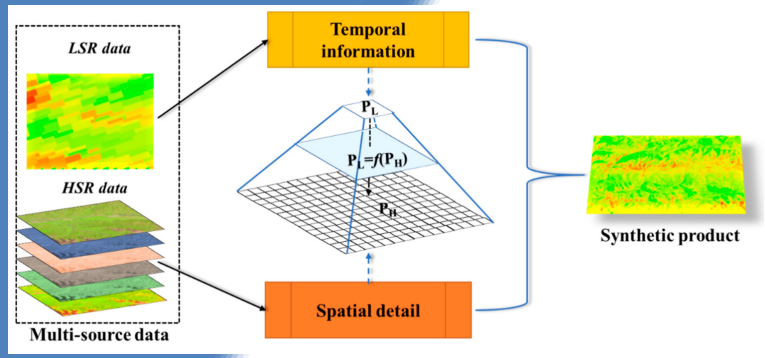
USING MACHINE LEARNING ALGORITHMS TO CHARACTERISE THE CARBON SYSTEM VARIABILITY DRIVERS IN THE EASTERN TROPICAL ATLANTIC

Dimitry Khvorostyanov[1], Nathalie Lefevre[1], Carlos Meja[1], Laurence Beaumont[2] and Urbain Koffi[3]

[1] LOCEAN / IPSL - CNRS/IRD/SU/MNHN
[2] Division Technique INSU
[3] LSPFA, ENS, Cote D'Ivoire

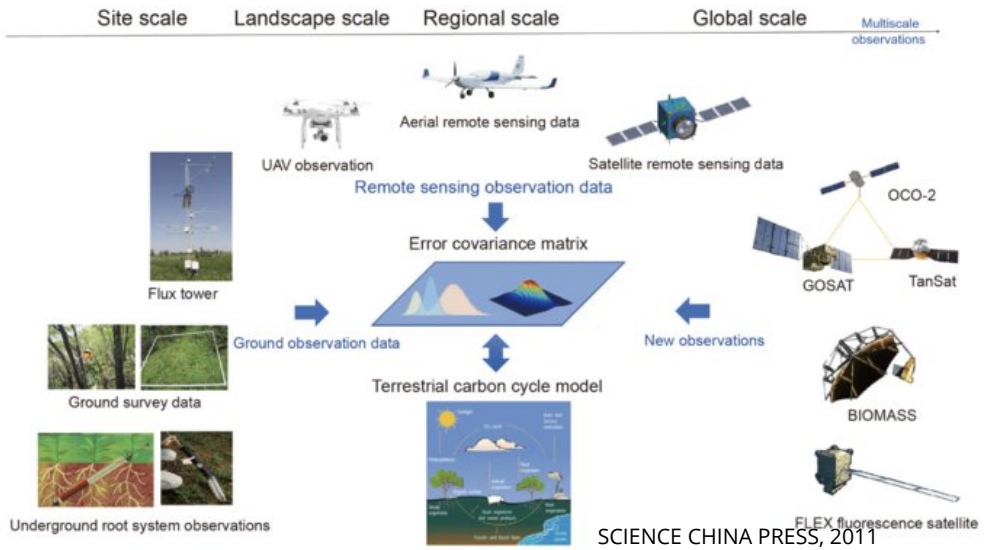
JOURNEE IA
ODATIS
04/06/2024

Approach & pertinence: Data fusion



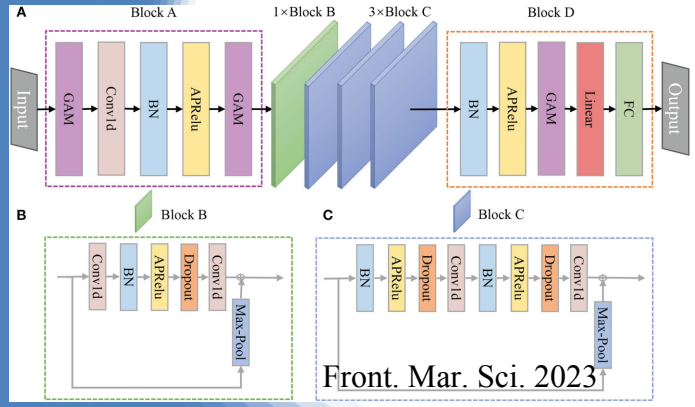
Retrieval of Land Surface Temperature from Landsat-8 and MODIS

