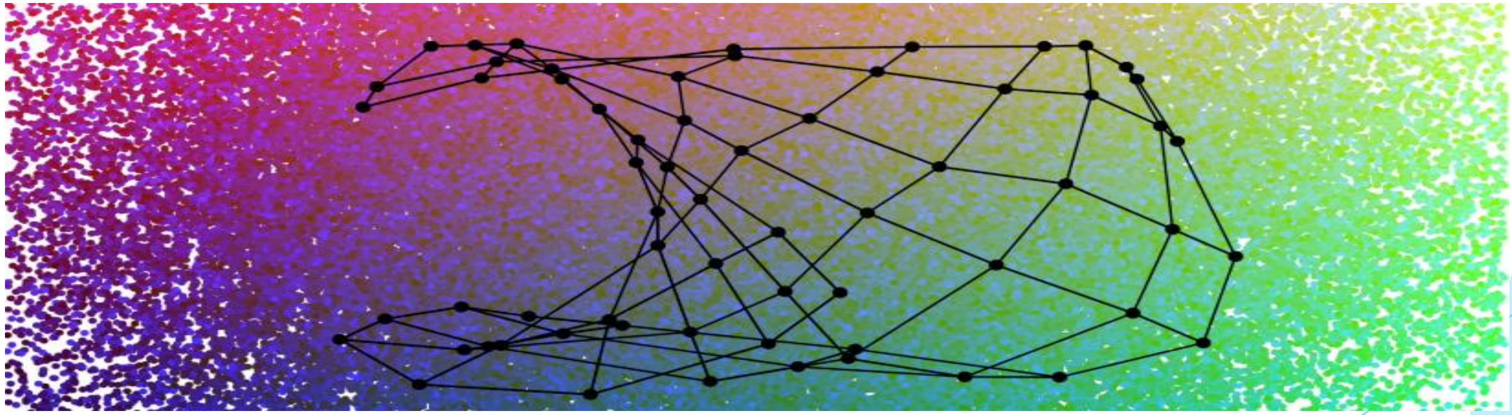
OCEAN CARBON FROM SPACE WORKSHOP
14-18 February 2022

Development of a Machine Learning approach for the estimation of the marine CO_2 Partial Pressure over the global coastal ocean in the frame of the CO_2COAST project



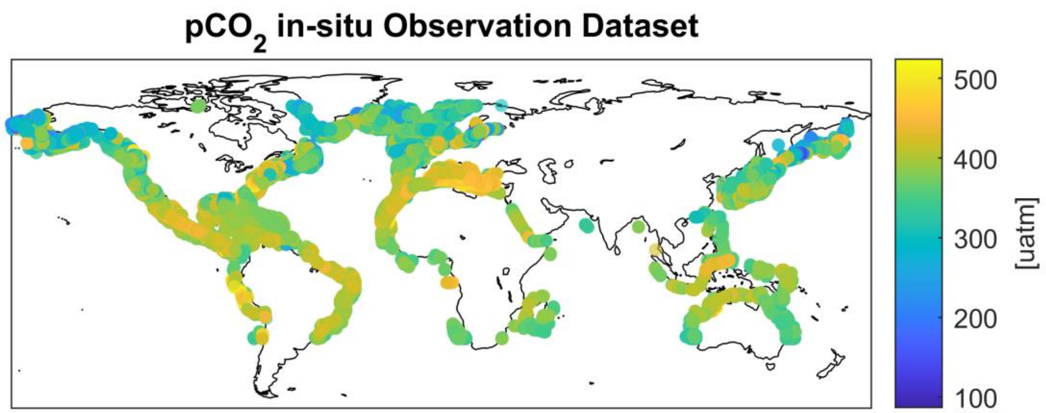
Roy El Hourany, Hubert Loisel, Daniel S.F. Jorge, Vincent Vantrepotte, Cédric Jamet, Sylvie Thiria, Ndeye Niang Keita, Julien Demaria, Marine Bretagnon, Antoine Mangin

While many improvements have been done for about 2 decades by the ocean color community in terms of algorithms development for the monitoring of coastal waters, the application of these algorithms is still limited to local or regional scales.

