A new model-data fusion approach: 2-step training procedure to integrate chlorophyll

Teresa Tonelli, Gianpiero Cossarini, Luca Manzoni, Gloria Pietropolli

OGS

NECCTON
NEW COPERNICUS CAPABILITY FOR TROPHIC OCEAN NETWORKS

1924 2024

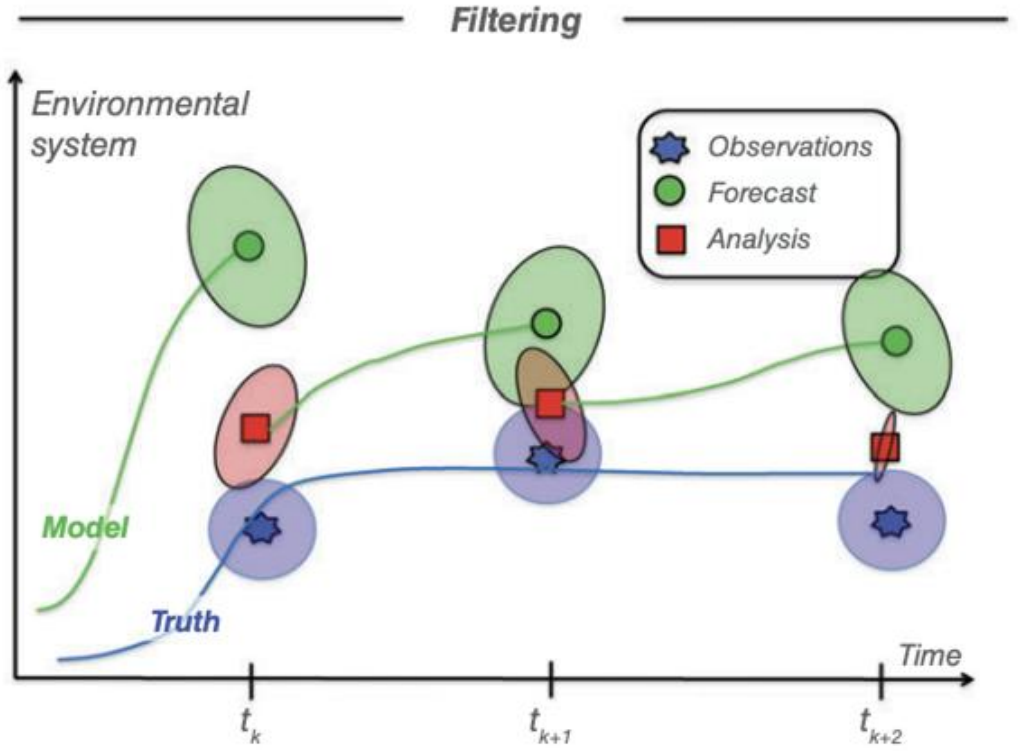
UNIVERSITÀ DEGLI STUDI DI TRIESTE

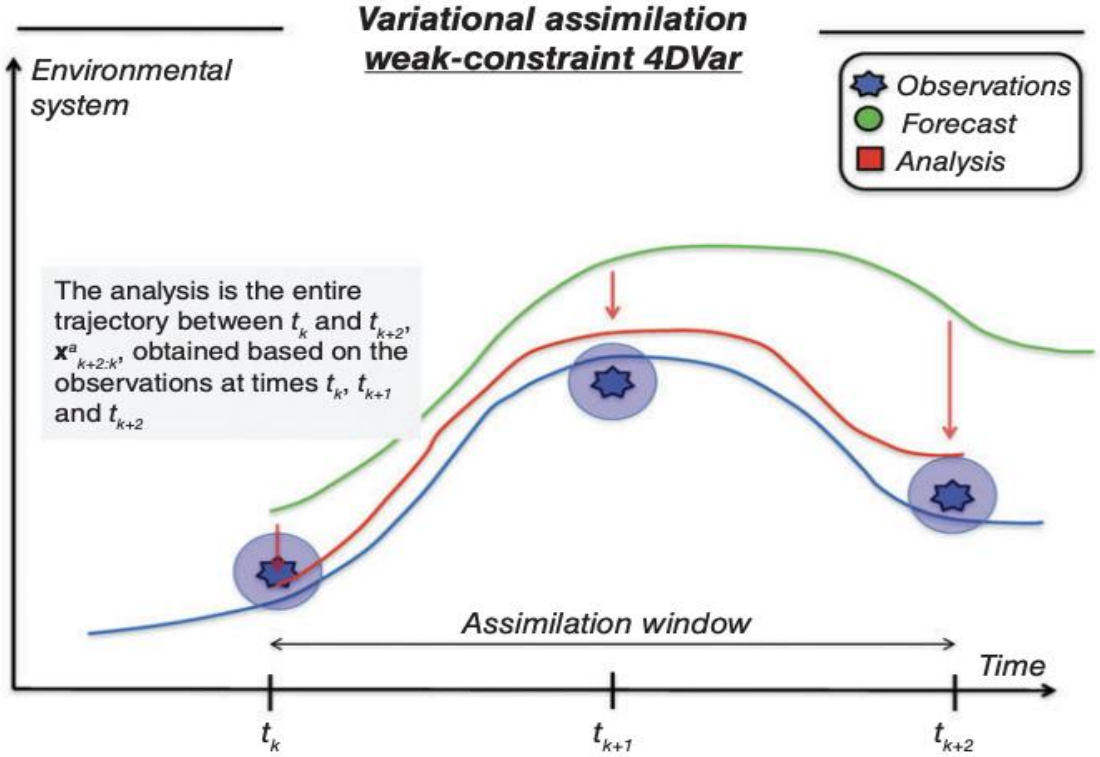
AI-TT Workshop, April 13, 2026

Background: Traditional Data Assimilation

- Traditional model and observations integration is done through **Data Assimilation (DA)**
- DA methods are computationally expensive
- DA methods require numerical model simulations at multiple time steps