Characterizing terrestrial carbon cycle

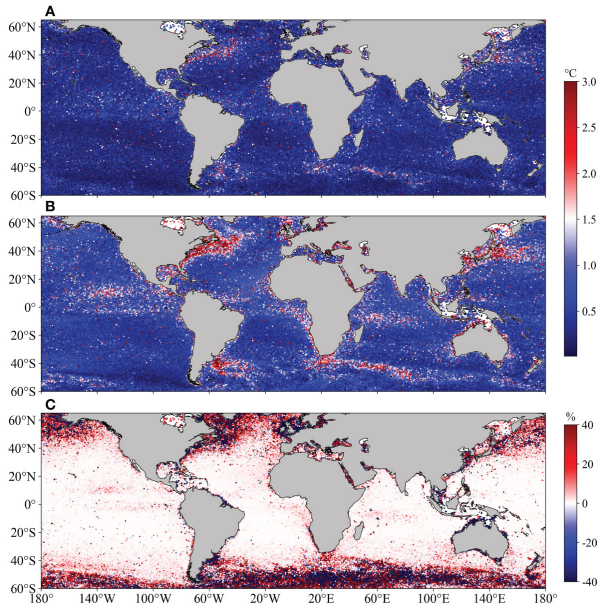


SCIENCE CHINA PRESS, 2011

Fusion of ocean data from multiple sources using deep learning. Wang et al, (2023)



Front. Mar. Sci. 2023



Increased Contextual Understanding

Enhanced Data Completeness

01 Enhanced Accuracy

02 Improved Predictive Capabilities

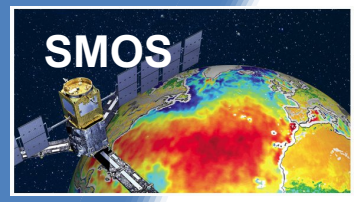
03

04

05 Improved Decision-Making

Contexte favorable pour des utilisations croisées de données satellite / *in situ*

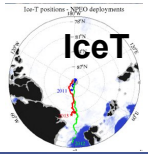
De nombreux projets pilotés par LOCEAN où sont produites et validées les données satellitaires et les données in situ

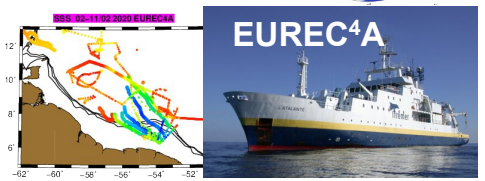


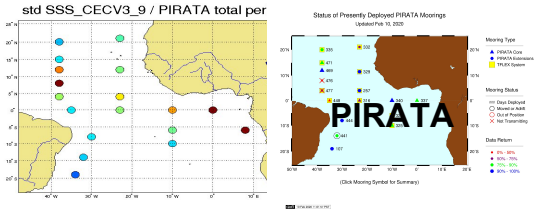
- Assurer le contrôle qualité des données in-situ et l'interprétation dans un contexte plus large grâce aux données satellitaires

- Valider les données satellitaires avec des données ir dans des régions spécifiques

- Faciliter des études de processus utilisant à la fois des données satellitaires et des donnés in situ



