 2000). These variables are critical for monitoring the role and response of ocean ecosystems to episodic events such eg. El-Niño and La-Niña, hurricanes and other tropical storms, or to long-term processes as they relate to the warming of our planet (Goes et al., 2005; Robinson, 2010).

Although oceanographic satellites are spatially robust and have provided high-resolution data for many oceanographic applications, most satellites are unable to see below one optical layer, making it impossible to directly measure any subsurface property. Recognizing the limitations of satellites to view the subsurface ocean, as well as the inadequacy of research vessels to provide data at spatial and temporal scales necessary for climate research, oceanographers have turned to unmanned technologies for sampling of the oceans (Roemmich et al., 2009). One of the most widely used technologies for collecting oceanographic data autonomously is the ARGO float which are part of a borad-scale array of temperature/salinity profiling floats. Starting with limited deployments, in the year 2000, the number of ARGO floats throughout the world's oceans has reached 3607 as of June 2014 (<http://www.argo.net>) (Fig. 2b).

The standard ARGO float operates in a park and profile mission. First the float descends to a target depth of 1000m to drift, and then again to 2000m from where it starts continuously recording the temperature and salinity of the water column (Fig. 2b). On ascending to the surface (typically over 10 day intervals) the floats transmit their location and data to shore via satellite communications. These freely drifting profiling now allow for continuous monitoring of the

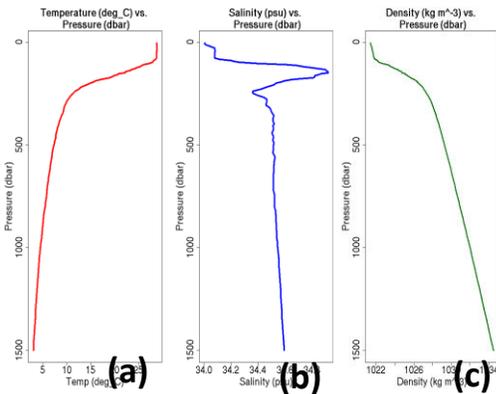


**Fig. 2a) Distribution of ARGO floats in the world's ocean, b) Schematic showing park and profiling mission of ARGO floats. The total number of floats in the ARGO network has reached 3607 floats as of 4<sup>th</sup> June 2014. (Courtesy [www.argo.ucsd.edu](http://www.argo.ucsd.edu))**

temperature, salinity, and velocity of different water masses in the upper 2000 m of the world's oceans. Efforts are currently underway to equip these floats with other sensors expanding their utility way beyond the temperature and salinity of the water column.

One of our proposed study areas involves development of methodologies to merge and integrate in-situ depth profile data from shipboard and ARGO floats with, remotely sensed data of sea surface properties. Our second study involves the development and application of statistical clustering techniques to data from ocean color satellites for identifying phytoplankton functional types from space. The availability of massive amounts of data from ARGO floats has opened the potential of extending satellite-derived ocean observations beyond the surface, to three-

dimensional (3D) representations of the global ocean. Accomplishing this task however, will require the development of techniques and methods that permit integration of surface data from satellites with sub-surface data collected by ships and by ARGO floats.



**Fig. 3 Typical profiles of a) temperature, b) salinity and c) density measured with an ARGO float**

There have been limited efforts in this direction mainly through the use of relationships between the density structure of different water masses and environmental variables measurable at the surface. When combined with pattern recognition techniques (Maji and Yahia, 2014), the ability to view the ocean in 3D will allow oceanographers to track the movements of water masses in different parts of the world's oceans. This is required for a better understanding of the relationships between atmospheric and oceanographic processes and their consequences for ecosystem dynamics. The ability to

combine and analyze massive amounts of data from satellites, ships and unmanned vehicles offers us the opportunity to greatly improve our capacity to understand and forecast oceanic variability, to understand how the oceans and atmosphere interact over time and space to influence life in the sea, climate variability and the overall health of our planet.

With oceanographic databases growing in size (from the giga to the peta-scale) each day; being updated with rapid frequencies (from the year/month to the day/hour scale) and involving complex structures and higher dimensionality (from one file/parameter to millions of files and dozens of parameters on multiple locations), there is a great deal of urgency to find ways for meaningful use of these massive and disparate amounts of data.