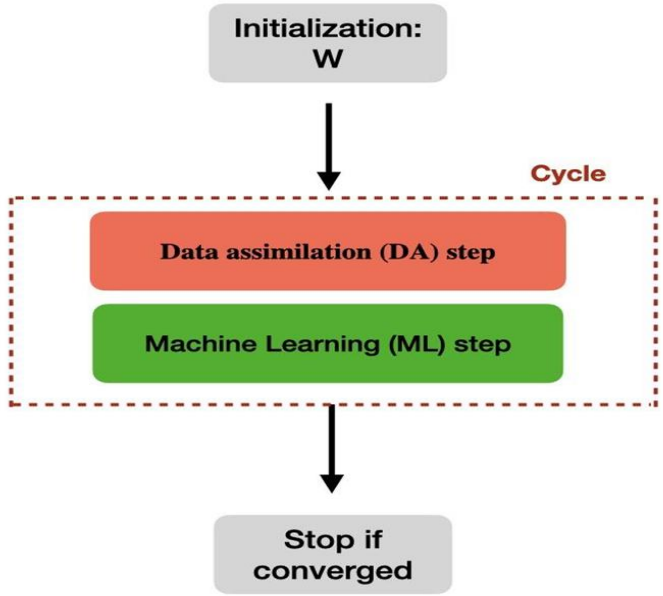
Carrassi, A. et al, (2018). “Data assimilation in the geosciences: An overview of methods, issues, and perspectives”





Model-data fusion

Example of model-data fusion: interaction between ML and DA



Data Fusion Approach: an alternative to standard DA, enabling the integration of observations and model data without the need to rerun numerical simulations.

Implementation of data fusion method through **neural networks**

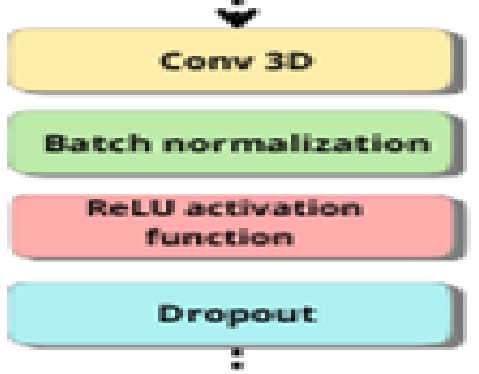
Objective: predicting biogeochemical variables from physical variables

List of physical variables
Salinity
Temperature
Sea water velocity (u and v)
Sea level anomaly

List of biogeochemical variables
Chlorophyll
Oxygen
POC
Nitrate

Convolutional neural networks

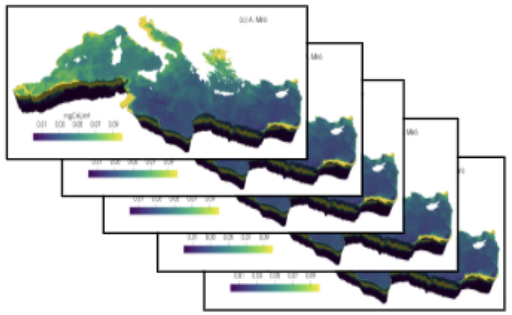
- Convolutional neural networks (CNNs) represent a suitable choice
  - Identify spatial correlation patterns
  - Reduce computational costs
- The CNN is composed of:
  - 6 hidden layers
  - 3D convolutional kernels
  - Batch normalization
  - ReLU activation function
  - Dropout
- The CNN input is a concatenation of 3D tensors while the output is a single 3D tensor



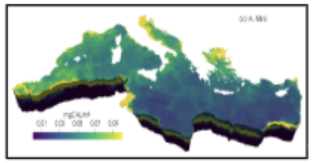
CNN and convolutional layer structure

Layer	Kernel Size	Stride	Padding	Output Shape	Operations
Input	–	-	-	[6, 30, 542, 221]	–
Conv3D	3	1	1	[9, 30, 542, 221]	BN, ReLU, Dropout (d_r)
Conv3D	3	1	1	[16, 30, 542, 221]	BN, ReLU, Dropout (d_r)
Conv3D	3	1	1	[32, 30, 542, 221]	BN, ReLU, Dropout (d_r)
Conv3D	3	1	1	[64, 30, 542, 221]	BN, ReLU, Dropout (d_r)
Conv3D	3	1	1	[128, 30, 542, 221]	BN, ReLU, Dropout (d_r)
Conv3D	3	1	1	[1, 30, 542, 221]	–
Output	–	-	-	[1, 30, 542, 221]	–

