Internship offer

Machine Learning for lake pollution forecast

Advisers: David Métivier, Isabelle Sanchez and Céline Casenave (INRAE researchers)

Length: 4 to 6 months

Location: UMR MISTEA (Mathematics, Informatics, and Statistics for Environment and Agronomy), 2 Pl. Pierre Viala, 34000 Montpellier, France

Teasing: Build fancy Machine Learning models (e.g., GAN, Physics Informed Machine Learning) for a real life environmental application with real data.

Contact: Please send your application with a CV and a few motivation lines to: David Métivier (david.metivier@inrae.fr), Isabelle Sanchez (isabelle.sanchez@inrae.fr) and Céline Casenave (celine.casenave@inrae.fr). Do not hesitate to ask questions if you are interested.

Context

More than half of the freshwater lakes and rivers of the world are polluted. The World Health Organization (WHO) declared microbial hazards, such as toxic cyanobacteria, to be "of public health importance". Cyanobacteria are photosynthetic bacteria playing a key role in the life cycle: they are primary producers, some of them can fix the atmospheric nitrogen, others form symbiotic relationships with other organisms e.g., plant, fungi, etc. However, their proliferation can also be harmful for aquatic ecosystems. Indeed, the population of cyanobacteria can grow very quickly and accumulate on the water surface, forming scums (see picture). This phenomenon, known as algal bloom, has important economic, ecological and health consequences.

In this context, it is important to develop some operational tools for the management of lake ecosystems. A first objective is to propose a short-term prediction tool that can be used as a warning system. The forecasting tool will enable managers to anticipate these restrictions and alert users. Such a forecasting tool has been developed for the experimental lake of Champs-sur-Marne (Île-de-France) and can be seen on `https://balneau-leesu-rec.enpc.fr/#!/dashboard` where three variables of interest are predicted three days ahead: water temperature, phytoplankton concentration and cyanobacteria concentration.

Objectives

The objective of this internship will be to improve the quality of the forecast using machine learning based methods.

As a starting point for the internship, we would like to build a Conditional Generative Model to predict the future distribution of the variable of interests conditionally on the past observations. Recent examples of GAN for time series generation can be found in the literature, e.g., [YJv19, EHR17, HR20, SS20] and will be the starting point of the internship. This will allow critical

prediction such as, **in three days with high probability, the lake will not be suited for swimming due to dangerous cyanobacteria concentration**.

Depending on the progress, we might also consider using the physical models available on the lake into Physical Informed Machine Learning methods.

The model will be tested and trained on the experimental lake of Champs-sur-Marne (Île-de-France) where high-frequency measurements (meteorology, water temperature, water quality) are collected by the LEESU (Water, Environment, and Urban Systems Laboratory, École des Ponts ParisTech) since 2017.

Required skills

The applicant should be a master student with skills in Machine Learning and taste for modeling and numerical simulation. A good level in programming is also required. DM is a **Julia** enthusiast, IS is an **R** expert and CC uses **Matlab**. All have experience with **Python**. We can discuss coding language together (we are open to different options).

Knowledge in ecology biogeochemistry and/or French is not required.

Terms of the internship

The duration of the internship will be at least 4 months up to 6 months and could start as early as February 2023 depending on the student's availability. The student will benefit from an internship grant (around 600 euros/month) as well as a reduced canteen rate.

References

[AOZ+09] Rita Adrian, Catherine M. O'Reilly, Horacio Zagarese, Stephen B. Baines, Dag O. Hessen, Wendel Keller, David M. Livingstone, Ruben Sommaruga, Dietmar Straile, Ellen Van Donk, Gesa A. Weyhenmeyer, and Monika Winder. Lakes as sentinels of climate change. *Limnology and oceanography*, 54(6):2283–2297, November 2009.

[EHR17] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs, December 2017.

[HR20] Moritz Haas and Stefan Richter. Statistical analysis of Wasserstein GANs with applications to time series forecasting, November 2020.

[SS20] Kaleb E. Smith and Anthony O. Smith. Conditional GAN for timeseries generation, June 2020.

[YJv19] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.