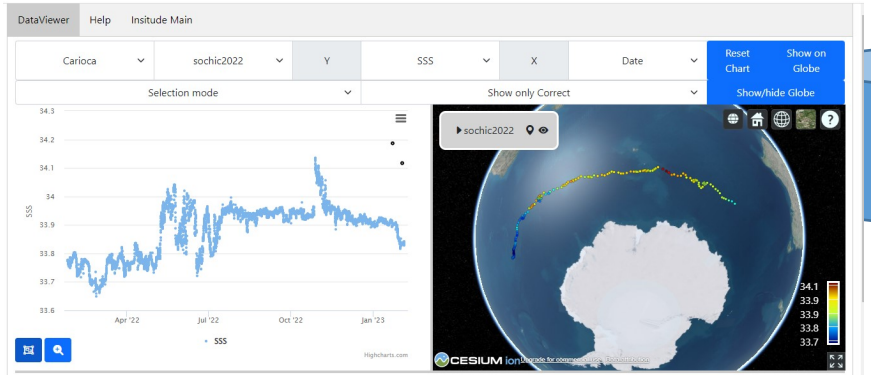




COCA
Coastal and Ocean Climate Applications

Argo

INSITUDE **platform facilitating cross-uses of** satellite / in situ **data**



Common DB

Coloc Builder

Colocalize & Visualize
- ✓ In situ
- ✓ Satellite
- ✓ ARGO

Coloc Validator

Specific Diantostics
- ✓ In situ
- ✓ ARGO

Meta Data

- ❖ Projects
- ❖ People
- ❖ Campaigns
- ❖ Buoys/Ships

Processing

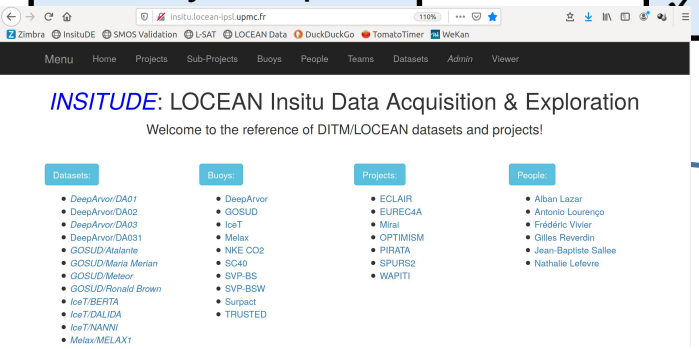
- ✓ Fetch Data
- ✓ Decode
- ✓ Compute
- ✓ Centralize
- ✓ Send & Store