Input and output structure



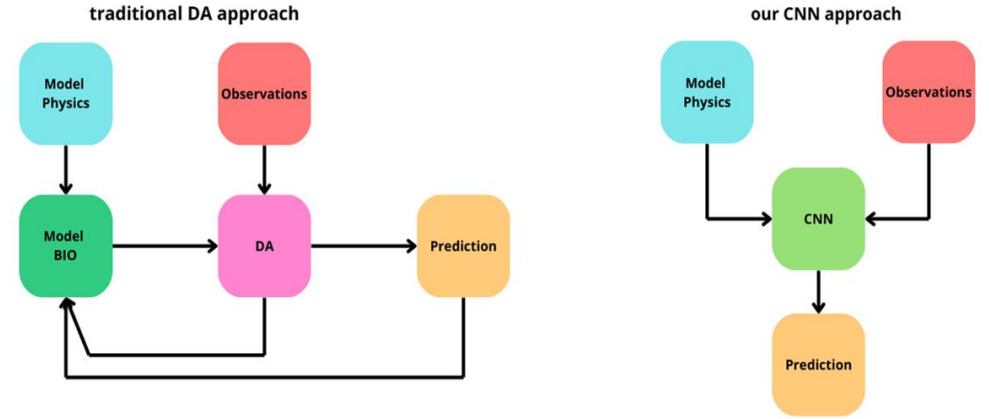
Input: PHY 3D output 1/12° x1/12° x 30 layers



Output: BIO 3D output 1/12° x1/12° x 30 layers

Model-data fusion through CNN: 2-step training procedure

Question: how can neural networks merge different data sources?



Our approach: training procedure composed of multiple steps:

- It splits different data sources and sequentially integrate them into the training procedure.
- It preserves data features during training, reducing the probability to lose information or misinterpreting it

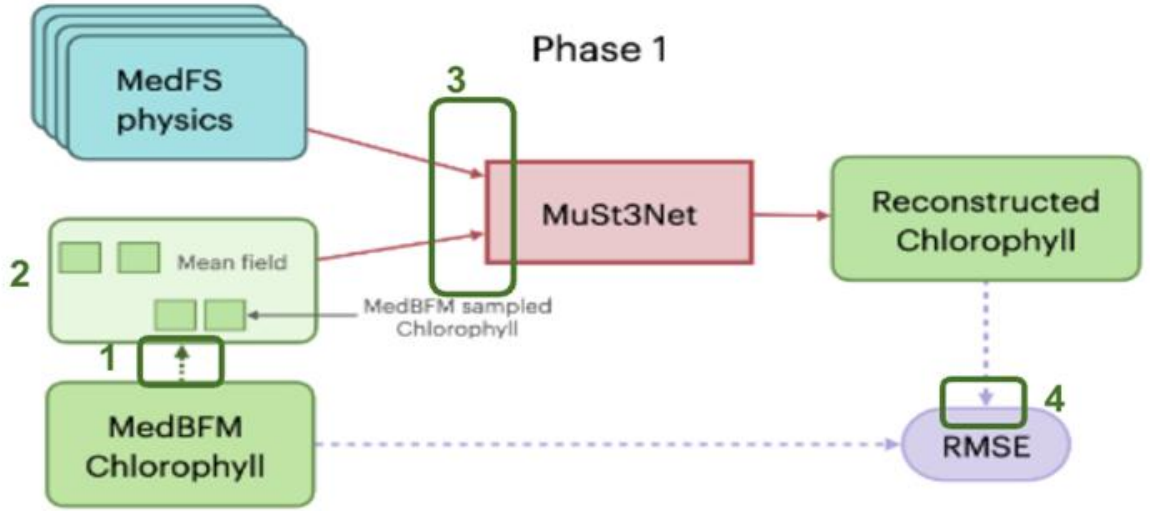
WORKFLOW:

1. the network **MuSt3Net** learns how to emulate the numerical model
2. the network **MuSt3Net** learns how to properly include BGC-Argo float data in an already trained framework

2 step training procedure

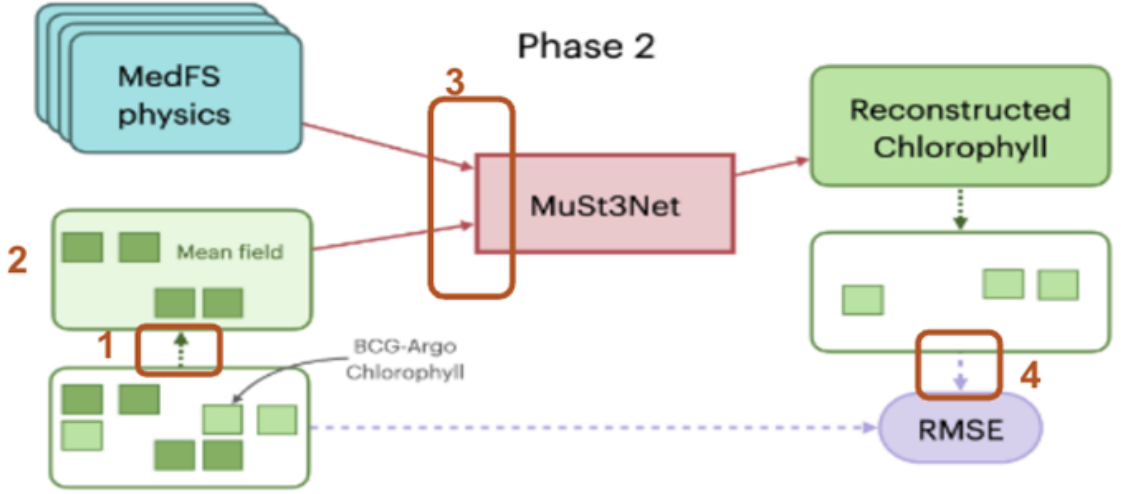
Phase 1:

