Phd proposal

End-to-end learning of geophysically-sound CNN representations from satellite-derived observation datasets to better inform sea surface dynamics

Phd Supervisor: Ronan Fablet, Prof. IMT Atlantique, Lab-STICC/TOMS, ronan.fablet@imt-atlantique.fr

Hosting research team: Lab-STICC/TOMS, AI Chair OceaniX (link)

Collaborators:

- J. Le Sommer, Scientist, CNRS, IGE
- B. Chapron, Senior Scientist Ifremer, Ifremer, LOPS
- A. Pascual, Senior Scientist, IMEDEA

Abstract. Artificial Intelligence (AI) technologies, models and strategies open new paradigms to address the modeling, simulation, forecasting and reconstruction of complex systems, including ocean-atmosphere dynamics. Due to the irregular space-time sampling of in situ and spaceborne observation data, most envisioned AI-driven strategies rely on learning representations from simulation data and applying these representations to observation data to inform the processes of interest. The applicability of such schemes may then strongly rely on the ability of simulation data to truly match observation data features, which may be questioned for numerous processes. The general objective of this PhD is to investigate the extent to which one may develop **fully observation-driven schemes in the context of the space-based observation of flying and future satellite missions such as SWOT mission**. From a methodological point of view, we propose to state this challenge for given geophysical processes or variables as a **joint end-to-end learning of a latent geophysically-sound representation and of the associated inversion scheme from irregularly-sampled observation dataset**. Using CNNs (Convolutional Neural Network), The taregted methodological contributions are regarded as key building blocks to revisit earth observation challenges, including among others operational satellite-derived geophysical products, data-driven schemes for inter-comparison studies between observation and/or simulation data,...
This PhD is part of the PhD program of AI Chair OceaniX (link) and will contribute of the collaborative framework of **Melody project (ANR MN 2020-2022)** with strong interactions between **Lab-STICC** (R. Fablet), **LOPS** (B. Chapron), **IGE/MEOM** (J. Le Sommer), **OceanNext** (C. Ubelman) and **OceanDataLab** (L. Gaultier). SWOT-related case-studies in Melody, e.g. wave-current separation, SWOT-derived SLA L4 products, will be the core application ground for the considered methodological developments, including the exploitation of SWOT fast-sampling phase data. The PhD candidate will benefit from the gathered multidisciplinary expertise of the supervision team in Ocean Science, Ocean Remote Sensing, Fluid Dynamics and Data Science.

Keywords: Observing systems, space oceanography, SWOT, irregularly-sampled data, data-driven methods, CNN, end-to-end learning, latent representations, inverse problems.

Scientific context and objectives

Understanding, modeling, forecasting and reconstructing fine-scale and large-scale processes and their interactions are among the key scientific challenges in ocean-atmosphere science. State-of-the-art approaches strongly rely on joint research effort in observing systems (e.g., in situ monitoring, satellite observations) and numerical simulations [e.g., 30]. The ability to relate models and observation data, though significant advances in data assimilation, remain open questions for numerous processes (e.g., small-scale parameterization, ocean-atmosphere interactions, biogeochemical ocean dynamics, climate-scale dynamics). Artificial Intelligence (AI) technologies, models and strategies open new paradigms to address these questions from the in-depth exploration of the existing observation and simulation big data [1,3,12,20-26].

Among others, the recent breakthrough in the resolution of fine-scale cloud processes in climate models [26] is a striking illustration. It further illustrates the typical learning-based paradigm for ocean-atmosphere processes, where a model or representation is learnt from simulation data. However, for numerous processes, on the one hand, the ability of model simulations to be fully representative of real dynamics is questionable and, on the other hand, one would expect to benefit from the existing observation datasets to extract computational representations. **The sampling patterns of these observation datasets** (e.g., irregular space time-sampling, partially-observed system,...) **raise issues which remain to be addressed to develop fully observation-driven and learning-based frameworks for earth science, including space oceanography.**

In this context, the general goal of this project is to address the following topical questions :

- **Can we develop fully-observation-driven learning-based paradigms from satellite-derived observation dataset, including synergies with other observation data (e.g.,** ARGO floats, buoys,...) **?**
- **Can learning-based paradigms better inform past and future dynamics from HR satellite-derived observations of the sea surface ?**

The methodological backbone underlying these topical questions is the **definition and identification of learning-based representations of geophysical dynamics**. In the framework of ANR project Melody[1] (2020-2023) and SWOT ST DIEGO[2] project, the proposed methodological framework will be demonstrated and implemented in the context of incoming SWOT mission towards informing past and future sea surface dynamics from HR SWOT snapshots. Case-studies will be designed based on OSSEs (Observing System Simulation Experiment) and real SWOT data.