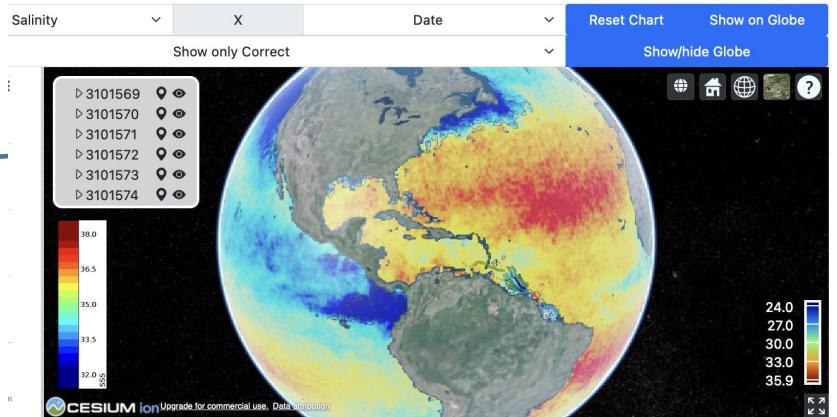
Data Viewer

- ✓ Interactive Visu
- ✓ Combine Data
- ✓ Assign QF
- ✓ Extract Valid

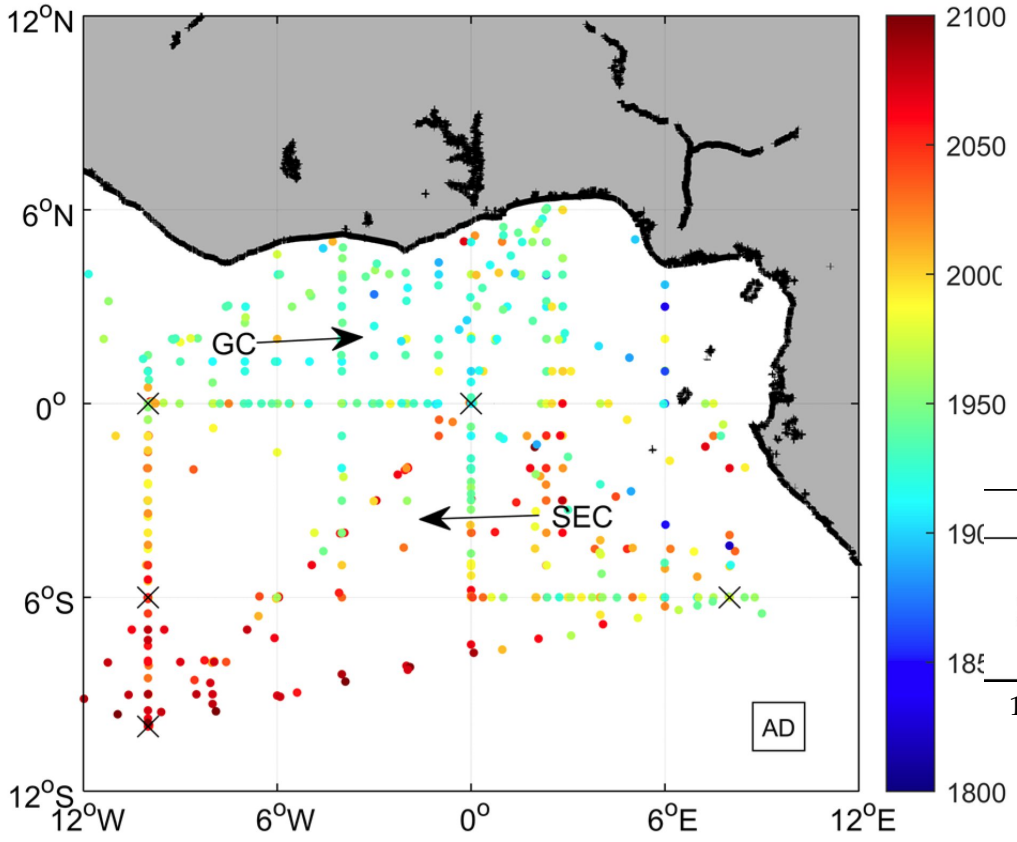
Process Studies Helper

- ✓ Correlograms
- ✓ Pair plots
- ✓ Regression plots
- ✓ Choice of variables





In situ data : EGEE 3 & PIRATA cruises & moorings



The Carbon system: 4 variables, 2 independent

- Fugacity of CO2 (fCO2)
- Dissolved inorganic carbon (DIC)
- Alkalinity (TA)
- pH

EGEE 3 & PIRATA FR-29 cruises: fCO2, DIC, TA measured
=>carbon system over-determined

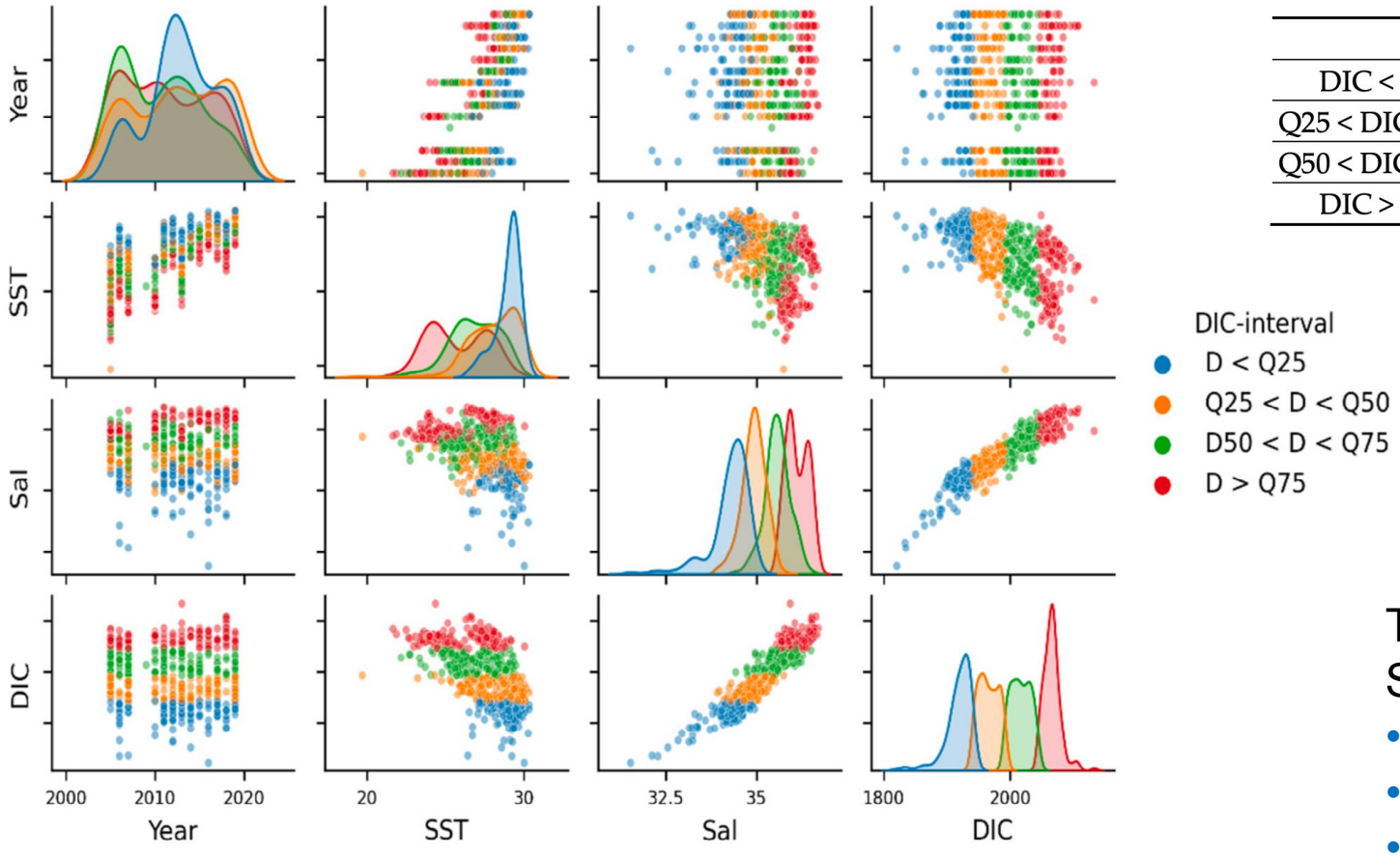
Pair	Calculated Parameter	RMSE	*r*
TA–DIC	fCO_2	11.7 µatm	0.90
DIC–fCO_2	TA	7.4 µmol kg^{-1}	0.99
fCO_2–TA	DIC	5.9 µmol kg^{-1}	0.99