1. Sub-sampling of profiles from BIO tensor
2. Fill in with mean-field value
3. Concatenation of physical and BIO tensors
4. Loss: comparison between 3D CNN output and 3D BIO ground-truth



Phase 2:

1. Sampling of train and test floats data
2. Fill in with mean-field value
3. Concatenation of physical and chlorophyll tensors
4. Loss: comparison between test profiles (model output profiles vs BGC-Argo profiles)



Experimental set-up

- Physical data: CMEMS physical reanalysis MYP-MED-006-004 (NEMO-OceanVar)
- Biogeochemical variables: Chlorophyll from CMEMS OGSTM-BFM (without DA)
- period: 2019-2021

- **Space and time resolution:**

Hyperparameter	Value
Latitude range	30°N –46°N
Longitude range	2°W–36°E
Depth range	0–300 m
Time range	2019–2021
Latitude resolution	8 km
Longitude resolution	8 km
Depth resolution	10 m
Time resolution	Weekly

- **Input pre-processing:**

Hyperparameter	Value
BFM profiles sub-sampling (per week)	200
BGC-Argo float profiles (per week)	40

- **Hyperparameters:**

Hyperparameter	Value
Epochs (Phase 1)	200
Epochs (Phase 2)	20
Learning rate	0.001 (both phases)
Hidden layers	6
Dropout rate	0.15
Activation function	ReLU
Kernel size	$3 \times 3 \times 3$
Trainable parameters	300097