Proposed approach

To address the topical objectives defined above, a key issue is the choice of the underlying representations of geophysical processes which we will rely on to develop and implement learning-based strategies. Neural network (NN) representations appear as a natural state-of-the-art choice for their genericity and computational efficiency [8,13,15,21-27]. Especially, neural ODE/PDE (Ordinary/Partial Differential Equation) and energy-based representations [8,13,23-25] are examples of state-of-the-art representations, which naturally arise as relevant representations for geophysical processes. Here, following our recent work [10], we propose to consider **energy-based representations** as they naturally apply to solving inverse problems as minimization issues, classically

$$(1)\ \widehat{X} = \arg\min_X U_P(X;\theta_P) + U_{Obs}(X,Y;\theta_{Obs})$$

where X is the state of interest, e.g., 2D, 2D+t and/or multivariate 2D+t fields, and Y an observation. U_P is a prior energy which encodes knowledge on X and U_{Obs} typically relates to an error criterion for some observation operator. θ_P and θ_{Obs} are respectively the parameters of these two energy priors. We may emphasize that data assimilation and optimal interpolation [6,7] involves such energy minimization formulations. Recently, we have shown that NN can embed energy priors U_P along with

[1] ANR Melody (2020-2023): Bridging geophysics and MachinE Learning for the modeling, simulation and reconstruction of Ocean DYnamics, PI: R. Fablet (Lab-STICC), co-PIs : L. Drebreu (INRIA GRA), P. Naveau (LSCE), J. Le Sommer (IGE).

[2] DIEGO (Data and dynamIcal synErGies for swOt): proposal to be submitted to SWOT ST call for proposals (PI: A. Ponte, Ifremer/LOPS) including a WP on learning-based SWOTdata analytics (co-PI: R. Fablet, IMT Atlantique, Lab-STICC).

the solver of the above minimization (for a known observation operator) such that we may state the following end-to-end learning issue

$$\widehat{\theta_P} = \arg\min_{\theta_P} \sum_i \left\| Y_i - \mathcal{S}_{\theta_P}\left(Y_i, U_P(;\theta_P), U_{Obs}(;\theta_{Obs})\right)\right\|^2$$

where $\mathcal{S}_{\theta_P}()$ is a NN-based solver[3] for minimization (1), i refers to the index of the training samples and the squared norm is only evaluated on the observed domain when considering irregularly-sampled training data as in [10].

Based on this general framework, this project will address the following questions:

- **Q1. Which geophysically-sound neural-network representations for sea surface dynamics ?**

  Within the proposed energy-based setting, we will explore different types of representations which amount to considering different parameterizations for energy U_P. Two categories of parameterizations will be of particular interest: parameterizations based on auto-encoders which generalize EOF [6] and parameterizations based on ODEs/PDEs. It may be noted that within the proposed energy-based setting the second category relates to Gibbs models, which are classical models in statistical physics. It may be noted that the considered energy-based settings can naturally embed multivariate energy terms, which may be interpreted as jointly accounting for different priors such as an underlying ODE and an energy conservation property or multi-scale/multi-source decompositions. Besides energy prior U_P, the definition of the associated NN-based solver for minimization (1) will also be considered. In addition to generic gradient-based algorithm using automatic differentiation tools embedded in NN framework, computationally-simpler and explicit iterative schemes will also be explored as proposed in [10]. We may point out that, for a trained energy prior U_P, the jointly learnt NN-based minimizer can be applied to any new observation dataset to produce gridded and interpolated fields. We may further emphasize that the trained representations (prior U_P and NN-based solver for minimization (1)) are expected to provide generic/reusable components of interest for other tasks than the one on which they were trained. This aspect is regarded as a key challenge to broaden the impact and exploitation of data-driven representations beyond a specific task.

