Institut des sciences du calcul et des données
SORBONNE UNIVERSITÉ

Job Offer

Job Summary

Title, Job Position	**Deep learning to predict plankton communities from space (PhD thesis)**
Research Field	Data science, Numerical Ecology, Remote sensing
Employer	Sorbonne Université, Institut des sciences du calcul et des données
Location:	LOV, Villefranche sur mer, France LPSM, Paris, France
Application Deadline / Timezone	18 June 2023
Type of Contract	36 months
Job Status	Full-time
Starting Date	01-10-2023

Hiring Organization

The student would work in Sorbonne Université, funded by the Institute of Computing and Data Science, between the sites of Villefranche-sur-mer (on the French Riviera, near Nice) and Paris, as well as collaborate with private companies (FOTONOWER and ACRI-ST).

Sorbonne Université (SU) was created on January 1st, 2018 from the merger of Paris-Sorbonne and Pierre and Marie Curie (UPMC) universities. As a public institution, it fulfills the public service calling of French higher education, research and innovation. SU is a multidisciplinary and research-intensive university with world-famous origins. The University's 53,500 students, 3,400 professor-researchers and 3,600 administrative and technical staff members who help it run every day contribute to a University that is diverse, creative, innovative, and with a global outlook.

The **Institute of Computing and Data Sciences** (ISCD; http://iscd.sorbonne-universite.fr/) is dedicated to exploring and developing the potential of computational and data-driven research and training across science, humanities and medicine at Sorbonne Université. The ISCD hosts the FORMAL team gathering oceanographers, mathematicians, and computer scientists to study the dynamics of life in the ocean (http://iscd.sorbonne-universite.fr/research/sponsored-junior-teams/formal-2/)

The **Laboratoire d'Océanographie de Villefranche** (LOV, Villefranche-sur-mer; https://lov.imev-mer.fr/) has expertise in plankton ecology and the associated data from HPLC and imaging, satellite imagery, some of the existing algorithms to predict phytoplankton community from space and general knowledge in machine learning and deep networks.

The **Laboratoire de Probabilités, Statistique et Modélisation** (LPSM, Paris; https://www.lpsm.paris) has expertise in mathematics, deep networks and particularly high-performance computing on GPUs.

The **FOTONOWER** company (Paris) has expertise in machine learning and computer vision, using CNNs, applied to environmental questions. The **ACRI-ST** company (Antibes) has a long history of working with satellite imagery to derive new fields at large scale.

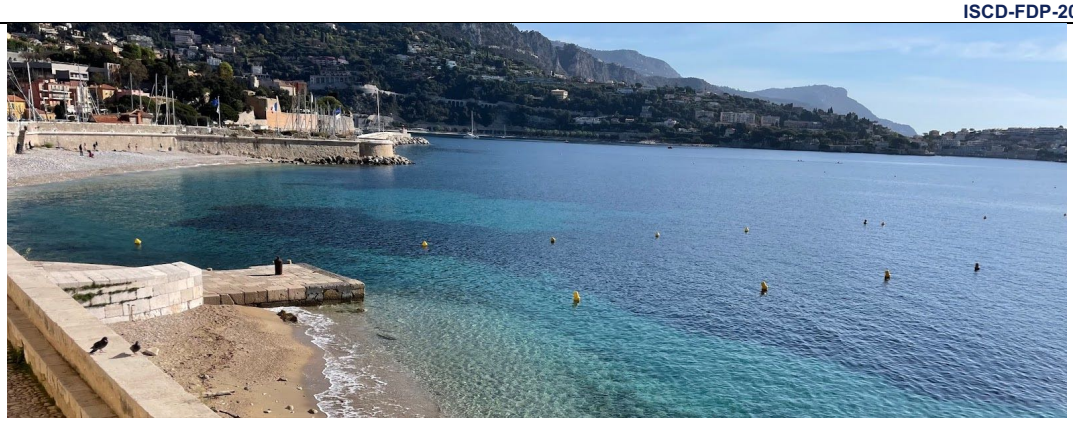

Fig 2: The pier of the Laboratoire d'Océanographie de Villefranche.

More specifically, the persons involved in the project would be:

- *Jean-Olivier Irisson, Associate Professor, LOV.* J-O is a computational ecologist, studying zooplankton through quantitative imaging and data science methods, in particular CNNs. He is particularly interested in the influence of meso and submesoscale processes on plankton.

- *Raphaëlle Sauzède, Research Engineer, IMEV.* Raphaëlle computes global ocean products merging satellite and in situ information through deep networks. She also has expertise on HPLC data, satellite imagery and phytoplankton distribution.

- *Lokmane Abbas-Turki, Associate Professor, LPSM.* Lokmane is one of the first academics to have programmed on GPUs and uses this skill to solve complex computational problems, in particular in finance.

- *Victor Retenauer, FOTONOWER.* Victor carried out a PhD in probabilities applied to finance and deep learning and started a company that uses machine learning tools for environmental applications, from garbage collection to plankton and beehiving.