CO_2Coast (2021-2025)

- Estimate the surface-ocean CO_2 partial pressure, pCO_{2w}, CO_2 flux, and associated uncertainties from satellite remote sensing over the **global coastal waters** at 1 and 4 km^2
- Assessment of the seasonal, inter-annual, and **last 25 years** trend evolutions of pCO_{2w} and CO_2 flux over global coastal waters.
- Analysis of the respective contribution of estuaries and coastal shelf waters, as well as the contribution of the different continental shelf types.

Global algorithm development; Data



Global pCO2 dataset (SOCAT: https://www.socat.info/):
~500 000 obs with ~90 000 presenting non-missing satellite matchups

Satellite Parameters (daily 4km resolution):
Rrs (6 wavelengths), SST, SSS, acdom, Chla
(from 09/1997: SeaWiFS, MERIS, and present (up to 12/2023) OCR (MODIS, VIIRS, OLCI) sensors + ACRI-ST products)

Challenges: High dimension data,
The relationship between variables might be regionally-dependent

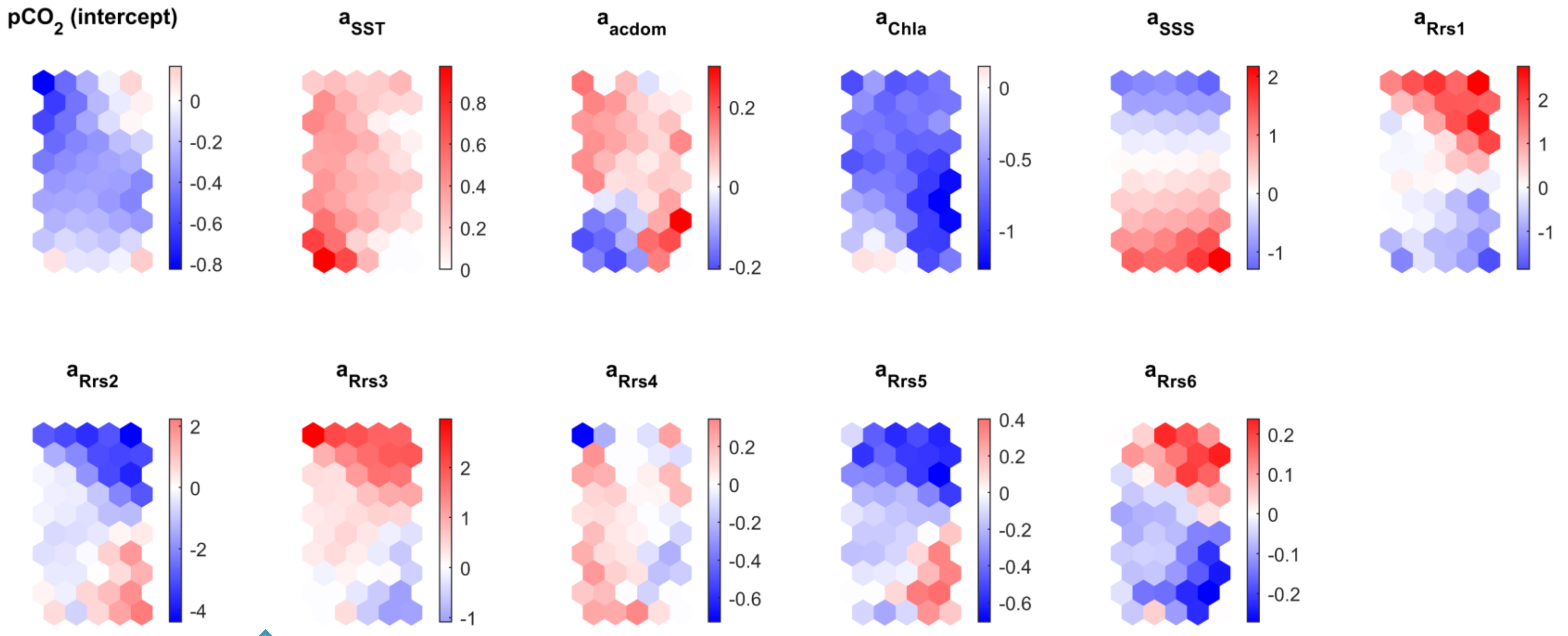
→ **Clusterwise Regression**

Clusterwise regression is used to classify observations into groups, and that the relationship between pCO2 and satellite parameters is different depending on the group. This process can identify and characterize regional relationships in a global dataset.

