**Deep Learning based Submesoscale Ocean Eddy Detection on the Amazon Web Service Cloud**

**Project lead**: Jianwu Wang, Associate Professor of Data Science, Department of Information Systems, University of Maryland, Baltimore County (UMBC), jianwu@umbc.edu, 410-455-3883

**Project partners**: Jinbo Wang, PODAAC Project Scientist and SWOT Scientist, NASA Jet Propulsion Laboratory (JPL), jinbo.wang@jpl.nasa.gov, 818-354-5936

**Proposed start and end date**: 08/01/2021-01/31/2022

**Budget Requested**: Total budget $10,000 with additional $5,000 AWS cloud credit. The budget will support a graduate student at UMBC to work on the project. Neither project lead at UMBC nor project partner at JPL has the complete technical skills to accomplish the proposed work. This program would provide a perfect opportunity for the two sides to collaborate on this interdisciplinary work. The itemized budgets based on deliverables are: 1) $2,500 for training data preparation and preprocessing, 2) $2,500 for deep learning (DL) model design and implementation, 3) $2,500 for DL based ocean eddy detection pipeline, 4) $2,500 for reproducible and scalable pipeline on AWS.

## Project Outline

**Project description**: Ocean eddies play an important role in ocean circulation and climate by transporting heat, salt, nutrients, and tracers such as CO2. While the mesoscale (~200-300km) eddies are important for horizontal transport, the submesoscale eddies (<100km) are important for both the horizontal and vertical transports. The vertical transport, especially in the upper ocean, is a crucial process in the oceanic heat and carbon uptake, which significantly modulates the Earth's climate variability.

Figure 1 shows an example of small-scale eddies revealed by VIIRS-NP sea surface temperature (SST) data in the California Current. The data is retrieved by the Visible Infrared Imaging Radiometer Suite (VIIRS) instrument at the joint NASA/NOAA Suomi National Polar-orbiting Partnership (Suomi NPP) satellite, and archived and distributed by NASA JPL PODAAC. It is clear that a global survey of these small-scale features is a daunting task for manual labeling.

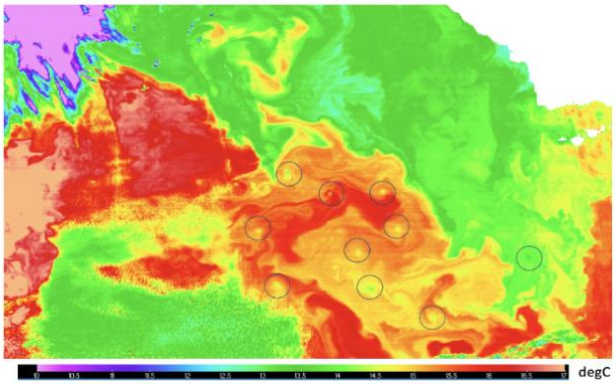


*Figure 1: Example of the coherent submesoscale eddies in/near the California Current marked by blue cycles.*

This project aims to solve the problem by applying deep learning (DL) techniques to SST images and build a cloud-native DL algorithm that can perform auto-identification. The overall deliverable is a **reproducible and scalable open-source deep learning pipeline for ocean eddy detection on AWS cloud**, which will be achieved by the following **four phases/deliverables**. Throughout the phases, our

deep learning model developers will work closely with our partners at NASA JPL to present the results and discuss how the DL model could be improved, such as by studying possible reasons of the failed predictions.

**Deliverable 1: labeled SST images for deep learning model training.** As shown in Figure 1, each SST granule/image might contain many submesoscale ocean eddies. Labeling ocean eddies requires some tagging software such as LabelImg and Imagelabeler. To simplify model training, we will also segment the original images/granules (3200 x 3200 pixels) into smaller sub-images (20 x 20 pixels) so each sub-image will only have at most one submesoscale ocean eddy. To avoid imbalanced training data, we will provide relatively same numbers of sub-images with and without ocean eddies. We plan to mainly use NASA VIIRS Suomi NPP sea surface temperature (SST) Level 2 data hosted at JPL PODAAC (doi: 10.5067/GHVRS-2PO61) for our labeling efforts. Specifically, we plan to have 100 labeled images and about 10 eddies in each image, which will produce about 1,000 eddy and 1,000 non eddy sub-images.

**Deliverable 2: deep learning (DL) model training and optimization.** Based on the training data received above, we will explore different DL models including Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) to measure the accuracy of these models in classifying whether each SST image in our training dataset contains ocean eddy or not. We will also explore whether unsupervised learning models such as Generative Adversarial Networks (GAN) could be used so that less labeled data is required. For each model, we will also tune model parameters and hyperparameters (model layer number, activation function, batch size, etc.) to find the configuration that can have the best performance.

**Deliverable 3: deep learning (DL) pipeline for ocean eddy detection.** We will build on top of our trained deep learning model to detect submesoscale eddies in the original satellite SST observation data. Because the sub-image sizes of the training data are much smaller than those of the original SST images/granules, to use our trained model in practice, we need to design a DL based pipeline to take original SST granules as inputs and output all identified submesoscale ocean eddies and their locations in each granule. One important step of the pipeline is to effectively partition each original SST granule into sub-images so we can directly apply our trained DL model for detection.

