Final Author comment for RC1

Review of egusphere-2025-2782 “Estimating the AMOC from Argo Profiles with Machine Learning Trained on Ocean Simulations” by Yannick Wölker , Willi Rath , Matthias Renz , and Arne Biastoch

My expertise as a reviewer is mainly focused on the data science part, the parameter estimation, machine learning, statistics, error handling, and significance considerations, in the frame of oceanographic research questions.

- Response: We thank the reviewer for their thoughtful and constructive comments on our manuscript. We very much appreciate the reviewer’s expertise in data science and machine learning within the context of oceanographic applications, which has helped us to clarify and strengthen the explanatory aspects of our study. We have revised the manuscript accordingly, as detailed below, and believe these changes have substantially improved the clarity of the paper.

General:

- The manuscript represents an important and interesting study on the potential of Argo floats to be used to estimate the AMOC, specifically the geostrophic part.

- I recommend publication after some minor clarifications, improvements and corrections.

- You mention the challenges of handling the irregular Argo data, which is reasonable. Then you overcome these difficulties with the embedding and graph-based NN approach, which is technically a very smart solution. However, you did not try an interpolation approach, bringing the Argo data on a regular grid and using those data as input for the feedforward NN. Thus, we don’t know if your graph-based approach is superior. In the end your results are good, which probably justifies your approach, however, for me it’s always the question if these results could have been achieved with simpler methods?

- Response: We thank the reviewer for this interesting comment. We tested their mentioned method in an earlier development phase and quickly moved on due to the following reasons. First, the classical binning comes with a bunch of assumptions about the spatial shape, temporal distance, and connectivity of binning cells. In the process of defining these, we found ourselves in a difficult situation, which would require a lot of testing. However, the results may be valid only in this particular study region. Using the graph-based approach with the learned clustering solves this in a pure data-driven fashion, which we envision to be more practical for future use and applicable for other tasks. Second, a feedforward NN (FFN) on top of a static binning was performing worse because the amount of Argos per bin is variable. As FFN have

a static and global connection, similar to the reason why Convolution Neural Networks are used in image processing, spatial or temporal shifts of correlations require explicit learning in an FFN, while this is inherent in GNNs for irregular topologies and CNNs for regular topologies. Overall, we think that a too deep investigation of the method is beyond the scope of this study, with our main goal showing the applicability of a method that is mostly data-driven and can therefore be reused in different scenarios. Based on the feedback of the reviewer also in the `Specific` section, we restructured the second part of the 'Processing of Argo profiles' section. We added the following text in line 246 to motivate our design choices more. "A static clustering on all data points, could create such a structured input. Around our targeted latitude this would amount to a zonal binning (Willis, 2010; Hernández-Guerra et al., 2010) but the set of Argo floats, especially for the shorter target time scales, show a heterogeneous distribution that would require carefully hand-crafted cluster boundaries. A classical binning comes with even more assumptions about shape, distance, and connectivity of the bins, which would require a thorough testing of hyper parameters. However, we aim in this work for a data-driven mapping function which identifies a structured numerical representation (n-dimensional vector)"

- Your approach, based on model data, shows that there is potential to reconstruct the AMOC utilizing Argo floats. However, for an application with real data, not enough data is available for a NN approach. So, what is not fully clear to me from the manuscript is, when we can reach “enough data”? Or, regarding to your discussions, is the only solution transfer learning, and enough data will not be available in a reasonable near future? Can you please clarify that?

- Response: We appreciate the reviewer's question as it is an important point. In general, the quality of training data is determined by how well it covers the expected states during inference. This means we would potentially reach "enough" training data when sufficient heterogeneous ocean states were observed, from which general knowledge can be extracted that is most likely to match the ocean state during inference. However, this statement is vague because the question of how the AMOC and its driver change in the future is under heavy discussion. The "enough data" also depends on the targeted time scale. While on short time scales, the current observational data contains different realisations of more frequent signals like the seasonal cycles, this is not true for yearly or decadal signals, which we would be most interested in by using Argo profiles. Based on this consideration, we found simulations to be a promising testbed for the question "What if we had plenty of observed years?" to test our reconstruction. For long time scales, we see the potential in the transfer learning approach with pre-trained reconstruction on large ensembles to cover more heterogeneous ocean states and a fine-tuning phase, much like the hyped foundation models. Considering our demonstrated performance, we see on shorter time scales the potential for training a hybrid reconstruction with assimilated simulation and Copernicus data. In the next decade, enough Argo measurements could be reached to train a stable reconstruction that would assist in scenarios where moored arrays like RAPID may have data gaps. We added to the paragraph with about the applicability and transfer learning methods a distinction between long and short timescales (l.583) "For a routine application \[...\] this study also identified limitations. The training within VIKING20X with virtual Argo profiles has

shown that an AMOC reconstruction on interannual or longer time scales requires large amounts \[...\]However, for shorter timescales and use cases like the filling of smaller temporal gaps the observational data within the next decade could be sufficient to train an AMOC reconstruction purely on real-world data."