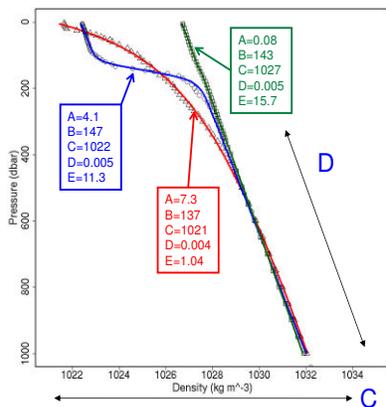
In the face of this ever mounting challenge, the cross-fertilization of Statistics and Oceanography appears to be an immensely promising area for inter-disciplinary research and for rapid advancements in oceanography. Environmental statistics has been utilized to address a wide variety of scientific uses, including air and water pollution, food security, conservation, epidemiology etc. Because environmental statistics covers a wide variety of statistical methodologies including time series, spatial statistics, spatio-temporal statistics, non-parametric, Bayesian statistics, generalized linear mixed models, decision theory, etc. it holds immense promise for collaborative research that involves the use of extremely large datasets (Wikle et al., 2013). A list of advancements in the field of oceanography detailed in Wikle et al. (2013), underlines the immense benefits of strengthening links between environmental statisticians and oceanographers, and emphasizes the value of cross-fertilization between these groups for resolving complex problems in earth sciences. In many of the works cited in Wikle et al. (2013), statistical "Bayesian hierarchical modeling" techniques were used to combine direct observations from a limited network of ocean buoys and ships with satellite images to improve predictions of sea surface temperature extremes and wind fields over the ocean. The ability to predict extreme swings in sea surface temperature and wind fields is the key to improving assessments of the frequency of hurricanes, storms as well as the distribution of plankton along the coasts of continental USA.

**Research goals and work plan:** As part of this proposal, we have identified two areas of oceanographic research both of which rely heavily on the use of large datasets and on statistical methods that will enable the integration of satellite, float and shipboard datasets for optimizing

their use. The first of these study areas involves development of methodologies to merge and integrate in-situ depth profile data from research ships and ARGO floats with, remotely sensed data of sea surface properties. The second involves the development and application of statistical clustering techniques to data from ocean color satellites to identify phytoplankton functional types from space. The overarching goal of these two study areas is to demonstrate the utility and benefits of cross-fertilization of statistical and natural sciences and potential the integration of the two fields offers to address complex oceanographic questions that would not be possible by current natural sciences methods alone. Currently most funding for oceanographic research from traditional sources such as NSF Geosciences or from NASA's Earth Sciences are directed more, towards process oriented studies rather than data mining and/or the development of data analysis methods.

**Goal–1 Development of statistical methods to integrate satellite, shipboard and float data:**

Oceanographic satellites are spatially robust and have provided high-resolution data for many environmental conditions that allow the use of relationships between the density structure of



$$\rho = \left( \frac{Az^E}{B^E + z^E} \right) + Dz + C$$

**Fig. 4 Profile of seawater density from ARGO float fitted with a modified Holling III curve. Parameters of curves were used to reconstruct 3-D density fields using satellite data. (Oliver and Irwin, 2008)**

different water masses and environmental variables measurable at the surface, coupled with pattern recognition techniques. This approach has been used in a previous study (Guinehut et al., 2004) which provides a theoretical proof-of-concept by reconstructing 200 m temperature fields using ARGO temperature profiles, satellite sea surface temperature (SST) and sea surface height (SSH) observations. Other studies (Nardelli & Santoleri, 2005) introduced multivariate empirical orthogonal functions and coupled pattern reconstruction techniques for reconstructing temperature profiles, also from satellite maps of SST and SSH by identifying variability patterns within the data set. While these studies focused on SST and SSH observations, recent work (Grossi, 2010; Oliver & Irwin, 2008) has demonstrated the potential application of this method for discriminating ocean biogeochemical provinces on global scales. In these studies, the authors relied on the application of parametric curve fitting techniques to ARGO float temperature data (Fig. 4). Parameters obtained from these curve fitting routines were then utilized to calculate vertical

profiles of water mass density from SST for different locations in the world's oceans and during different times of the year with a very low root mean squared error (RMSE) (Grossi, 2010). Using this approach of combining satellite observations of SST with information on the vertical structure of the oceans over time (Irwin & Finkel, 2008) it was shown that the large oligotrophic regions (ocean deserts) of the ocean were growing in size as a result of global warming. Time series analyses show of the density profiles showed significant trends of increasing water column stratification in these ocean deserts (Irwin & Oliver, 2009).