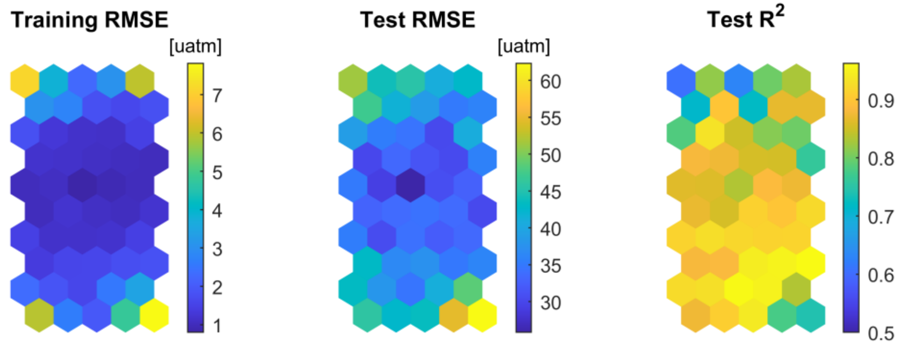
Clusterwise regression using SOMClusterwise

- Classify individuals into G groups while having the best model within each of the groups.
- Classification and learning carried out simultaneously

Each cluster of the SOM is characterized by an intercept and a slope for every variable [pCO2= intercept+Σa_{var} * var].
The resultant slopes shows how each variable affects the variation of the pCO2 in each cluster



alpha and intercept coefficients for each cluster on the SOM grid (normalized scale), Ex: SST always affects positively the pCO2 variability. But SSS shows negative and positive influence, depending on the cluster.



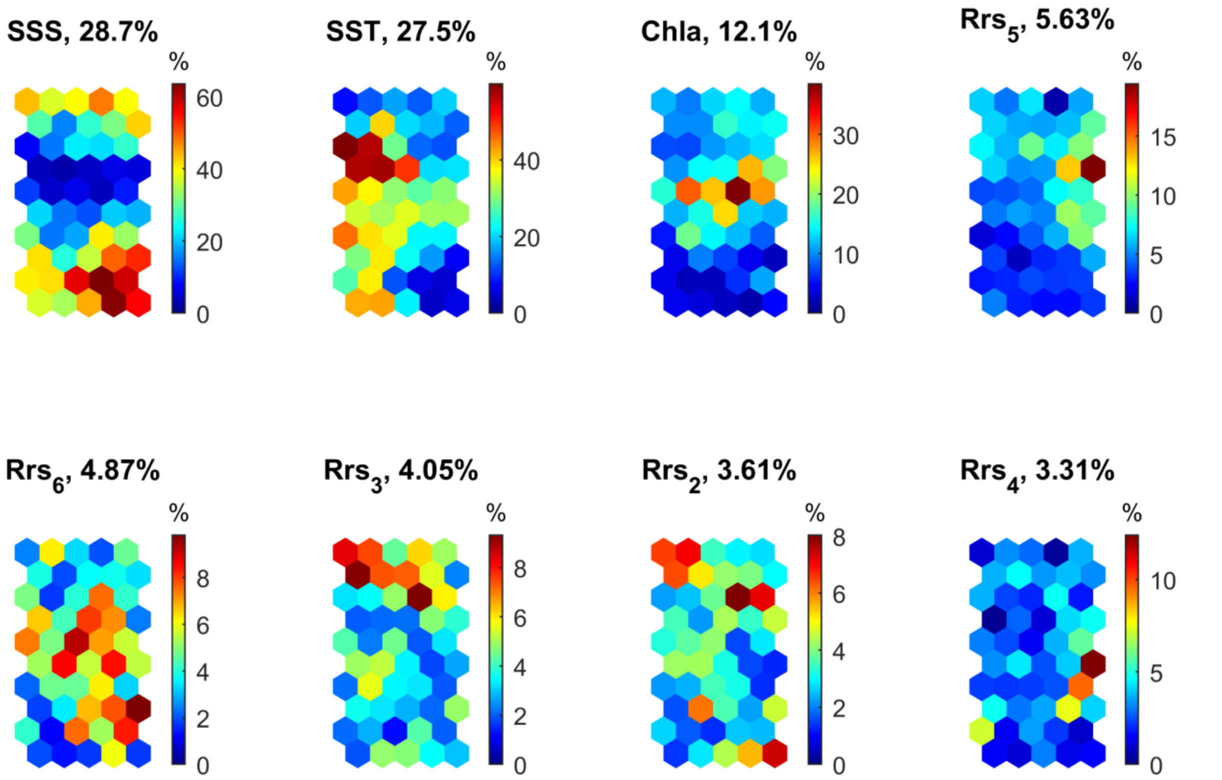
Test results; R2 and RMSE for each cluster on the SOM grid

PS: Training done with 9K obs and tested with 80K obs

The importance of Satellite-derived variables used to accurately estimate pCO2 values vary from a cluster to another. This result upholds the regional aspect of the relationship between pCO2 and the oceanic parameters.