119 concomitant measurements of DIC, TA and fCO_2 are used to check the consistency of the carbon system

Distribution of DIC observations (2005–2019)
and location of the five PIRATA moorings

Lefèvre, N.et al, Ocean Circulation Drives the Variability of the Carbon System in the Eastern Tropical Atlantic. *Oceans* 2021, **2**, 126-148

"Features" and regression models



	Condition	N	N_{Train}	N_{Val}
DIC < Q25	DIC < 1941.1	159	132	27
Q25 < DIC < Q50	1941.1 ≤ DIC < 1992.2	159	132	27
Q50 < DIC < Q75	1992.2 ≤ DIC < 2044.6	158	132	26
DIC > Q75	DIC ≥ 2044.6	161	135	27

Data selection to ensure homogeneous representation of data in the validation dataset

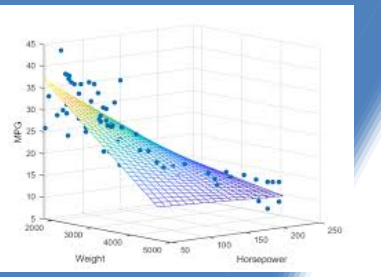
Target variable : DIC

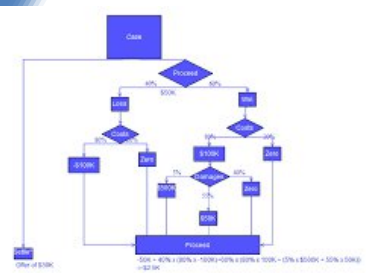
Selected regression parameters (*features*):

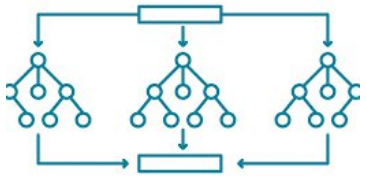
- SSS, SST
- Year (interannual trend)
- sin(DoY), cos(DoY) (intra-annual variations) -- NN only

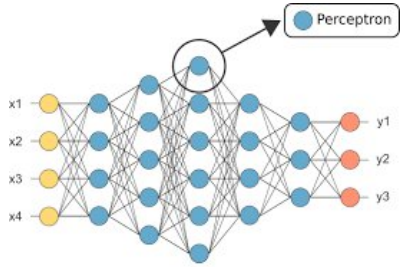
Scatter plots between variables and density plots for each variable (on the main diagonal). Data are colored according to the intervals of the quantiles of DIC (25%, 50%, 75%).

"Features" and regression models









Selected regression models:

- Multivariate linear regression
- Decision Tree
- Random Forest
- Feed Forward Neural Network