One of the limitations of curve fitting routines for ocean profile data is their inflexibility. A handful of parameters that are obtained are not universal to the dataset and at times inadequate descriptors of the shape of the profile especially if they are different. Our plan here is to cast the fitting as a 1-D function via non-parametric methods, a topic that has received significant attention in the statistical community over the past decades (Wahba, 1990; Wasserman, 2006).

Non-parametric procedures are more flexible; the parameter spaces may be potentially infinite dimensional (akin to flexibility) and one determines “best” functions that fit the data under some structural assumptions (akin to dimensionality reduction) like: smoothness, monotonicity of the function and other shape constraints like convexity and combinations thereof. The profiles of temperature and density (Figs 3a & c) for instance, can be described by smooth monotone functions (decreasing and increasing respectively) as a function of depth. Monotone function fits to a finite sample of data-points are typically non-smooth, i.e., piecewise linear (Ayer et al., 1955; Meyer & Woodroffe, 2000). The functions appearing in the data seem to be smooth in addition to being monotone. This makes the problem more challenging and requires procedures via which one can fit functions that are both smooth and monotone to data (Dette et al., 2006; Pal & Woodroffe, 2007; Ramsay, 1998; Turlach, 2005). Profiles of nutrients in seawater are akin to profiles of temperature and density and thus could be described by monotone functions. The profile of salinity shown in Fig. 3b is a mixture of a unimodal component, followed by a decreasing trend. As a first order approximation, we may use non-parametric splines, local linear regression type fitting procedures. It will be interesting to investigate how the fit may be improved using combinations of a unimodal function and a decreasing function with an unknown structural break-point that needs to be estimated from the data. The same holds for profiles of chlorophyll in the water column.

**Development of statistical methods for detection of PFTs from satellite data:** The oceanographic community remains heavily invested in developing methods to identify and discriminate Phytoplankton Functional Types (PFTs) from remotely-sensed ocean color data records because they are relevant proxies of ecosystem functioning and the response of ocean biology to climate change. PFTs are conceptual groupings of phytoplankton species, which have an ecological functionality in common (either in terms of the food web or biogeochemical cycles). Examples include atmospheric nitrogen fixers (e.g. *Trichodesmium*), calcifiers (coccolithophores), and silicifiers (e.g., diatoms), etc. These groupings are not necessarily related to physiological characteristics, but are often based on functionality and importance of the organisms to biogeochemical processes, for example, export of organic carbon to the deep ocean vs. local recycling (Sathyendranath et al., 2004).

Satellite ocean color instruments measure the spectral radiance emanating from the top of the atmosphere ( $L_t(\lambda)$ ;  $\mu\text{W cm}^{-2} \text{ nm}^{-1} \text{ sr}^{-1}$ ) at discrete visible and infrared wavelengths. Atmospheric correction algorithms are applied to  $L_t(\lambda)$  to remove the contribution of the atmosphere from the total signal and produce estimates of remote-sensing reflectances ( $R_{rs}(\lambda)$ ;  $\text{sr}^{-1}$ ), the light exiting the water mass normalized to a hypothetical condition of an overhead Sun and no atmosphere. Bio-optical algorithms are applied to the ( $R_{rs}(\lambda)$ ;  $\text{sr}^{-1}$ ) to produce estimates of marine geophysical properties, such as the near-surface concentrations of the phytoplankton pigment chlorophyll-a (Chl *a*;  $\text{mg m}^{-3}$ ), of organic carbon, particulate inorganic carbon, etc (Werdell & Bailey, 2005).