- It is important to thoroughly always differentiate between the real elements in this study (AMOC, Argo, etc.) and the simulated. Please check all text.
    - Response: We thank the reviewer for the valuable feedback and agree that a sharp line between the observational and simulated data is essential. We changed the text in the introduction to make a clear statement, that this study is inspired by real-world observation strategies but uses only simulated ocean data for the reconstruction. (l. 94) "In this work, we demonstrate how and to what extent the AMOC can be reconstructed in an ocean model setting from simulated measurements that mimic widely available observational products using supervised machine learning. All reconstruction inputs are extracted from the ocean simulation together with virtual Argo floats which have the same spatio-temporal distribution as in the real-world, and then tested against the total AMOC calculated on the same simulation.". Additionally, we made sure to be consistent with the wording "virtual Argo profiles" whenever we refer to Argo profiles extracted from a simulation.

Specific:

- Line 9-10: Add "... AMOC can be potentially data-drivenly ..."
    - Response: We thank the reviewer for their comment and changed the abstract accordingly to "...AMOC can be potentially reconstructed by Argo profiles in a data-driven fashion"

- Line 21: Are you referring to the North Atlantic Deep Water? But that is colder and saltier not fresher, or?
    - Response: Here we refer to NADW which is typically fresher (about 35 psu) compared to the Gulf Stream (36-37 psu) flowing above.

- Line 27: You say that ocean and climate models often fail to simulate the AMOC, but nevertheless you go for a full model analysis to draw inference on the real world.
    - Response: We thank the reviewer for this valuable feedback. Our goal is to mention the deficits of models to point out the extra work that would be necessary for an inference in the real world. The goal of this work is to use models as a physically consistent testbed for the AMOC reconstruction. The bias of the AMOC in an ocean model comes from the global balance which is out of scope for our AMOC reconstruction. The other benefit to ocean simulations is the longer data horizon than the real world. The AMOC biases in the ocean models are introduced by global balances. We changed the text to make clear that, despite the inherent errors of ocean models, these are a valid way to test our reconstruction assumptions.

- Line 53: Again make clear that you are not using real Argo.

- Response: We understand that we have to be clearer in the distinction between real Argo measurements and the “virtual Argo profiles”, which we used from ocean simulations. We change the sentence and added the word concept, highlighting that the idea is motivated from real Argo profiles. Also, we specified that we use "simulated Argo floats from ocean models".

- Line 93-95: I understand that the authors are going for a NN approach and real data is too limited in this case. However, why do the authors think that simpler approaches like linear regression may not be sufficient?
  - Response: We thank the reviewer for this important comment. In our view, simpler approaches would require structured data, which does not hold for the spatial and temporal distribution of Argo profiles. Similar to our answer the reviewer’s third general comment. We refer to this observation in the previous paragraph (l. 71) "So far, these approaches rely on spatial and temporal binning \[...\] Most machine-learning approaches require structured input data \[...\] ". Overall, the non-linear mechanism essentially allows for learning implicit data imputations, which would need to rely on subjective choices in linear approach and the binning of Argo profiles. The addition of the linear approach would not be possible in an objective sense (because of all the required choices) and would substantially expand the manuscript in length and scientific content beyond the current scope.

- Line 110: This sentence is confusing, what do you mean by “widely available observations”? If I understood correctly you are not using real observations.
  - Response: We acknowledge the unclear statement and thank the reviewer for their finding. The text was removed in a process of removing redundancy for the sake of preciseness. The current paragraph is: (l. 94) “In this work, we demonstrate how and to what extent the AMOC can be reconstructed in an ocean model setting from simulated measurements that mimic widely available observational products.% using supervised machine learning. Reconstructing the AMOC on different time scales from 10 days up to five years, we find that the importance of geostrophic transport in the ocean interior becomes more pronounced, with longer time scales. To capture this, we use virtual Argo profiles as our observational input, leveraging their insight into the ocean interior while addressing their sparse and irregular sampling with a graph-based neural network approach.”