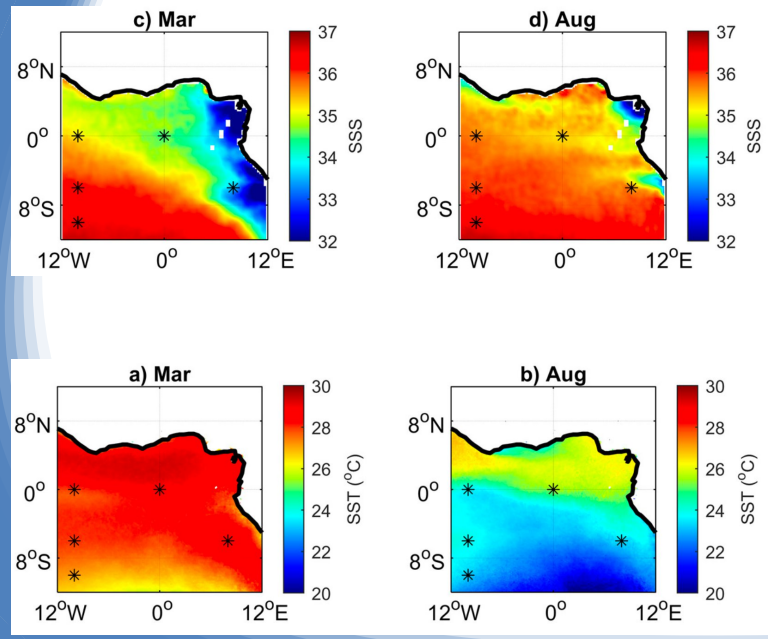
Mooring or Cruise	Regression Method	RMSE (μmol kg^{-1})	r	N	Time Period
6° S, 10° W	MLR	7.7	0.96	6611	2006–2017
	DT	14.1	0.87		
	RF	9.8	0.94		
	NN	7.3	0.97		
6° S, 8° E	MLR	14.8	0.98	239	2017–2019
	DT	25	0.95		
	RF	24	0.95		
	NN	12.8	0.99		
EGEE 3	MLR	11.8	0.97	6895	2006
	DT	14	0.96		
	RF	10	0.98		
	NN	9.3	0.98		
PIRATA FR-29	MLR	8.1	0.99	4462	2019
	DT	11	0.98		
	RF	9	0.98		
	NN	9.4	0.99		

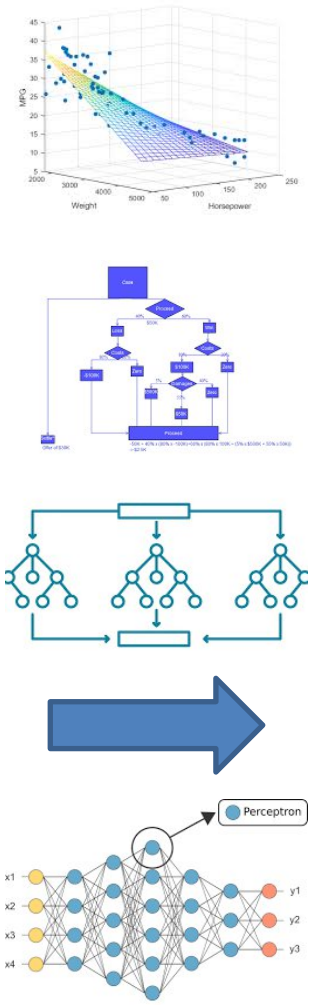
- NN usually performs slightly better but with no significant improvement
- So we use the MLR, which is the simplest and easily interpretable
- This implies that dependencies of DIC on its main predictors are mostly linear
- SSS has the largest impact, with relative importance more than 90% (quantified by RF and DT)
- Impact of the SST is about 5%

Regression model applied to MODIS SST and SMOS SSS

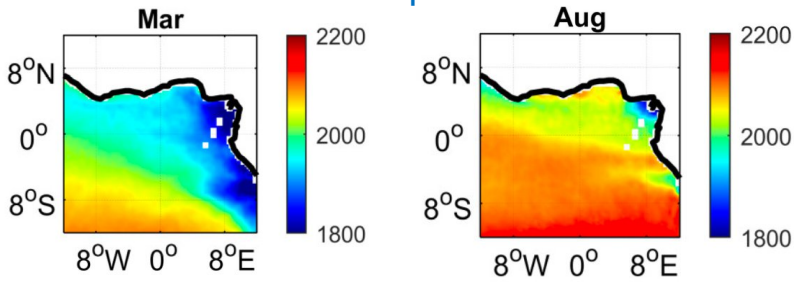
Year

sin(DoY)
cos(DoY)