Apart from Chl *a*, each PFTs can be distinguished by its unique pigment type. Marine  $\text{N}_2$  fixers for instance, are rich in phycocyanin and phycoerythrin, while PFTs like diatoms and dinoflagellates are rich in the golden brown pigment fucoxanthin. Since these pigments absorb different wavelengths of visible light, the  $R_{rs}(\lambda)$ ;  $\text{sr}^{-1}$  signatures of the light emanating from the surface of the ocean varies with the PFTs present in the water. In recent years there have been a variety of bio-optical and ecological approaches designed to exploit these differences in the  $R_{rs}(\lambda)$ ;  $\text{sr}^{-1}$  signatures to derive PFTs from satellite ocean color fields (Ciotti & Bricaud, 2006; Hirata et al., 2008; Sathyendranath et al., 2004). These approaches can be broadly classified into two groups, the abundance methods, which assign empirically-derived thresholds on the

magnitudes of satellite derived on the  $R_{rs}(\lambda)$ ;  $\text{sr}^{-1}$  to each PFT, and the spectral group, which interpret radiometric spectral shapes for inferring PFTs (Werdell et al., 2005). In general, the former, exploit observed relationships between the trophic status of the environment and the type of phytoplankton expected to be present, whereas the latter exploit differences in the optical signatures of specific size classes or functional groups to distinguish between phytoplankton types. Unlike abundance methods, spectral methods can in principle discriminate between phytoplankton groups with common biomass, provided the groups present have contrasting optical signatures within the spectral bands detected. However spectral methods remain confounded by natural variations in the spectral characteristics of PFTs due to their growth status, nutrient availability, and ambient light history. In practice, most spectral methods do not make use of local oceanographic information such as measurements of photosynthetically active radiation, mixed layer depth's sea surface temperature, and nutrient concentrations, that can provide indications of phytoplankton physiological state.

One of the goals in this study is to develop a classification scheme for discriminating PFTs from satellite ocean color datasets using methods that take into account not only the spectral quality of the light emanating from the surface of the ocean, but information on the environment, such as SST, nutrient conditions obtainable from satellites. Our plans involve combining these maps of PFTs with other satellite data products that can aid in: 1) identifying water masses that a particular PFT is associated with, 2) establishing the physiological status of the PFTs and 3) the likely impact of the physiological status of a given PFT on the spectral signatures of surface waters that are visible from space. Once our methods for detecting PFT fields have been developed, we plan to use different boundary detection techniques (Maji & Yahia, 2014) to track the onset, evolution and fate of PFTs in the aquatic environment. The kinds of statistical techniques that we envisage using for the PFTs are multivariate clustering techniques and analysis of similarity techniques that take into account the spectral characteristics of each pixel within an image. We postulate that the problem will eventually reduce to that of comparing stochastic processes, corresponding to locations in terms of their spatial autocorrelation (Viladomat et al., 2014). For example, a spatial correlation equal to zero implies that there is no association between pixels within an image, whereas non-zero implies there is some similarity between the processes. We also intend to borrow approaches from bioinformatics, such as genetic expression algorithms and other objective mathematical constructs that are now routinely being used for transforming genetic data continuums into meaningful categories. Because of the inherent complexities associated with deriving PFTs from massive amounts of ocean color data, we will focus on the Arabian Sea from where we have obtained extensive shipboard biological, bio-optical and environmental data from our projects funded by NSF and NASA, water column profile data from ARGO floats, ocean color and SST data from MODIS-Aqua and other data from satellites. Our goal is to show that satellite ocean color data when used in conjunction with other environmental data from ships, floats and satellites can be used to confirm our recent findings that the Arabian Sea is experiencing a loss of biodiversity (Gomes et al., in review). Both the research topics identified, involve mining of massive oceanographic datasets and blending of ideas from the natural sciences as well as statistical sciences, both goals, achievable within a period of 16 months. Currently funding from traditional sources such as NSF Geosciences Division or from NASA's Earth Sciences Program for data mining or for the development of data analysis methods is unlikely as most projects supported by these agencies are geared towards process oriented studies. Two graduate students, the first Jan-Erik Tesdal, 1<sup>st</sup> year Graduate Student from the Dept. Earth Environmental Sciences, and the second, to be named from the Dept. of Statistics, will be involved in this project.