- Line 113: You are not using Argo data!
- Line 110-124: Start the whole paragraph with explaining that you use simulated data.
  - Response: We agree with the reviewer that this paragraph requires a more transparent communication of the used values. We moved the sentence "All reconstruction inputs are extracted from the ocean simulation..." from the next paragraph to the top of this one. Also, we made sure to refer to virtual Argo profiles for our reconstruction, because these are extracted from the ocean simulation with spatial and temporal distribution from real Argos.

- Page 5, Figure 1: c,d,e: Shouldn’t the middle x-label be the other way round 2024/1958?
  - Response: We agree with the reviewer and changed the x-axis for the x-ticks in the middle to '2024/1958'.

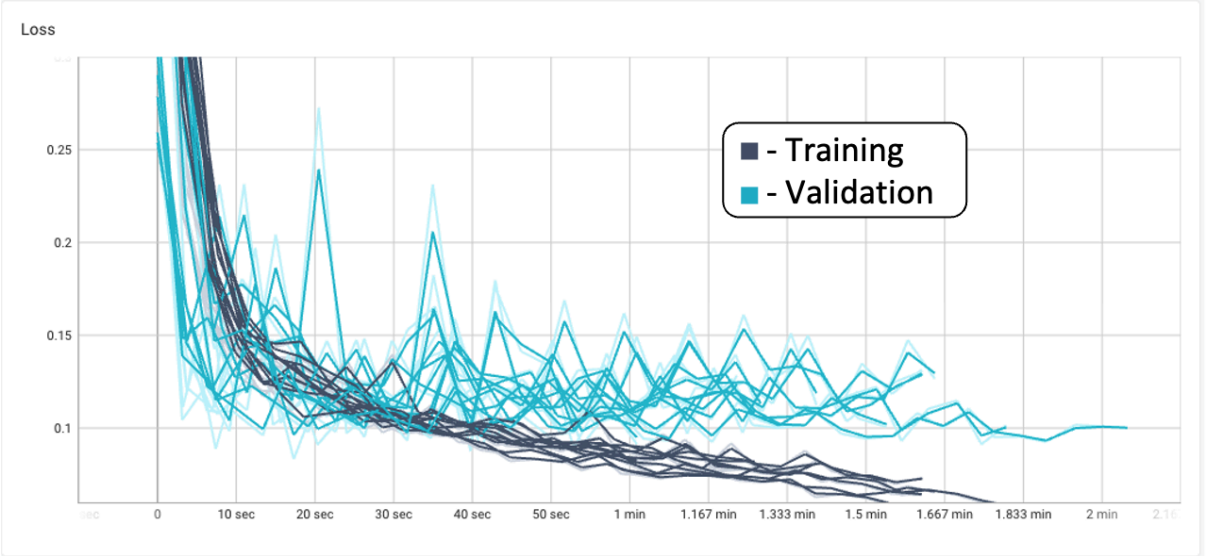
- Line 182: First Argo floats have been deployed since 1997.
    - Response: We thank the reviewer for pointing out our inaccuracy with the start of the Argo deployment. Based on our cited literature Riser et al. 2016 we changed the first deployment year to 1999.

- Line 284: Difficult to understand. What is the “trained reconstruction”? And what means “the trained reconstruction is able to reconstruct...”?
    - Response: We changed the wording to stress that we mean by a trained reconstruction the neural network with the optimised weights from the training process. The idea is to avoid confusion in the general audience between ocean model and the often used models term in machine learning. We decided to make a clearer statement here and refer in the following manuscript to the trained reconstruction. We changed the wording into (l. 229)"...summarized as learning a set of parameters, represented by a trained reconstruction model (following referred to as trained reconstruction), which uses..."