<u>Importance of variables;</u>

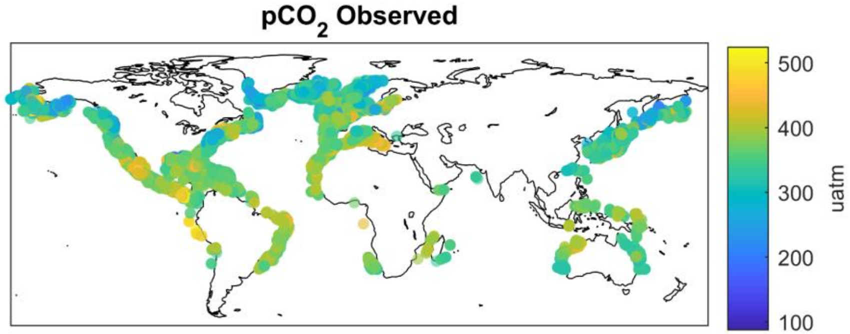
Random Forest analysis applied in each SOMCluster

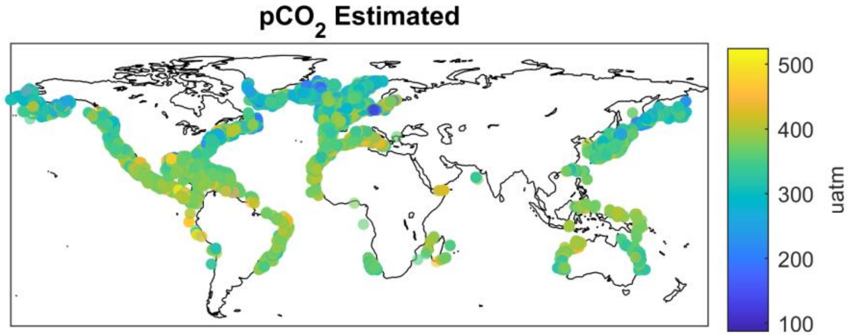


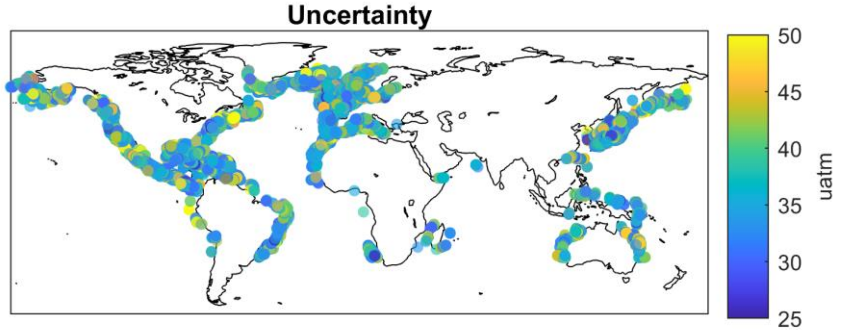
Variables' importance (%) shown on the SOM grid for each cluster.

The variables were ordered on this figure following an increasing overall mean importance shown in each plot's title