Offer Description

Scientific context

Marine **plankton**, the organisms drifting along ocean currents, comprise hundreds of thousands of taxa and over 116 million genes. They play **crucial roles** in the functioning of the Earth: they contribute half of the photosynthesis on the planet, store carbon over climatic scales through the biological carbon pumps, and are a source of food for the rest of the marine ecosystem. Knowledge of their **distribution** and diversity in relation with the physical and biogeochemical conditions of their environment is therefore **essential**. It would allow us to better estimate their **biomass** and their roles in **biogeochemical cycles**. In the context of **global changes** in water masses, it would help anticipate changes in the planktonic communities they host.

However, because most planktonic organisms are small and diluted, estimating their distribution at the scale of the world's oceans is **difficult**. They are usually sampled from ships, using nets and bottles. The resulting samples are sorted back in the lab, in a time-consuming process that requires deep expertise. Some **semi-automated methods** accelerate this process (often at the expense of taxonomic resolution). For phytoplankton (vegetal), **High Performance Liquid Chromatography** from filtered sea water samples provides concentrations of pigments that allow estimating the total biomass and broad composition of the sample. For zooplankton (animal), **intelligent cameras**, deployed from ships or mounted on automated robots, take vast amounts of images and extract objects of interest that can be classified to quantify the composition of the community.

Those samples are still few and far apart compared to the size of the oceans and are only representative of a given point in space and time. Conversely, **satellites** take images of the ocean frequently and over wide swaths. Ocean colour satellites, for example, can provide **near-complete coverage** of the surface of the ocean every week. From the signal they record, over multiple wavelengths, new quantities such as total chlorophyll concentration, amount of particulate matter, or even phytoplankton community

composition can be inferred [Uitz 2006] and, this time, over the world's ocean. To do so, **regression models** are calibrated from satellite records co-located with in situ samples [Sauzède et al 2016]. Those models often exploit optical properties to physically relate the reflectance recorded by the satellite to the composition of the sample. For zooplankton, no such model exists but similar, **machine-learning based**, regressions have been carried out [Drago et al 2022].

More recent **deep learning** techniques should also help predict the distribution of plankton. To predict communities, it is first necessary to take the **correlations** among the concentrations of different species into account, since they are a property of the ecosystem. To do so, the **choice of the loss function** is critical. Instead of specifying it fully, which may not be possible, a **Generative Adversarial Network** (GAN; [Gui et al 2023]) can be used, pitching a generator network that predicts composition based on satellite data against a discriminator network, which actually sets the loss function for the generator. Furthermore, the breadth of data at our disposal makes **transfer learning** strategies important so that already trained models can be reused [Weiss et al 2016]. For example, a model trained for phytoplankton could be fine-tuned for zooplankton; a model trained on one time point can provide a good initialisation for other time points; etc. Finally, because the statistical distribution of plankton concentrations is typically long tailed (i.e. contains **very extreme values**), it will be important to challenge models using the asymptotic results from extreme value theory [de Haan and Ferreira 2006] or non-asymptotic analysis [Duttweiler 1973].

The goal of this PhD thesis is to predict the composition of phyto- and zoo-plankton communities worldwide and dynamically by leveraging databases of thousands of semi-automated in situ measurements, diverse sources of satellite imagery and advanced deep learning techniques.

The **versatility** of this subject, in terms of the possible **developments** in mathematics/computer science, of the variety of **data** available and of the possible **applications** to various questions in oceanography, makes it a great opportunity for Master 2 students that seek to gain expertise in a very current research topic and continue their research career either in the public sector or in the private one.

PhD project

A key point for this project is the recognition that the community of plankton in a given parcel of water cannot be predicted from the properties of that parcel alone and, instead, necessitates to **take its context into account**. Indeed, the space and time scale of the population dynamics of phyto- and particularly zooplankton go beyond what happens in the 4×4 km pixel imaged at a given instant by an ocean colour satellite. Indeed, these **communities are shaped by mesoscale processes** such as fronts or eddies, which accumulate biomass or stimulate its production, and by their dynamics, which stir and elongate plankton patches over scales of dozens of kilometres. Pixels at the core of a patch, at the edge of an eddy or at the extremity of a filament have different histories and likely host different communities. Therefore, it is necessary to detect the spatial and temporal context of the pixel of interest.

Now, instead of regressing a vector of concentration of plankton groups (Y) on a vector of values of reflectance at given wavelengths (X_λ), we now need to regress it on a **four-dimensional hypercube** of reflectance ($X_{\lambda,x,y,t}$), which has dimensions **latitude and longitude** (x, y) and **time** (t), in addition to **wavelength** (λ; Fig 1). To capture the spatiotemporal structure within this hypercube while keeping the size of the regression model manageable, we can summarise the input through convolutional and pooling layers, before entering a Multi-Layer Perceptron, hence creating a Convolutional Neural Network. Similarly, the input can be cut into smaller patches, the relationships and relative importance of which would be captured by a Transformer-based architecture.