**Deliverable 4: reproducible and scalable deep learning (DL) pipeline execution on AWS cloud.** We will deploy our DL-based ocean eddy detection pipeline to AWS cloud to achieve reproducible and scalable pipeline execution. We will use proper cloud services provided by AWS, such as Simple Storage Service (S3), Elastic Compute Cloud (EC2) and Virtual Private Cloud (VPC). PODAAC Cloud already migrated a lot of its data into AWS S3, including the SST data (doi:10.5067/GHVRS-2PO61, EarthData login required) we will mainly use in this proposed project. We will create a virtual network using AWS VPC to have an isolated and secure network environment, then start one or more GPU virtual machines

via AWS EC2 service within the virtual network. Then we will run our DL pipeline in the cloud, access SST data from AWS S3 and output predicted ocean eddies for each SST image. To achieve scalability, we will utilize the Horovod distributed deep learning package to speed up the training and process on multiple GPU nodes. To achieve automation, we will utilize the Boto library, which is an integrated interface to AWS cloud services. Specifically, our implementation of the whole pipeline will be Python scripts with parameters to be configured by users such as their cloud credential, GPU virtual machine type and number. To achieve reproducibility, other users will be able to have exactly the same virtual environment and the same DL model by running the script with the same parameters.

**Project objectives (learning & technical)**: From the learning opportunities provided by this proposed project, we aim to provide (1) a curated training dataset of small-mesoscale (<150km) oceanic eddies identified from high-resolution SST data, and (2) a cloud-native, deep-learning-based open-source pipeline that can automatically identify such eddies from the SST images on PODAAC cloud. Its **innovation** lies in the integration of techniques (deep learning, cloud computing, scalability and reproducibility) to provide an end-to-end solution for the real-world Earth science challenge.

**Project significance and impact**:

**Project significance and impact for physical oceanography.** Observing small-scale oceanic structures from a global perspective is crucially important for monitoring their influence on the climate system such as the heat and carbon budget. They are also critical for integrating air-sea coupling into numerical modeling. NASA's upcoming Surface Water and Ocean Topography (SWOT) mission targets these small-scale oceanic processes, and will provide a comprehensive view of the associated circulation fields at small scales. SWOT is a pathfinder mission carrying a new generation altimeter, a Ka-band Radar Interferometer that is completely different from the conventional nadir altimeter. Validating such new measurements is an important first step toward further scientific research and applications. It is a difficult task to validate oceanic eddies at SWOT scales (15-150km) from a global perspective. In-situ campaigns that focus on small-scale ocean circulation are often limited to small regional settings. However, because oceanic eddies often imprint on sea surface temperature as local anomalies, satellite sea surface temperature images can provide such a validation database from a global view.

Because a global survey of these small-scale features is a daunting task for manual labeling, by applying DL techniques to SST images, the results will not only directly benefit the SWOT oceanography community, but also will benefit the research, field campaigns, and applications that target small-mesoscale ocean eddies.

**Project significance and impact for applying deep learning (DL) techniques in Earth sciences.** DL is a promising technique and has already revolutionized many fields, and is increasingly being used in remote sensing applications. Meanwhile, Earth remote sensing data now grows at an astronomical pace as

satellite instruments become more and more powerful. Further, both NOAA and NASA are committed to migrate their Earth remote sensing data to public cloud providers like AWS. This proposed project would greatly promote better adoption of DL and cloud computing techniques for Earth sciences by assessing and demonstrating: 1) how DL models can help Earth science application, 2) how DL pipelines can run natively and efficiently at AWS cloud, 3) how to make a cloud-native DL pipeline reproducible.

**Description of key project steps and timeline**: We plan to finish the first version of each of the four deliverables identified above within one month and then refine the tasks in the last two months. The exact timeline is: 2021/08: Data labeling, 2021/09: DL model implementation, 2021/10: DL pipeline for ocean eddy detection, 2021/11: DL pipeline on AWS, 2021/12-2022/01: refinement of the four tasks.

**Description of additional funding currently supporting this work**: Both Jinwu Wang and Jinbo Wang are co-leads of the NASA Machine Learning Capacity ESDSWG working group. Jianwu Wang is supported by his NASA ACCESS award (80NSSC21M0027) to participate in the working group. Jinbo Wang is the PODAAC deputy project scientist supported through PODAAC.

## Outreach

**What groups/audiences will be engaged in the project?** We will primarily engage the 15+ members in the NASA Machine Learning Capacity ESDSWG working group including Chandana Gangodagamage at NASA GSFC and Ziheng Sun at George Mason University. We will present our results in its monthly meetings and seek feedback.

**How will you share the knowledge generated by the project?** We will share our results to other NASA Earth Science Data System Working Groups (ESDSWG) via its meetings. We note that we have received approval from Steve Olding, the ESDSWG coordinator, for this proposal's submission.

**Description of *who* (agencies/individuals) should be aware of this project, i.e. potential outreach targets.** We hope to reach out to additional agencies/individuals such as NOAA on our results by organizing a session/tutorial at ESIP meetings (see below for details).

## Project Partners

**Description of project partners (agencies/individuals) and their involvement**: 1) Data scientists at NASA JPL PODAAC including Ed Armstrong, Ben Holt and Jorge Vazquez, who will mainly help provide labeled data and feedback on DL model results. 2) UMBC machine learning researchers including Sanjay Purushotham who will help provide guidance on DL model design and evaluation.

**How will this project engage members of the ESIP community**: First, we will engage the ESIP community by proposing a session/tutorial on scalable machine learning at AWS at the 2022 ESIP Winter meeting. Second, we will present/discuss our results at the ESIP machine learning cluster and the ESIP cloud computing cluster. Third, we will open source our implementation at ESIP GitHub repository.