- **Q2. Can we develop fully-observation-driven learning-based paradigms from satellite-derived observation dataset, including synergies with other observation data (e.g.,** ARGO floats, buoys,...) **?**

  We will first apply the proposed energy-based framework to the learning of computationally-efficient representations of sea surface dynamics from irregularly-sampled observation datasets, which are typical of satellite-derived L2 or L3 products due to the orbit, revisit time and swath of the satellite as well as to the sensitivity to atmospheric conditions (e.g., clouds, heavy rains…) as illustrated below. The objective here will be to develop and implement the proposed energy-based framework introduced above so that both the NN-based energy prior U_P and the associated solver for minimization (1) can be jointly learned from gathered observation datasets. Transfer learning (e.g, using some pre-trained model as initialization) may also be investigated. Using OSSEs, we will design benchmarking experiments according to different metrics (e.g., reconstruction and forecasting performance). Besides the quantitative comparison of the proposed learning-based schemes with state-of-the-art model-driven and data-driven ones, these benchmarking experiments will also allow us to evaluate observation-only learning strategies in the context of flying/incoming satellite missions. Similarly, they will provide the basis for the evaluation of the added-value of multimodal[4] synergies during the learning stage and/or the reconstruction issue.

[3] We may refer the reader to [10] for a more detailed description of the considered framework, including the design of NN-based solvers for minimisation (1) using gradient-based or fixed-point schemes.

[4] By multimodal, we refer here to both to multi-tracer/multi-sensor dataset, including satellite-derived and in situ data, as well as synergies between observation and simulation data.

- **Q3. Can learning-based paradigms better inform past and future dynamics from HR satellite-derived observations of the sea surface ?**

  The second topical objective aims to address the conditional reconstruction and/or simulation of past and future from satellite observations of sea surface geophysical fields. Here, the key idea is to apply the proposed energy-based framework while considering states X which decompose into different components (e.g., different tracers, multi-scale decomposition). Given a trained energy prior U_P, the conditional reconstruction and/or simulation will be stated as solving for minimization (1) where observation Y is only available for one or a few time steps. A typical example will be the conditional forecasting of the SLA (Sea Level Anomaly which relates to geostrophic sea surface currents) given HR SLA and SST (Sea Surface Temperature) snapshots. We may point out that observation term in (1) may be different from the one used during the training of energy prior U_P. Besides, GAN learning strategies[5] [13] may be here highly-relevant to deliver realistic simulations.

The demonstration and evaluation of the proposed objectives and approach will rely on and contribute to **case-studies designed in the framework of ANR project Melody with a focus on the incoming SWOT mission.** SWOT mission will provide for the first time 2D snapshots of ocean circulation with a spatial resolution of ~15km. Beyond **strong collaborations with physical oceanography teams** (IGE: J. Le Sommer; LOPS: B. Chapron; IMEDEA: A. Pascual), the focus given to SWOT mission is further supported by the participation to DIEGO SWOT ST proposal[2]. Envisioned case-studies comprise both the characterization and reconstruction of sea surface currents and the separation of wave and current signatures in SWOT data. Data synergies of interest include: SWOT data and nadir altimetry data, SWOT data and HR snapshots of passive tracers (e.g., SST, ocean colour), SWOT data and in situ current data (e.g., drifters). **OSSEs using SWOT simulator from NALT60 data** [30] will be initially considered. From the second year of the PhD, we aim to apply and evaluate the proposed framework to real SWOT data. Especially, we expect SWOT data acquired during the fast-sampling phase to be of key interest both for learning purposes as well as to complement OSSE-driven evaluation with cross-validation experiments with real data. In this respect, the ongoing collaboration with IMEDEA (A. Pascual) may be highly relevant to complement the proposed case-study with additional in situ data (e.g., glider data, cruise data).

