A review of **Correction of Air-Sea Heat Fluxes in the NEMO Ocean General Circulation Model Using Neural Networks**
By Andrea Storto et al.

Submitted to Geoscientific Model Development
Manuscript gmd-2024-185, https://doi.org/10.5194/gmd-2024-185

The authors present a data-driven method for "correcting" the air-sea fluxes in the state-of-the-art NEMO ocean general circulation model. The method is mimicking a sea-surface relaxation scheme that is commonly used in NEMO to constraint the sea-surface temperatures (SSTs) to external datasets, which would most often by remote-sensed satellite products. In practise, the "classical" SST relaxation method is enforced as a corrective term on the air-sea nonsolar heat flux, which is part of the model's surface boundary condition for the temperature.

The authors present an Artificial Neural Network (ANN) that mimicks this corrective heat flux term described above, and then its use and integration within the NEMO ocean model in place of the "direct" sea-surface relaxation, which requires knowledge of future SSTs, so is not possible for doing forecast. On top of it, the authors also illustrate the versatility of their methods by using their method over the 1960s, for which the observational network that would be used for constraining the model is sparse. Their results are promising and show that the ANN provides a viable alternative to classical SST relaxation, which could be used in other contexts.

The paper is well-structured, to the point and yet exhaustive. It presents a data-driven alternative to a classically used method, having the huge advantage of being predictive and not relying on an external dataset, therefore being applicable in other contexts, like forecasting and/or reanalyses over periods with poor observational coverage. The paper is quite technical, therefore making it a good candidate for GMD. I think that there are a few points on which the paper could be more accurate, and some scientific details which could be revisited. Therefore, my recommendation would be to ask the authors to take the following comments into account. Hopefully, this would make the manuscript stronger and better for the scientific community.

Major comments:

1) The manuscript directly introduces the method as a heat flux correction, as the title suggests. Technically, it is indeed the case, but the end goal is building a method for constraining the sea-surface temperature and reduce model biases. It is unclear whether this correction is due to fluxes being wrong – the only tangible thing is that these fluxes "suit" the model better. On a related note, the method might seem convoluted: why not simply adding an extra tendency on the SST? I think one of the main reasons, which the authors keep implicit, is that they can then directly, online integrate their ANN into the model via Fortran90 bindings. This is a strength which could be more put to the front. The method seems convoluted, but it is for a good reason, which is its integration into the classical model.

2) The variable importance scores (VIS) provide interesting insight, but it is tricky. The authors do acknowledge that air-sea interactions are strongly nonlinear, so that separating one variable from another is artificial. That said, acknowledging that the exercise is difficult to start with, and it is not possible to do a clear cut, I think that the variables that are checked against in Figure 2a) are too numerous, and some of them are too close one to another, namely:
    a. 10m wind and wind stress;
    b. Runoff, freshwater flux and latent heat – runoff is a freshwater flux, and latent heat is proportional to one component of EMP, which is essentially the freshwater flux (outside of river deltas).

3) Data-driven methods trained for RMSE reduction are known to be overtly smooth, and that problem is not easy to tackle. I would be curious to know whether the ANN also suffers from it, e.g. by seeing comparisons of instantaneous SSTs with their method vs classical relaxation.

4) Regarding section 3.2 (retrospective simulations), I think that the results are interesting but that maybe comparing them to a classical SSR run (e.g. to COBE, which is available over that period) would have been insightful. We know that the available products (like COBE) have limitations – does your method have benefits compared to directly using that?

5) There are too many commas in many sentences. I suggest the authors go over the manuscript with that in mind and reduce the number of commas.

Minor / specific comments

Title: is it really correcting the air-sea flux? Also please specify NEMO model version.

L. 48 – 49 : give the absolute start year instead of saying "last 15 years".

L. 51: I would add that calibrating reanalyses can be prone to double correcting some effects; i.e., deliberately introducing errors in atmosphere reanalyses that compensate systematic ones in bulk formulae.

L. 69: I think you're using a global configuration, it would be worth mentioning. Is the grid ORCA1? I'm not sure I understand what 1/3º - 1º means...

L. 72: the river discharge from land

L. 74: a TKE scheme (Gaspar et al., 1990 : https://doi.org/10.1029/JC095iC09p16179)

L. 75: I would put the CIGAR reanalysis sentence before the model is described, in the line "We use the CIGAR parameters which is described in depth in Storto and Yang (2024); below we briefly describe some of its specific parameter choices"

L. 83: I think the word “assimilation” is misleading, since I would expect “assimilation” to refer to using an observation operator and directly assimilation measurements… maybe “corrective methodology”?

L. 93 – 95: I do not understand this sentence. What does “whereas” mean here? Where is the contradiction between the first and second parts? It would also be good to explain the readers what “unrolled” mean. I feel like the manuscript would be clearer if it first gave the predictors (the list of six variables); and then say that the predictors have information about the geographical locations. And after that, explicitly write that the SST observations are not included as predictors, which is the point of your study.

L. 103: I would say, “assuming it is nominally valid at the end of each daily window (midnight UTC)”. The average itself is defined over a time interval, but your inference assumes it is valid as an instantaneous snap.

L. 103 – 105: please clearly explain what the REF model does in the presence of sea ice. “the use of the sea-ice mask in the construction of the Qrp field” is vague. If in the presence of sea ice, no correction is applied in REF, then say it explicitly. Does “no sea ice predictors are used” mean that the ANN is simply not used in the presence of sea ice? It would be worth clarifying.

L. 111: we progressively improve the performances…

Table 1: in my opinion, the VIS explanations belong more in the main body text than in the figure caption.

L. 131: The nonlinear character doesn’t make the problem irreducible in itself. It is the **cross-variable** nonlinearities which entangle the processes that makes it complicated. I think “strong multivariate nonlinearities of air-sea interactions (e.g., wind stress depending on both near-surface winds and local temperatures via nonlinear bulk formulas)” would be more accurate.

L. 146: the sensible heat flux related to mesoscale has already been made l. 137.

L. 152: please specify that NEMO is also implemented in Fortran.

Figure 4: I think those are seasonal climatologies, right? If so, please specify.