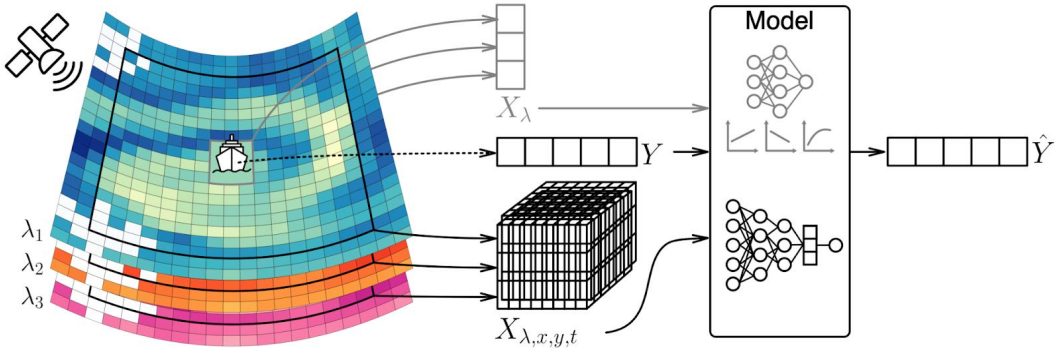

Fig 1: Predicting plankton composition sampled by a ship (Y) from reflectance at different wavelengths measured by satellite ((X_λ). Instead of using values from a single pixel (grey pathway) we propose to consider the context of that pixel in space (x, y) and time (t) hence working with the hypercube $X_{\lambda,x,y,t}$ as input (black pathway). Summarising this input for regression can be done through a Convolutional Neural Network or a Vision Transformer instead of a simple Multi-Layer Perceptron or multiple regressions.

The PhD student will work to collate the thousands records of phyto- and zooplankton communities available from databases hosted at LOV, to split them in appropriate training and test sets, to set up the modelling infrastructure, to train models for phytoplankton and zooplankton and finally to exploit the dynamic, worldwide fields that the models will produce to provide new insights on plankton biomass and/or diversity.

More details are available upon request.

Bibliography

Drago L, Panaïotis T, Irisson JO, Babin M, Biard T, Carlotti F, Coppola L, Guidi L, Hauss H, Karp-Boss L, Lombard F. Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning. Frontiers in Marine Science. 2022 Aug 9;9.

Duttweiler DL. The mean-square error of Bahadur's order-statistic approximation. The Annals of Statistics. 1973 May 1:446-53.

Gui J, Sun Z, Wen Y, Tao D, Ye J. A review on generative adversarial networks: Algorithms, theory, and applications. IEEE transactions on knowledge and data engineering. 2021 Nov 23.

Haan L, Ferreira A. Extreme value theory: an introduction. New York: Springer; 2006 Jul.

Sauzède R, Claustre H, Uitz J, Jamet C, Dall'Olmo G, d'Ortenzio F, Gentili B, Poteau A, Schmechtig C. A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient. Journal of Geophysical Research: Oceans. 2016 Apr;121(4):2552-71.

Uitz J, Claustre H, Morel A, Hooker SB. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. Journal of Geophysical Research: Oceans. 2006 Aug;111(C8).

Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016 Dec;3(1):1-40.

Profile Requirements

Required Education Level

The PhD is at the intersection of **marine ecology** and **data science** and we would welcome Masters students from both backgrounds. While the topic offers the potential for innovations in the context of deep networks, the results are geared towards the application in marine ecology (i.e. would be mainly published in marine ecology journals).

Students with a background in **marine ecology** would be expected to know the general ecology of phyto- and zoo-plankton, the functioning and role of mesoscale processes, the variables used to characterise the

surface of the ocean and, ideally, have some experience with satellite imagery, including ocean colour.

Students with a background in **data science** would be expected to know the principles and tools of machine learning, the functioning of Convolutional Neural Networks and/or Transformers in particular, possess enough mathematical knowledge to understand some details of them, as well as, ideally, some experience with the definition and implementation of custom processes within a deep network.

Students from a given background must be interested in **learning** what is specified in the **other background**, through discussion with the supervision team and independent reading. All candidates must have **strong computer programming skills** and the ability to manipulate **large datasets**.

Required Languages

Scientific & technical English (B2 level for written and oral). French would be a plus, but not mandatory.

Work Location

Institutes

Institut des sciences du calcul et des données. Project-Team FORMAL.

Country: France

Location: LOV, 181 chemin du Lazaret, 06230 Villefranche sur mer

How to apply?

Required Application Materials

1. Cover letter with research interests and motivation for the topic
2. Most recent curriculum vitae
3. Names and contact for one to two referees

How to submit

Interested candidates can contact jean-olivier.irisson@imev-mer.fr for more information and should submit the required application materials by email at the same address with the title “ISCD PhD Application - FORMAL”.

Selection Procedure

Selection process

Candidates are evaluated by faculty reviewers. Reviewers will evaluate candidates according to their academic accomplishments and their potential for research.

The selection process is based on an evaluation of the submitted material followed by an interview with short-listed applicants.