Review of "A Glass-Box Framework for Interpreting Source-Term–Related Functional Modules in a Global Deep Learning Wave Model"

This manuscript addresses a timely and critical challenge in the field of Artificial Intelligence Oceanography: the lack of scientific trust and physical consistency in data-driven "black-box" models. The authors propose a "Glass-Box" dissection framework and apply it to the OceanCastNet (OCN) v1.0 model. They argue that the model's processor autonomously learns an emergent functional partitioning that mirrors the physical source terms of third-generation numerical wave models.

Overall, shifting the focus from simply improving predictive metrics to mechanistically interpreting the latent space of geoscientific AI models is exactly what the community needs. Translating neural network pathways into the language of physical oceanography is commendable. However, I have substantial concerns regarding the methodological robustness, scientific rigor, and possible circularity of the interpretation. In its current form, it remains ambiguous whether the model has truly learned emergent physical mechanisms analogous to wave-model source terms, or whether the reported modules reflect post-hoc interpretations of statistical patterns learned from the training data. Before the manuscript can support its central claims, the authors need to provide stronger causal, quantitative, and physically independent evidence that these functional groups correspond to meaningful physical processes rather than artifacts of architecture, training data, or the chosen interpretation framework.

Major comments:

1. The proposed glass-box framework appears to rely strongly on the specific architectural design of OCN v1.0, especially the explicit partition of the 768-dimensional latent space into eight feature groups and the separation between intra-group AFNO2D filtering and inter-group Channel MLP mixing. Therefore, it is unclear whether the reported interpretation is a general property of data-driven wave models with similar input-output variables, or whether it is mainly a consequence of this particular AFNO-based architecture.

   More importantly, the manuscript does not demonstrate whether the identified physical roles of specific groups are robust to different random seeds, independent training realizations, checkpoints, model widths, or numbers of feature groups. Because latent representations in neural networks are generally not uniquely identifiable, the physical function assigned to Group may change under retraining, even if the final predictive performance remains similar. Without such robustness tests, the conclusion that OCN has autonomously learned stable source-term-related modules is not sufficiently supported.

The authors should repeat the analysis across multiple independently trained models and report whether the same functional partitioning emerges consistently. If the group identities are not stable, the claims should be reframed as an interpretation of one trained OCN v1.0 instance rather than a general glass-box discovery.

2. The manuscript contains many figures based on group indices, feature indices, weight magnitudes, and qualitative heatmaps. While these visualizations may be useful for internal diagnosis, several of them provide limited physical information to the reader and do not by themselves offer strong evidence for the proposed source-term interpretation. In particular, figures that mainly show "Group xx" or "Feature xx" patterns without clear quantitative metrics, uncertainty estimates, or physically independent validation may not be sufficiently informative for the main text.

   This issue affects the readability and scientific focus of the manuscript. The central claim of the paper is not simply that different feature groups behave differently, but that these groups correspond to physically meaningful source-term-related modules. Therefore, the main figures should prioritize evidence that directly supports this claim, such as quantitative comparisons, causal perturbation tests, cross-seed consistency, regime-dependent responses, or validation against physically defined diagnostics.

   I suggest that the authors substantially streamline the figure presentation. Figures with limited interpretive value, purely diagnostic feature/group visualizations, or redundant qualitative heatmaps could be moved to the Supporting Information. The main text should focus on a smaller number of figures that directly establish the physical meaning, robustness, and causal relevance of the identified modules.

3. Since OCN does not explicitly predict separate wind-sea and swell components, the current interpretation of specific groups as wind-input- or swell/climatology-related modules would be much more convincing if it were evaluated under physically defined sea-state regimes. I suggest that the authors use available reanalysis products, such as ERA5 wind-sea and swell diagnostics, to calculate the swell energy proportion, for example (H_{swell}^2/ H_{total}^2), and separate the analysis into wind-sea-dominated and swell-dominated regions or samples. The authors could then examine how the contributions, activations, or ablation impacts of Group 3, Group 4, and Group 1 change across these regimes. If Group 3 is truly analogous to wind input, its relative contribution should be stronger in wind-sea-dominated conditions; if Group 4 represents propagation, climatology, or swell-related information, its contribution should become more important in swell-dominated regions.

Minor comments:

1. The caption of Figure 10 appears inconsistent with the text. The figure is described as the functional analysis of Group 4, but the caption states that the ablation experiment shows the output when "Group 1" is removed. This should be checked and corrected.

2. In Section 3.3.1, when discussing the Group 4 ablation, the manuscript later states that "Group 1 provides an essential, foundational steering pattern" (line 208). This appears to be a type inconsistency, since the section is about Group 4. The authors should carefully check all group-number references throughout the manuscript.

3. The term "physical energy" used in the Average Physical Contribution Evolution analysis (line 97) may be misleading, because the decoded latent contribution is not necessarily physical wave energy. The authors should use a more neutral term such as "decoded contribution magnitude." And the normalization procedure (line 98) should be described more explicitly.

4. The polar plots for mean wave direction, such as Figure 11i and similar panels, should clearly specify the angular convention. In particular, the authors should indicate what 0° represents, for example geographic north and whether the plotted wave direction denotes the direction waves come from or the direction waves propagate toward.