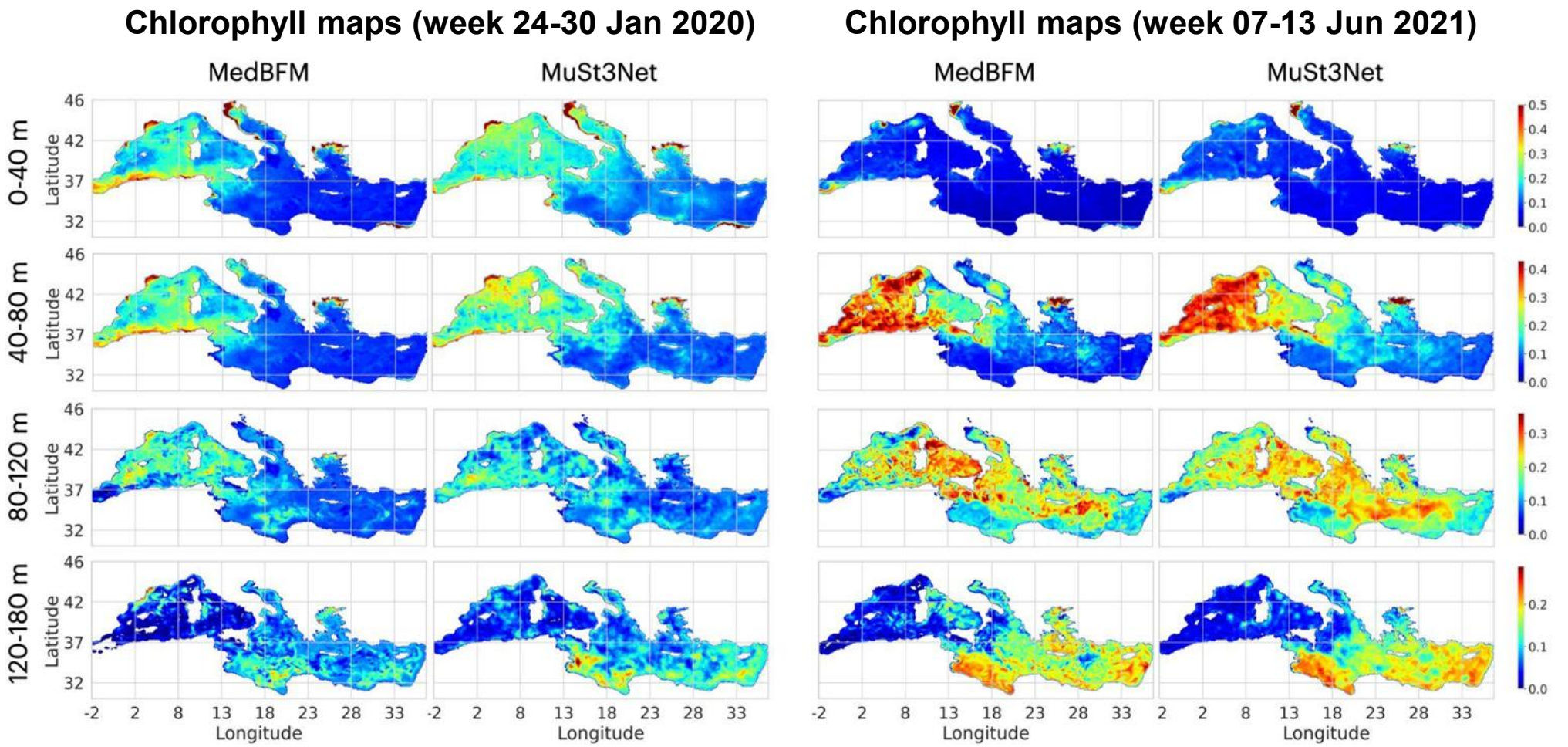
Results: overview

- **Maps Phase 1** : MuSt3Net works as an emulator of the OGSTM-BFM model
- **Maps Phase 2**: MuSt3Net integrates chlorophyll BGC-Argo float
- **Profiles**: details of integration of single profiles on different sub-basin (NWM, SWM, TYR, ION, LEV)
- **Error evaluation**: statistics on the 2-step integration procedure quality
- **Hovmöller**: MuSt3Net reconstructs also the temporal dynamics of chlorophyll

Results Phase 1: OGSTM-BFM model vs MuSt3Net maps

Phase 1: Comparison between OGSTM-BFM chlorophyll maps and MuSt3Net chlorophyll maps

Similar chlorophyll patterns: MuST3Net properly reproduces chlorophyll patterns, both in open sea and near coasts



RMSEs

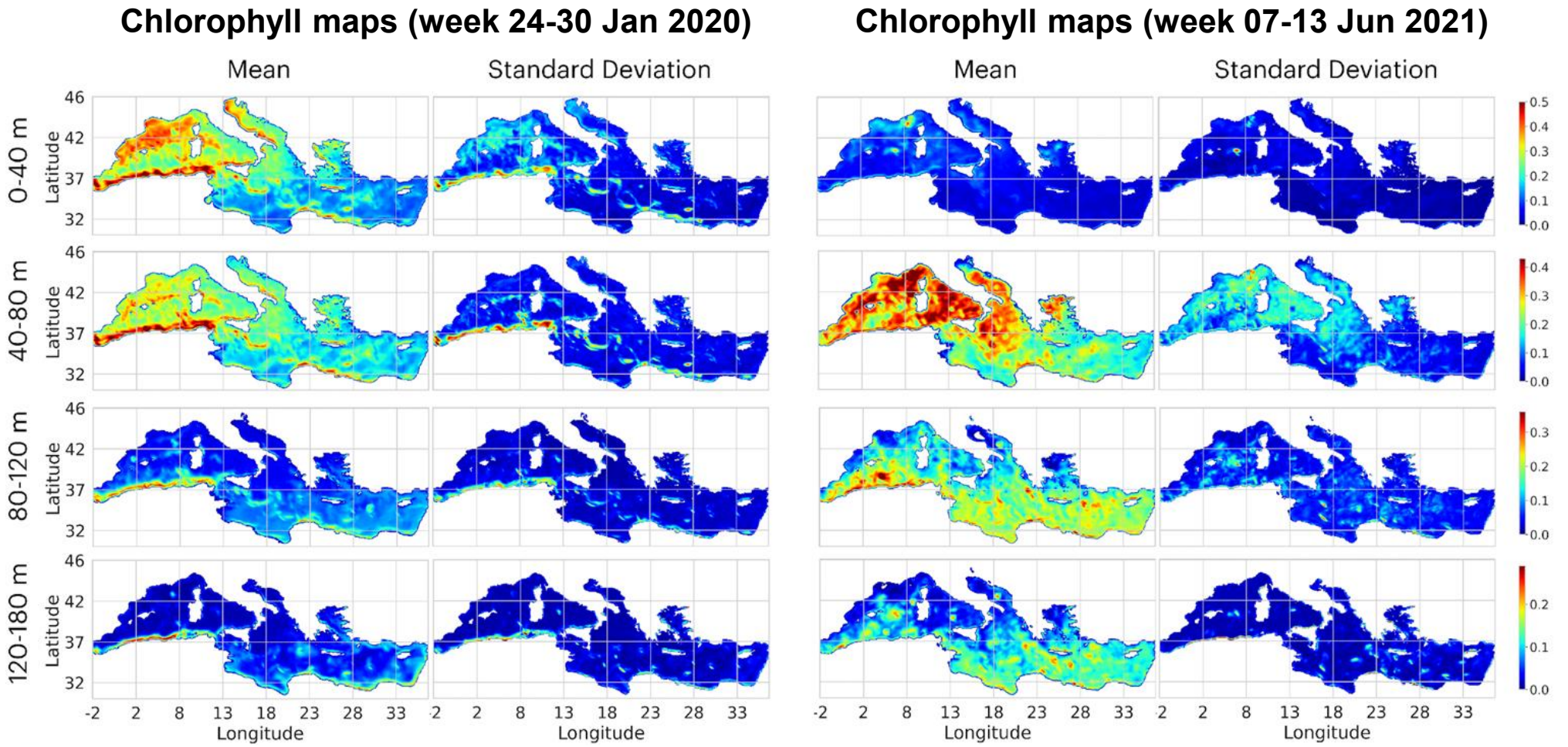
Seasons	Loss (mg m^{-3})
Winter	0.064
Spring	0.093
Summer	0.068
Fall	0.051