Proposed workplan

The planned workplan is as follows:

M1-M2	Review of the state-of-the-art (deep learning, inverse problems, SWOT mission)
M3-M12	Development of the proposed methodological framework (Q1) Design and implementation of SWOT OSSE case-studies
M13-M24	Focus on Q2 Application to SWOT fast sampling phase data
M25-M30	Focus on Q3 Application to real SWOT data
M31-M36	Preparation of PhD manuscript

Supervision

This PhD proposal will be supervised by R. Fablet (Prof. IMT Atlantique, Lab-STICC, Brest). The PhD candidate will strongly benefit from the interdisciplinary environment established in Brest between LOPS (B. Chapron) and Lab-STICC teams in the framework of EUR ISBLUE (Interdisciplinary graduate School for the Blue planet). This also includes interactions with SMEs and larger companies (e.g., OceanDataLab, Eodyn, CLS). Through Melody project, the PhD shall also interact and

[5] GANs are Generative Adversarial Networks [13]. They are among the state-of-the-art deep learning models for simulation purposes.

collaborate with IGE and OceanNext (J. Le Sommer). Especially, short-term visits in Grenoble will be planned for the design and implementation of SWOT OSSEs shared with MEOM/OceanNext. In the framework of OSTST MANATEE, a collaboration with IMEDEA (A. Pascual) is also envisioned regarding the SWOT case-study region

Expected contributions and outreach

We expect from this PhD proposal different types of contributions:

- **Scientific contributions in terms of journal papers and conference communications** in the field of **data science, remote sensing and space oceanography** corresponding both to new learning-based schemes and models for the analysis and reconstruction of spatio-temporal processes, and more particularly sea surface dynamics;
- **Algorithms and codes** using deep learning frameworks (eg, tensorflow), which we expect to be of interest to develop an open source SWOT data analytics toolbox;
- **Datasets and associated benchmarking experiments** which will be used to distribute a SWOT data challenge within the framework of Melody project.

In addition to these contributions, this PhD will also contribute to outreach activities co-animated by Lab-STICC and LOPS (e.g., doctoral course on Data Science for Geoscience, summer schools on Ocean Remote Sensing Synergy).

We may also emphasize that we expect the proposed models and algorithms to be of broad interest beyond the considered SWOT case-studies, which include applications to satellite-derived geophysical products from irregularly-sampled earth observation datasets (e.g., land surface temperature, ocean colour...) and a variety of geospatial applications. Such applications may provide the basis for spin-off activities (e.g., internships, new collaborations with SMEs,...).

Candidate profile

The targeted PhD candidate shall have a MSc and/or engineer degree in Data Science or Artificial Intelligence with a strong interest in environmental sciences, possibly acknowledged by previous activities or experience. A dual degree in ocean science and data science as promoted by Isblue MSc program would be of key interest. Besides a strong theoretical background, computer skills, including first experience in using state-of-the-art deep learning frameworks (e.g., tensorflow, pytorch) and programming environment (e.g., python, git server), will be particularly expected.

References

[1] Brunton et al. PNAS, 113(15), 2016.
[2] de Bezenac et al. ICLR, 2018.
[3] Chapron et al., QJRMS, 2017.
[4] Chen et al., NIPS 2018.
[5] Chung et al., NIPS, 2015.
[6] Cressie et al., Wiley, 2011.
[7] Evensen, Springer, 2009.
[8] Fablet et al. IEEE TCI, 3(4), 647-657, 2017.
[9] Fablet et al. EUSIPCO, 2018.
[10] Fablet et al., Arxiv, 2019.
[11] Fraccaro et al., NIPS, 2016.
[12] Gaultier et al., JAOT, 2016.
[13] Goodfellow et al. NIPS, 2014.
[14] He et al., CVPR, 2016.
[15] Karpatne et al., IEEE TKDE, 29(10), 2017.
[16] Lguensat et al. MWR, 145(10), 2017.
[17] Lorenz. J. Atm. Sc., 20(2), 1963.
[18] Lu et al. ICML, 3282–3291, 2018.
[19] Nelson et al., Preprint, 2018.
[20] Nguyen et al. IEEE DSAA, 2018.
[21] Nguyen et al. Preprint, 2019.
[22] Ouala et al., Remote Sensing, 2018.
[23] Ouala et al., ICASSP, 2019.
[24] Ouala et al., ICASSP, 2019.
[25] Ouala et al., Preprint HAL, 2019.
[26] Rasp et al., PNAS, 2018.
[27] Rousseau et al., JMIV, 2019.
[28] Sévellec et al. Nat. Comm., 9(3024), 2018.
[29] Zhao et al.. Nonlinearity, 29(9), 2016.
[30] NATL60 ocean simulations.