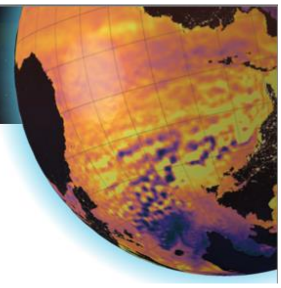

Enhancing BGC-Argo Chlorophyll-a Data Quality and Uniformity Using Machine Learning

Phytoplankton biomass, forming the foundation of the oceanic food web, is commonly estimated from the concentration of chlorophyll-a (Chla). In vivo chlorophyll-a fluorescence (fluo), a proxy for Chla, is one of the most frequently measured biogeochemical properties in the ocean, especially since the integration of miniaturized fluorometers onto autonomous platforms such as BioGeoChemical-Argo (BGC-Argo) profiling floats. Over the last decades, the number of fluo profiles has more than doubled thanks to increasing deployments of these autonomous platforms compared to historical measurements from oceanographic vessels, establishing BGC-Argo as a cornerstone of the global biogeochemical observing system. However, the conversion of fluo into Chla is not straightforward as it is influenced by multiple factors, including the composition and physiological status of phytoplankton communities. Accurate calibration of fluo into Chla is therefore both challenging and critical for optimal utilization of the rapidly increasing amount of fluo data. Significant efforts are being made by the Argo Data Management Team (ADMT) to calibrate and qualify fluo data measured from BGC-Argo floats to deliver Chla with the highest possible accuracy. Despite these efforts, the current BGC-Argo Chla dataset encompasses large regional biases, particularly in high latitude environments such as the Southern Ocean. Recent advances in machine learning methods have shown significant potential to improve the global and regional quality and coherence of the BGC-Argo Chla dataset. In particular, these techniques offer new tools to estimate poorly resolved or unmeasured variables (e.g., radiometry), allowing a better and systematic calibration of fluo into Chla, and provide an innovative framework for evaluating the accuracy of the fluo-derived Chla data against the most accurate in situ reference Chla measurements. Thus, using machine learning, the errors associated with Chla estimates can be reduced by more than 100% in the Southern Ocean. Moreover, the new framework to evaluate the accuracy of the BGC-Argo Chla dataset at global and regional scales is crucial for better defining uncertainties in the BGC-Argo dataset. In conclusion, this study highlights the potential of machine learning-based techniques to enhance the BGC-Argo global Chla dataset. Providing high-quality BGC-Argo data is essential not only for scientific research but also for operational oceanography. BGC-Argo is becoming a fundamental program for data assimilation into biogeochemical models, making the accuracy of its data crucial for the development of future digital twins of the ocean.

R. Sauzède (Institut de la Mer de Villefranche, CNRS/Sorbonne Université) , C. Schmechtig (OSU ECCE Terra, CNRS/Sorbonne Université), P.R. Renosh (Laboratoire d'Océanographie de Villefranche, CNRS/Sorbonne Université), J. Uitz (Laboratoire d'Océanographie de Villefranche, CNRS/Sorbonne Université) and H. Claustre (Laboratoire d'Océanographie de Villefranche, CNRS/Sorbonne Université)