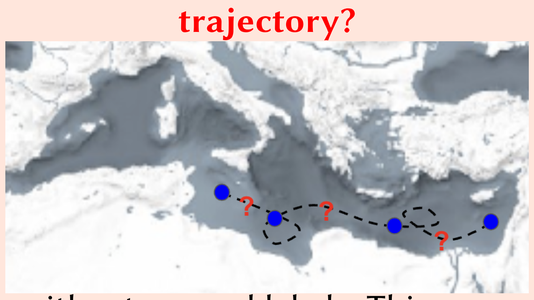
Filling the Ocean's Gaps: a Self-Supervised Neural Network for Argo Profiles Data Augmentation

Teresa Tonelli (OGS, UniTS), L. Manzoni (UniTS), G. Cossarini (OGS)

Self-Supervised 1D-UNet for Argo Profiles Data Augmentation

Expanding the volume and quality of observations is essential to support the growing range of Argo-driven models and applications. Developing methods to generate augmented observations offers a promising tool to mitigate this gap.
This works develops a 1D-UNet, a CNN-based encoder–decoder network with skip connections, that generates augmented profiles of biogeochemical variables by interpolating existing ones, thereby yielding new measurements for intermediate time steps and spatial locations.
Given the profiles at time t_0 and t_1, the network is trained to predict profiles at intermediate time steps $t_{\frac{1}{3}}$ and $t_{\frac{2}{3}}$, which correspond to profiles in intermediate points along the float trajectory.

Can we fill the gaps between Argo profiles along a float trajectory?



The novelty of our approach lies in the **Self-Supervised Learning** technique, which lets the network learn directly from the input data without manual labels. This structure, implemented here through a **Dual Cycle Interpolation**, enables the generation of augmented profiles using only existing observations, mitigating the need for extensive labelled data, accelerating model convergence and improving its generalisation capability.

Self-Supervised Learning: the Dual Cycle Interpolation

- Our 1D-UNet is trained in a self-supervised mode, exclusively relying on pre-labeled profiles at regular time-steps.
- The 1D-UNet self-supervised training is based on a cycle consistency approach: executing an operation and its inverse consequently, the network can assess the consistency of its computation by evaluating the discrepancy between the reconstructed data and the original input.

Structure of Dual Cycle Integration

- **Stage 1**: given a time interval δt and 2 sequence of consecutive profiles $[x_i, x_{i+\delta t}]$ and $[x_{i+\delta t}, x_{i+2\delta t}]$, the network M predicts the intermediate profiles:

$$[x^{[1]}_{i+\frac{\delta t}{3}}, x^{[1]}_{i+\frac{2\delta t}{3}}] = M(x_i, x_{i+\delta t})$$

$$[x^{[1]}_{i+\frac{4\delta t}{3}}, x^{[1]}_{i+\frac{5\delta t}{3}}] = M(x_{i+\delta t}, x_{i+2\delta t})$$

- **Stage 2**: the same **stage 1** procedure is applied on profiles with a time window shifted of $\delta = \delta t/3$:

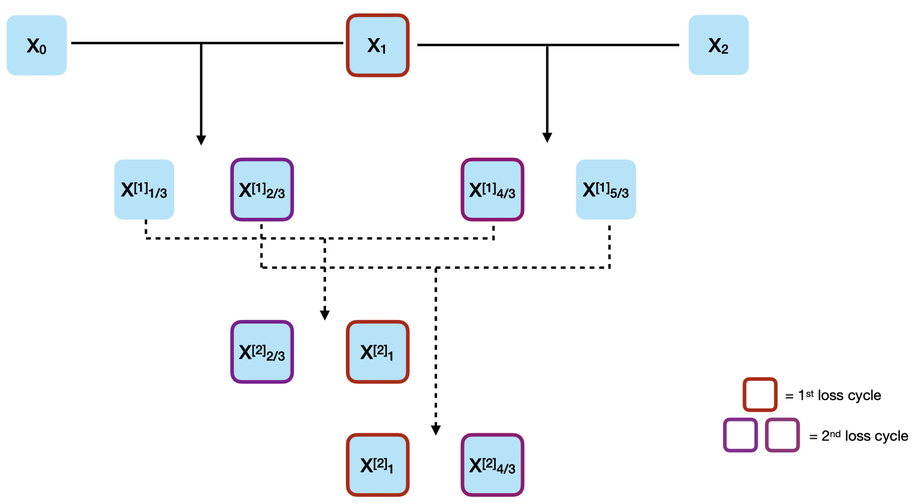
$$[x^{[2]}_{i+\frac{2\delta t}{3}}, x^{[2]}_{i+\delta t}] = M(x_{i+\frac{\delta t}{3}}, x_{i+\frac{4\delta t}{3}})$$

$$[x^{[2]}_{i+\delta t}, x^{[2]}_{i+\frac{4\delta t}{3}}] = M(x_{i+\frac{2\delta t}{3}}, x_{i+\frac{5\delta t}{3}})$$

- The cycle consistency requires that $x_{i+\delta t}$ predictions should match the original $x_{i+\delta t}$ data; additionally, intermediate profiles computed multiple times throughout the cycle should remain consistent.