## References:

- Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W., Silverman, E., 1955. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 26, 641-647.
- Dette, H., Neumeyer, N., Pilz, K.F., 2006. A simple nonparametric estimator of a strictly monotone regression function. 469-490.
- Dickey, T.D., Bidigare, R.R., 2005. Interdisciplinary oceanographic observations: The wave of the future. *Scientia Marina*, 69, 23-42.
- FAO.2010.The State of World Fisheries and Aquaculture (2010). Rome, 197pp.
- Goes, J. I., Thoppil, P. G., do R Gomes, H., & Fasullo, J. T. (2005). Warming of the Eurasian landmass is making the Arabian Sea more productive. *Science*, 308, 545-547.
- Goes, J. I., T. Saino, J. Ishizaka, C.S. Wong, Y. Nojiri (2000) Estimating sea surface nitrate from space by compound remote sensing. *Geophysical Res. Letters*, 27, 1263-1265.
- Gomes, H. do R., J. I. Goes, S.G. Prabhu Matondkar, S. G. Parab, S. Basu & P. Thoppil (in review), Massive outbreaks of *Noctiluca* blooms in the Arabian Sea linked to the spread of hypoxia, *Nature Communications*.
- Grossi, M., 2010. Predicting three-dimensional ocean density structure from remotely sensed surface observations. University of Delaware.
- Guinehut, S., Le Traon, P., Larnicol, G., Philipps, S., 2004. Combining Argo and remote-sensing data to estimate the ocean three-dimensional temperature fields—A first approach based on simulated observations. *Journal of Marine Systems*, 46, 85-98.
- Hirata, T., Aiken, J., Hardman-Mountford, N., Smyth, T., Barlow, R., 2008. An absorption model to determine phytoplankton size classes from satellite ocean colour. *Remote sensing of environment*, 112, 3153-3159.
- Irwin, A.J., Finkel, Z.V., 2008. Mining a sea of data: Deducing the environmental controls of ocean chlorophyll. *PLoS One*, 3, e3836.
- Irwin, A.J., Oliver, M.J., 2009. Are ocean deserts getting larger? *Geophysical research letters*, 36.
- Levitus, S., Antonov, J.I., Boyer, T.P., Stephens, C., 2000. Warming of the world ocean. *Science*, 287, 2225-2229.
- Maji, S. K, Yahia, H. M (2014) Edges, Transitions and Criticality, Pattern Recognition, DOI :10.1016/j.patcog.2013.12.013
- Meyer, M., Woodroffe, M., 2000. On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, 1083-1104.
- Munk, W., 2000. Oceanography before, and after, the advent of satellites. *Satellites, oceanography and society. Elsevier*, 1-4.
- Nardelli, B.B., Santoleri, R., 2005. Methods for the Reconstruction of Vertical Profiles from Surface Data: Multivariate Analyses, Residual GEM, and Variable Temporal Signals in the North Pacific Ocean. *Journal of Atmospheric and Oceanic Technology*, 22, 1762.
- Oliver, M.J., Irwin, A.J., 2008. Objective global ocean biogeographic provinces. *Geophysical research letters*, 35.
- Pal, J.K., Woodroffe, M., 2007. Large sample properties of shape restricted regression estimators with smoothness adjustments. *Statistica Sinica*, 17, 1601.
- Ramsay, J., 1998. Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 365-375.
- Robinson, I.S., 2010. *Discovering the Ocean from Space: The unique applications of satellite oceanography*: Springer.
- Roemmich, D., Johnson, G.C., Riser, S., Davis, R., Gilson, J., Owens, W.B., Garzoli, S.L., Schmid, C., Ignaszewski, M., 2009. The Argo Program: Observing the global ocean with profiling floats.
- Sathyendranath, S., Watts, L., Devred, E., Platt, T., Caverhill, C., Maass, H., 2004. Discrimination of diatoms from other phytoplankton using ocean-colour data. *Marine ecology progress series*, 272, 59-68.
- Turlach, B.A., 2005. Shape constrained smoothing using smoothing splines. *Computational Statistics*, 20, 81-104.
- Viladomat, J., Mazumder, R., McInturff, A., McCauley, D.J., Hastie, T., 2014. Assessing the significance of global and local correlations under spatial autocorrelation: A nonparametric approach. *Biometrics*.
- Wahba, G., 1990. Spline models for observational data, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. *SIAM, Philadelphia*.
- Wasserman, L., 2006. *All of nonparametric statistics*: Springer.
- Werdell, P.J., Bailey, S.W., 2005. An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote sensing of environment*, 98, 122-140.
- Wikle, C.K., Milliff, R.F., Herbei, R., Leeds, W.B., 2013. Modern Statistical Methods in Oceanography: A Hierarchical Perspective. *Statistical Science*, 28, 466-486.