Geographical areas	Loss (mg m^{-3})
NWM	0.087
SWM	0.096
TYR	0.091
ION	0.052
LEV	0.060

Results Phase 2: MuSt3Net maps

Phase 2: reconstruction of chlorophyll maps after the Argo-float data integration

Similar chlorophyll patterns: MuSt3Net properly integrates BGC-Argo floats without losing information gained from phase 1 (BIO emulator)

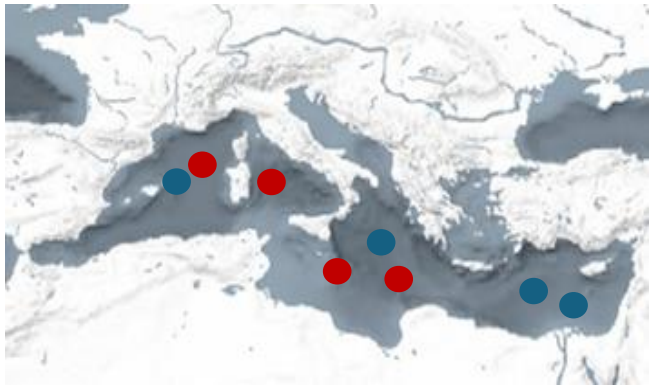


RMSEs

Seasons	Loss (mg m^{-3})
Winter	0.065
Spring	0.099
Summer	0.071
Fall	0,055

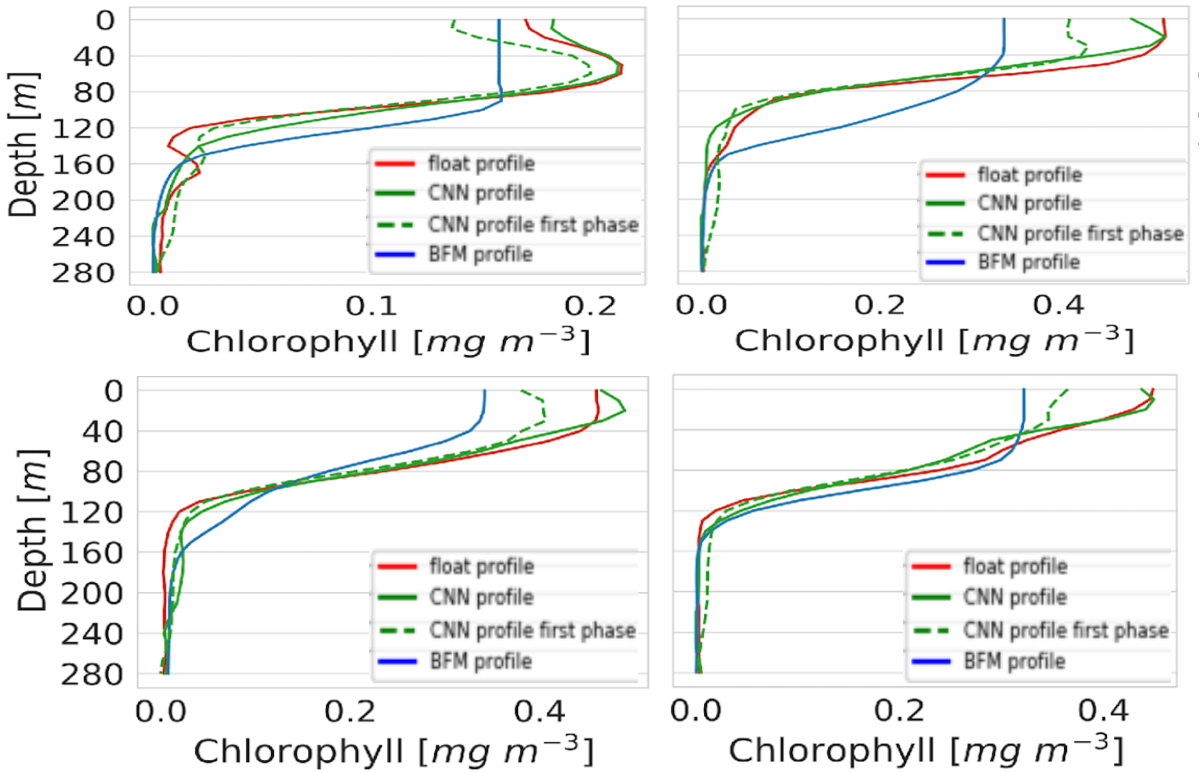
Geographical areas	Loss (mg m^{-3})
NWM	0.086
SWM	0.098
TYR	0.094
ION	0.053
LEV	0.065