Dual Cycle Interpolation Framework

Structure of the Dual Cycle Training procedure to generate augmented Argo-float



Discussion and Future Works

Purpose and Advantages

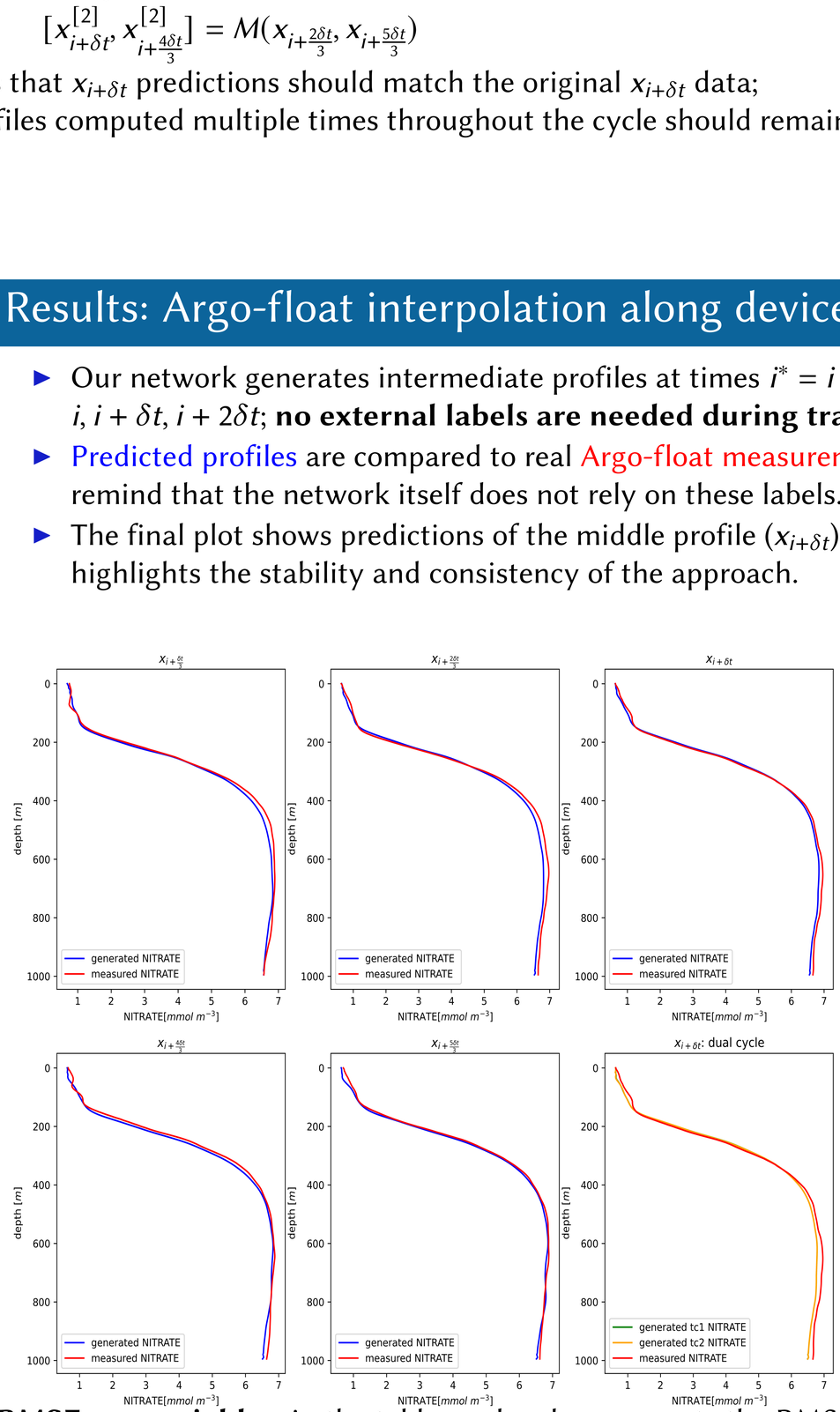
- This method allows the duplication of Argo-float data through the generation of additional profiles in unseen spatial locations.
- More coverage for biogeochemical variables.
- Our approach does not depend on δt.