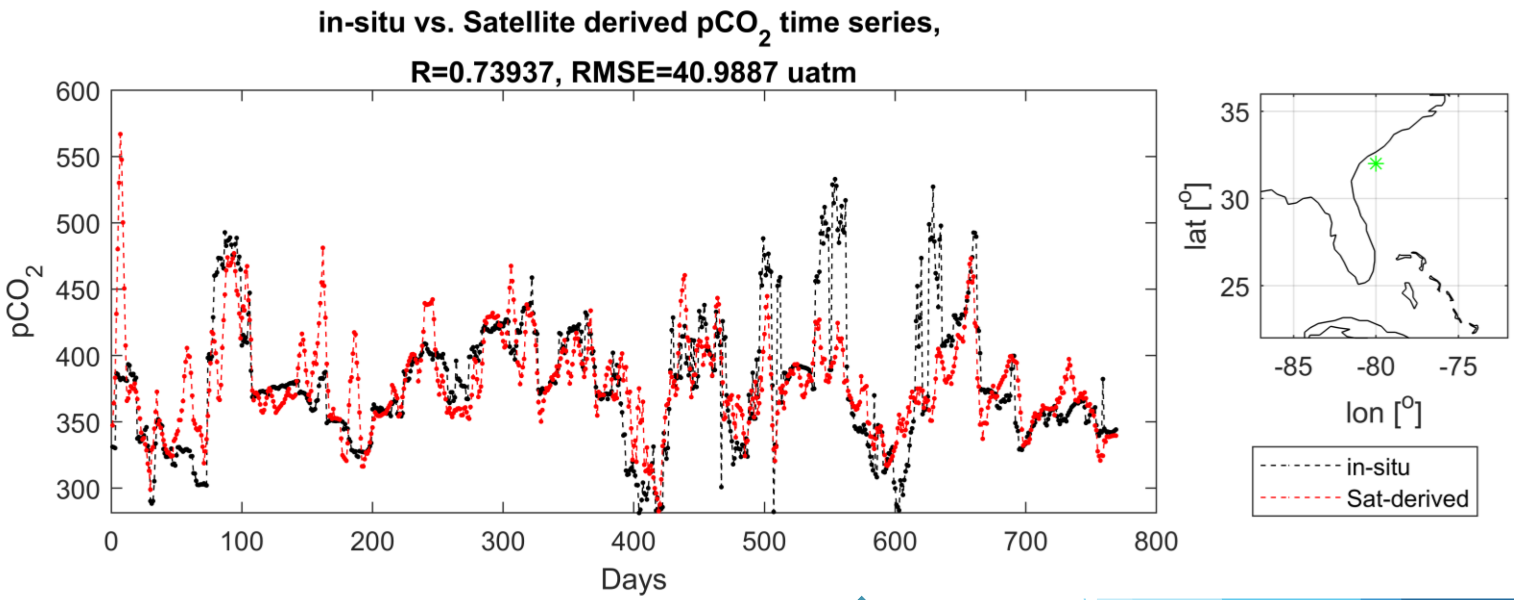
Tests applied at the global and temporal scales highlight a significant accuracy while estimating pCO2 from satellite data based on a clusterwise regression approach.

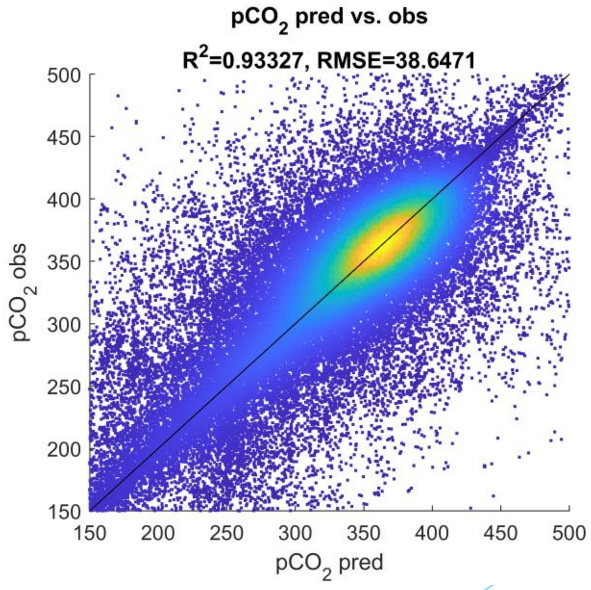






PS: Training done with 9K obs and tested with 80K obs





Time series validation

Global validation and associated uncertainties.

Using more observations to train the algorithm could help reducing further the RMSE values.

Next steps:

- Algorithm adjustments: Incorporating more datapoints in the training process, choice of optimal hyperparameters (number of clusters and the number of training iterations) to get the best significant models with the least error possible for each cluster,
- Development of a quality control mask and Geographic interpretation of the clusters,
- Production of a daily Satellite-derived global dataset

future aims:

- Assessment of the seasonal, inter-annual, and last 25 years trend evolutions of pCO2w and CO2 flux over global coastal waters.
- Analysis of the respective contribution of estuaries and coastal shelf waters, as well as the contribution of the different continental shelf types.

Thank You