- Line 303-327: You are saying that if you neglect the spatial info on Argo data, you can utilize a suitable neural network architecture. In the following you say you keep the spatial component using SUSTeR. I do not understand what in the end you do. In addition, understanding SUSTeR and explanations about traffic are not helpful. I suggest to remove this explanation and refer to the publication. Instead please make clearer what you have actually done in the end.
    - Response: We thank the reviewer for this comment. To the first part of the comment, we removed the sentence about the Transformer and LSTM to avoid confusion about our work. Making it clear that we aim to incorporate both the spatial and temporal components of the virtual Argo floats. We tried to give an example that keeping only the temporal structure would be borderline structured and could be processed by sequential models, making the point that the important spatial structure would still be missing. (l.241) "Argo profiles pose a demanding challenge for the design of neural networks, as neural networks are normally designed to handle structured inputs. For spatio-temporal data points, structured input requires a constant number of measurements from constant locations with a static topology (e.g. a grid structure, or a graph), which is both not true for moving Argo floats."
    - Furthermore, we removed the technical part of the traffic prediction comparison and added a brief intuition of why SUSTeR is a good match for the Argo problem. We want to mention the different applications to build a bridge for those who will take a look at the SUSTeR paper. (l.260)"SUSTeR set out with the goal to handle unstructured traffic observations and find a general representation of city traffic, much like the unstructured virtual Argo profiles with the goal to find a general representations of the ocean state." Lastly, we rewrote how SUSTeR works and provide a more intuitive description of the learnable clusters, which can dynamically group the virtual Argo profiles.

- Line 329-- Sect. 3.2: Regarding the training procedure of a NN, it would be interesting to see a “loss curve”. Often these loss curves are given for the performance of the model on the training set, during the training, as well as on the validation set (unknown).
    - Response: We value the feedback of the reviewer and discussed that thoroughly. We see only limited benefit in including the training/validation loss curves, as they did not

show an interesting pattern. The training uses the early stopping strategy, stopping with a patience of 10 epochs by a maximum of 50 epochs for each experiment. We think that the manuscript would not benefit from analysing the training process in such a technical way. However, we did so during the development to check for problems during the training process. We added to this Response Letter the training curves (train & validation) for the 11 ensemble members of the seasonal time scale from the experiment in section 4.1:



- Line 374, Eqn. 4: I think the denominator is not Var(y), but the total sum of squares \sum{ (y_i-\overline{y})2 }.

- Response: We thank the reviewer for pointing this out. The reviewer is correct that in Eq. 4 the denominator should be the total sum of squares, \sum_{i = 1}^{N} (y_i - \overline y)^2, rather than Var(y). We have corrected this in the revised manuscript.

- Page 16, Fig. 4: I suggest to plot the reconstruction curve (blueish) on top of the ground truth (green) to better see it.

- Response: We thank the reviewer for this good idea. We changed this and all other plots that contain reconstruction time series accordingly.

- Line 595: I guess you again randomized the Argo input data, not leaving it really out? Please mention in the text.

- Response: We value the feedback of the reviewer, finding this unclear execution. We added the following sentence to describe the experiment in more detail. (l.489) "We retrained the reconstruction in a neural network without argo-related inputs. This removes the influences of the Argo profiles completely from the reconstruction. " By performing a complete retraining, we decoupled the influence from the Argo profiles completely. In this way, it is not necessary to find a strategy (like randomized input) to remove the features, because it was needed for Figure 7 in which we only decoupled the evaluation influence.

- Line 619: If I understand correctly by reading the full paragraph, the reason for no added value of deep Argo is probably just caused by not having enough training data. Thus the influence of deep Argo stays rather unknown. If that is true, please mention already here.
    - Response: We thank the reviewer for their feedback. We changed the latter part of the sentence and added a statement about the missing heterogeneous training data for the deeper layers. (l. 510) "Across all timescales, we found that the added value is limited by the seen training data, which did not cover heterogeneous ocean states in the deeper layers, leaving the possible influence of Deep Argo floats undetermined in our training setup."

Corrections:

- 41: Rewrite this part, which sounds strange “... cables that measurement ...”

- 48: Space missing “... balance(Mc ...”

- 155: Figure ??

- 259: “... an high ...” → “ ... a high ...” and “... an dedicated ...” → “... a dedicated

- 309: (?

- 396: “... due to larger ...”

- 399: “The the ...”

- 400: “... limits the compare ...” → sounds strange, please rewrite.

- 427: “data(Jiang” → space

- 478: “... and the also the ...”

- 536: “brach” ? → branch

- 542: “... due to the ...”

- 543: “... transport(Mc” → space

- 630: Change “We test if the test data lays within the training data and its ...” to “We investigate if the test data lie within the training data and if its ...”

- 706: “... amounts of diverse ...”, delete: “... set of ...”

- 736: “... mentioned,the ...” → space

- Response: We thank the reviewer for carefully pointing out the spelling mistakes and rephrasing suggestions and once again for their careful evaluation and constructive feedback. The revisions have helped us to improve the presentation and methodological description of our work. We hope the changes meet the reviewer’s expectations and that the revised manuscript has an improved quality.