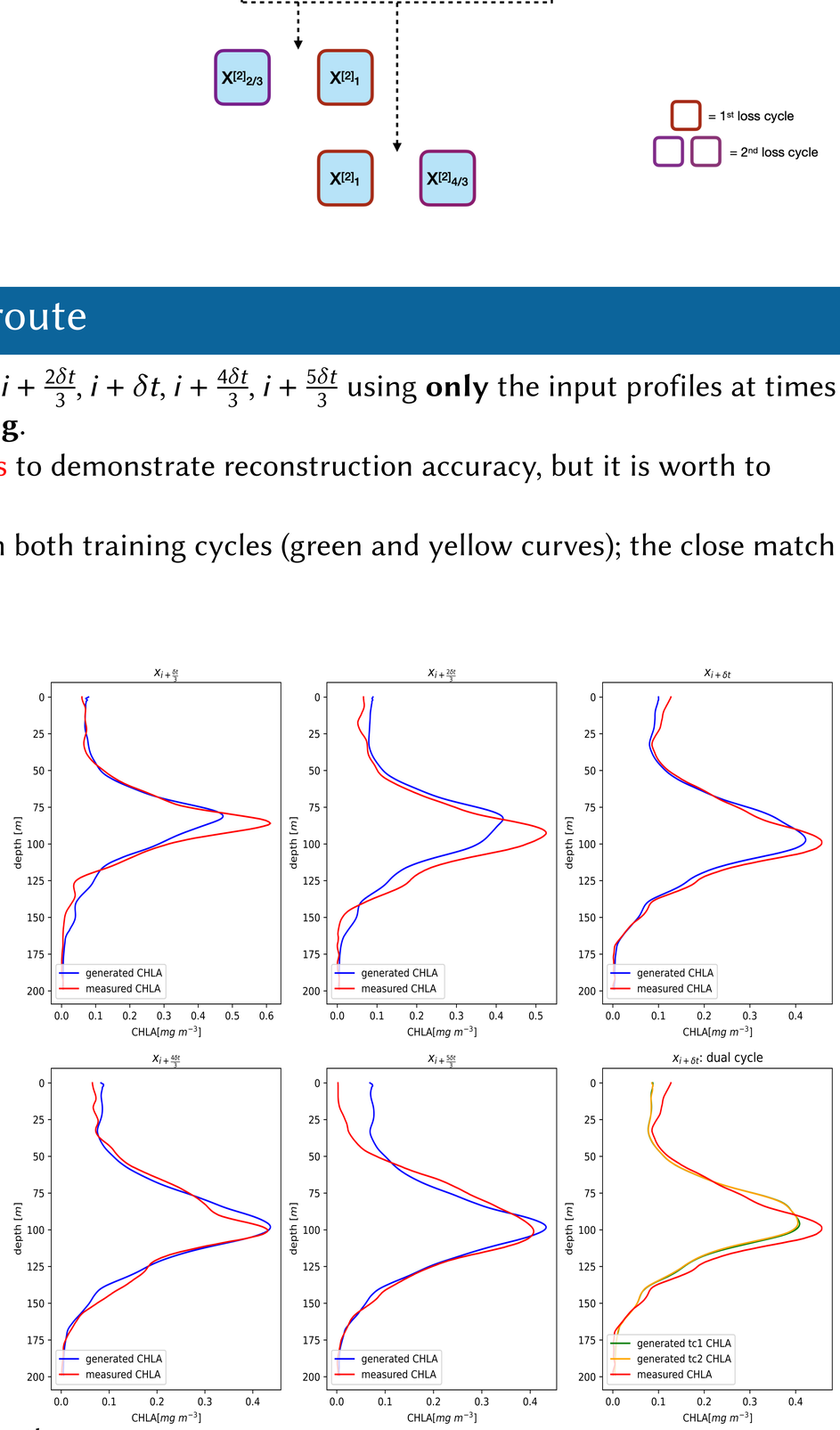
Validation:

- External validation through additional Argo-float in the Mediterranean Sea.
- Comparison with linear interpolation to check the effectiveness of our approach.

Results: Argo-float interpolation along device's route

- Our network generates intermediate profiles at times $i^* = i + \frac{\delta t}{3}, i + \frac{2\delta t}{3}, i + \delta t, i + \frac{4\delta t}{3}, i + \frac{5\delta t}{3}$ using **only** the input profiles at times $i, i + \delta t, i + 2\delta t$; **no external labels are needed during training**.
- Predicted profiles are compared to real Argo-float measurements to demonstrate reconstruction accuracy, but it is worth to remind that the network itself does not rely on these labels.
- The final plot shows predictions of the middle profile ($x_{i+\delta t}$) from both training cycles (green and yellow curves); the close match highlights the stability and consistency of the approach.





RMSE per variables: in the table, each column reports the RMSE over the entire dataset for a given intermediate position (e.g, $x_{i+\frac{\delta t}{3}}, x_{i+\frac{2\delta t}{3}}, \ldots$), computed by averaging the errors across all samples at that position.

	$x_{i+\frac{\delta t}{3}}$	$x_{i+\frac{2\delta t}{3}}$	$x_{i+\delta t}$	$x_{i+\frac{4\delta t}{3}}$	$x_{i+\frac{5\delta t}{3}}$
nitrate ($mmol\ m^{-3}$)	0.199	0.170	0.122	0.198	0.173
chla ($mg\ m^{-3}$)	0.048	0.056	0.025	0.042	0.049