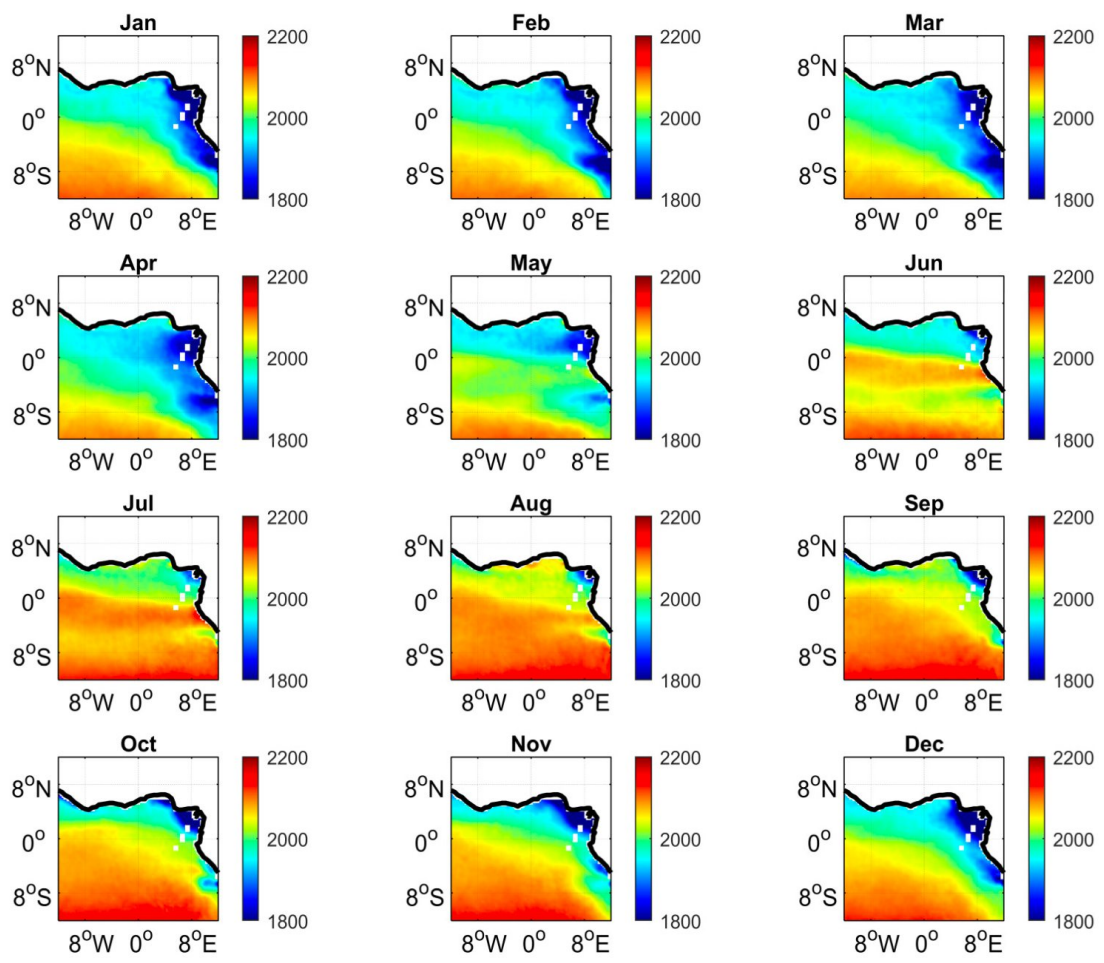




$DIC_{predict}$



Regression model applied to MODIS SST and SMOS SSS



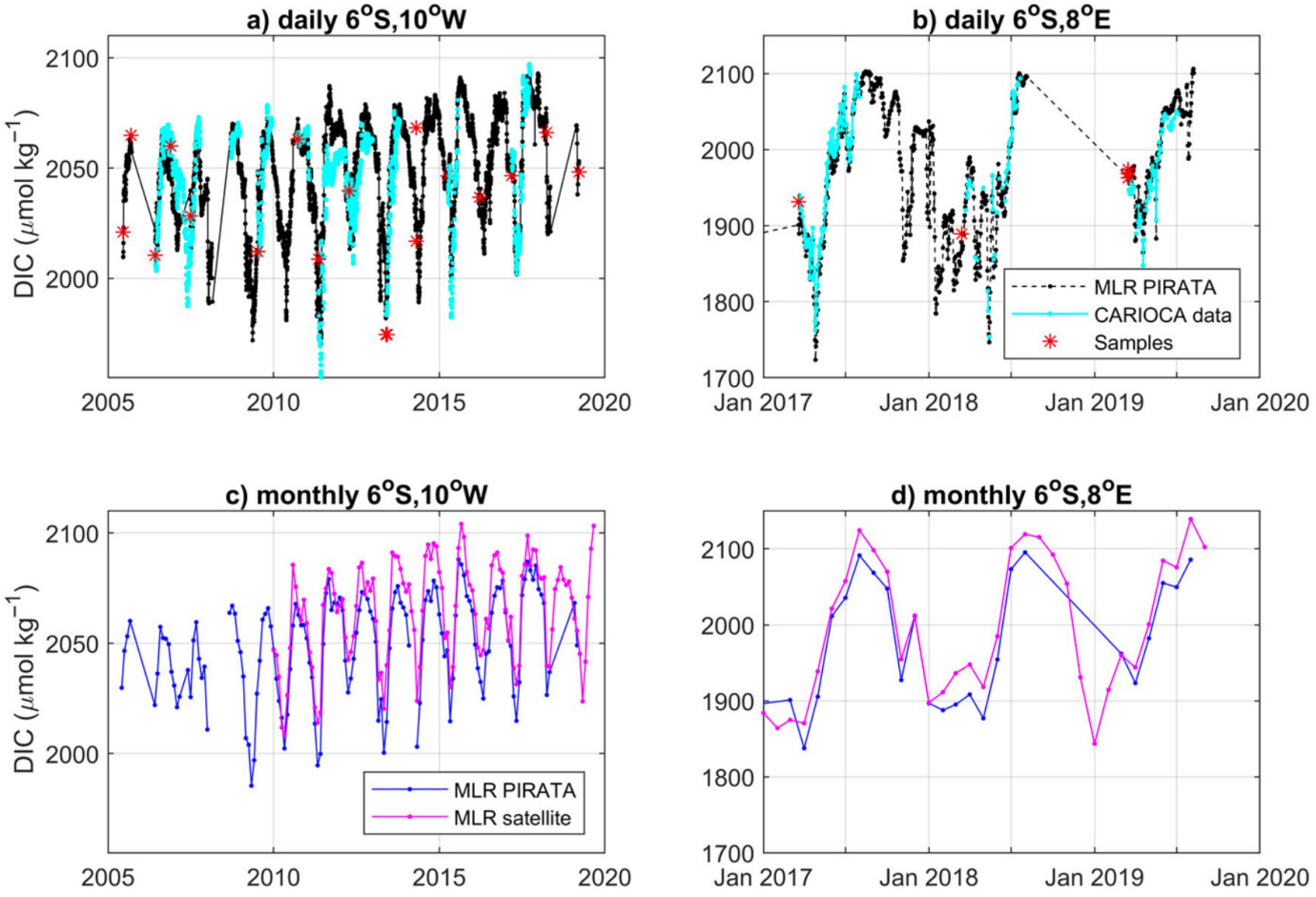
Climatologies of DIC

✓ Applying the regression model to satellite data allows to quantify physical processes related to:

- Upwelling zones & periods
- Water masses
- Seasonal dynamics

Climatology of DIC (µmol kg^{-1}) from linear regression model with the monthly fields of MODIS SST and SMOS SSS.
Climatology period: January 2010 to September 2019

Year-to-Year Variability : Measurements vs MLR Model



Validation of the MLR model with focus on physical features, in addition to the overall validation scores

- ✓ MLR captures well the high frequency variations of DIC
- ✓ overestimates the DIC concentrations in some years
- ✓ Timing of the decreases and increases in DIC is in good agreement between the DIC calculated by the MLR and the DIC calculated from underway fCO2

(a) Daily DIC determined from the multiple linear regressions (MLR) using SST and SSS recorded at the 6° S, 10° W mooring since 2006 (in black) and calculated from measured seawater fCO2 and TA–SSS (in cyan) at 6° S, 10° W The DIC samples are in red. (b) as in a) for the mooring at 6° S, 8° E from 2017 to 2019. (c) Monthly DIC calculated with the MLR using SST and SSS recorded at the mooring (in black) and using satellite SST and SSS (in blue) collocated at 6° S, 10° W. (d) same as in (c) for the mooring at 6° S, 8° E from 2017 to 2019

Conclusions

- Performing studies of oceanographic phenomena using machine learning still requires knowledge of the science behing the processes, e.g.:
  - DIC strongly depends on the water masses and increases over time due to the atmospheric CO_2 increase.

- Combining data from complementary sources (*data fusion*) using machine learning allows insights on the processes otherwise impossible or harder to detect from a single source, e.g.:
  - In the northern part of the basin, relatively fresh and warm waters are associated with lower DIC concentrations compared to the southern part where colder and saltier waters are enriched in CO_2
  - East of 10°E, the strong influence of the Congo plume is evidenced with a different relationship which corresponds to conservative mixing between the river and oceanic waters.

Merci !

Des questions?