Results: chlorophyll profiles

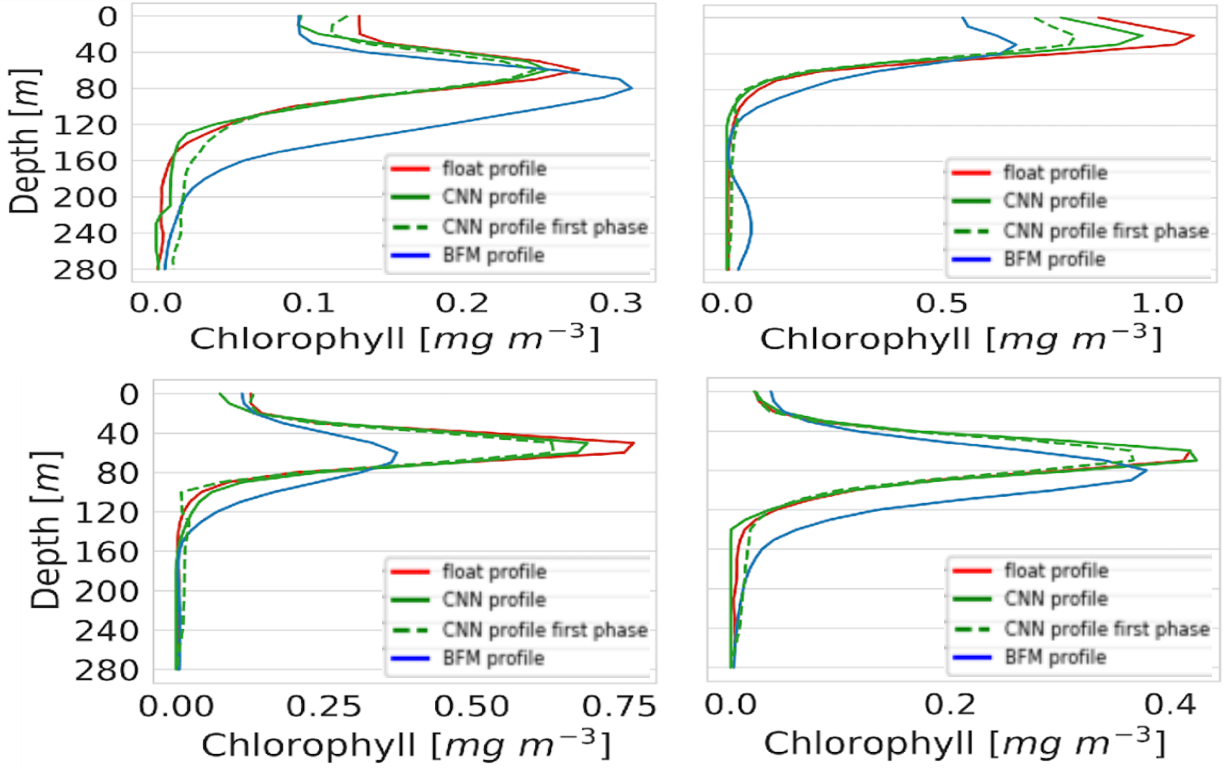


- **Comparison on chlorophyll profiles**: checking the ability of CNN of properly integrate floats data and to reproduce shapes of surface layer bloom and deep chlorophyll maximum
- CNN integrates float data into an already trained framework

Profiles of winter weeks (4ª, 10ª) in the NWM, ION and LEV Sea



Profiles of a summer week (17ª, 24ª) in the NWM, TYR and ION Sea



Results: profiles quantitative evaluation

Comparison of **RMSE**, **MAE**, **BIAS** and **r** across **Seasons** (Table 1) and **Regional Seas** (Table 2)

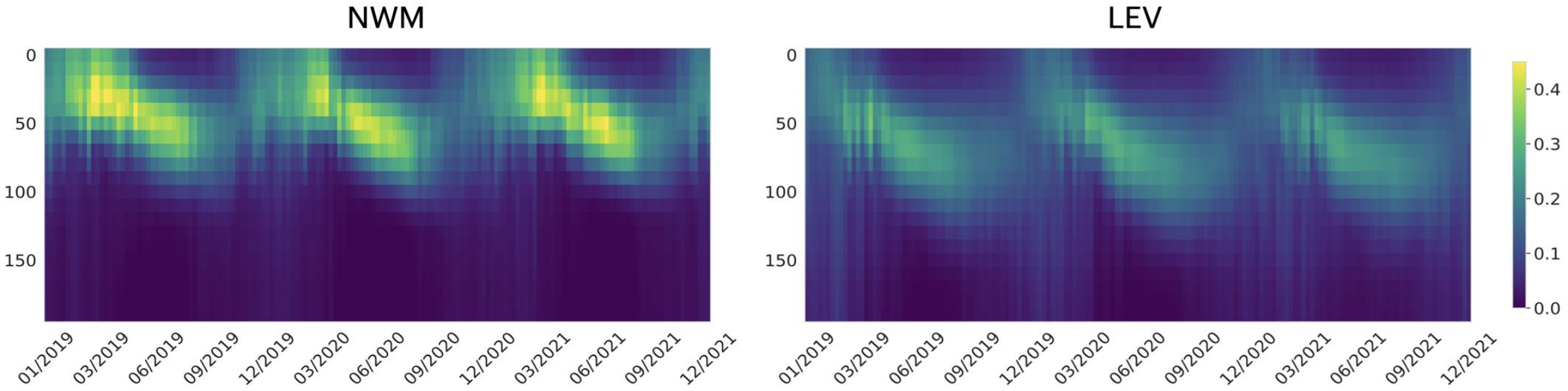
- Higher accuracy observed in Summer and Fall: challenges in predicting surface blooms
- Lower errors in Ionian and Levantine Seas: higher Argo-float coverage, lower chlorophyll variability

Season	RMSE $[mg\,m^{-3}]$		MAE $[mg\,m^{-3}]$		BIAS $[mg\,m^{-3}]$		r	
	Train	Test	Train	Test	Train	Test	Train	Test
Winter	0.065 ± 0.012	0.095 ± 0.009	0.060	0.066	-0.023	-0.025	0.99	0.99
Spring	0.099 ± 0.015	0.124 ± 0.017	0.056	0.064	-0.016	-0.021	0.97	0.96
Summer	0.071 ± 0.010	0.078 ± 0.012	0.043	0.045	-0.002	-0.003	0.99	0.98
Fall	0.055 ± 0.008	0.049 ± 0.005	0.034	0.031	-0.001	0.001	0.99	0.99

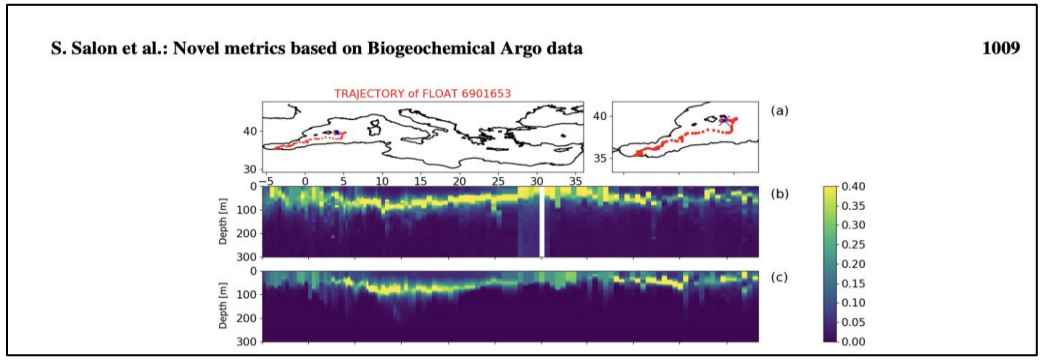
Season	RMSE $[mg\,m^{-3}]$		MAE $[mg\,m^{-3}]$		BIAS $[mg\,m^{-3}]$		r	
	Train	Test	Train	Test	Train	Test	Train	Test
NWM	0.086 ± 0.029	0.096 ± 0.032	0.049	0.059	-0.015	-0.024	0.98	0.97
SWM	0.098 ± 0.031	0.086 ± 0.034	0.059	0.050	-0.027	-0.019	0.99	0.99
TYR	0.094 ± 0.038	0.099 ± 0.042	0.063	0.063	-0.028	-0.025	0.97	0.94
ION	0.053 ± 0.020	0.055 ± 0.019	0.043	0.044	-0.003	-0.003	0.98	0.99
LEV	0.065 ± 0.018	0.066 ± 0.021	0.045	0.046	-0.006	-0.004	0.99	0.99

Results: Hovmöller plots

Although CNNs focus on spatial reconstruction, our method also accurately resolves the temporal dynamics of chlorophyll at basin scale:



- Higher chlorophyll values in western basin than in eastern basin
- Correct DCM and superficial blooms



Salon et al. 2019

Conclusion & future works

2-step CNN: a viable approach for model and data (BGC-Argo) fusion to predict BGC variables

- **PRO**:
  - it predicts the whole 3D domain and preserves spatial information
  - faster time to solution: 1d for 1Y ModelBFM+DA float -> 2h for 1Y trained CNN
- **CONS**:
  - quality depends on availability of observed profiles
  - ground-truth quality difficult to assess, however uncertainty based on ensemble

- **Architecture improvements**: different architecture, which integrates seasonality
- **Adopting different strategies:** training the CNN on a coarse resolution, and including super-resolution steps
- **Other variables**: test the prediction method to the other biogeochemical variables (oxygen, nitrate, bbp700/POC)
- **Other data:** integration of satellite data

Publication: “Tonelli, T., Cossarini, G., Manzoni, L., & Pietropolli, G. (2026). Two-phase CNN for model data fusion: Predicting 3D chlorophyll-a in the Mediterranean Sea. *Ocean Modelling*, 102707.”

Thank you for your attention

Ocean Modelling
Volume 201, April 2026, 102707

Two-phase CNN for model data fusion: Predicting 3D chlorophyll-a in the Mediterranean Sea

Teresa Tonelli [b a], Gianpiero Cossarini [b], Luca Manzoni [b a],
Gloria Pietropolli [a]

https://www.sciencedirect.com/science/article/pii/S1463500326000314

NECCTON
NEW COPERNICUS CAPABILITY FOR TROPHIC OCEAN NETWORKS

HorEU project

OGS

1924 2024

UNIVERSITÀ DEGLI STUDI DI TRIESTE

RMSE across depth layers: A comparative analysis of 1 and 2 phase

- **Phase 2,** which integrates BGC-Argo float data, consistently reduces RMSE across all depth layers compared to the emulator trained solely on MedBFM data (**Phase 1**).

- RMSE varies with depth, showing higher values near the surface and around the deep chlorophyll maximum (DCM).

These patterns highlight the increased difficulty of modeling **surface layers** and the **DCM**.