Understanding AI Ocean Forecasting Models Through Controlled Perturbation Experiments

Julien Seillade, Zakaria Aissa-Abdi, Long Li, VAN GENNIP Simon, GAUDEL Quentin, Anass El Aouni

Mercator Ocean International

Abstract

Recent advances in AI have enabled data-driven ocean forecasting models to achieve skill comparable to, and in some cases exceeding, traditional process-based numerical models. Beyond conventional statistical metrics, several AI systems now reproduce physically meaningful diagnostics such as geostrophic currents, mixed layer depth, and other process-oriented indicators. Yet a fundamental question remains largely unanswered: what do these models actually learn? Are they discovering physically consistent relationships, relying on statistical correlations, or combining both in ways that remain opaque? In this work, we present a first systematic framework for profiling AI ocean forecasting models to better understand the information they exploit and the mechanisms underlying their predictions. Rather than evaluating only forecast accuracy, we probe the models through controlled perturbation experiments and analyse the resulting forecast response. These experiments reveal the sensitivity of the models to different physical state variables and provide insight into the relative importance of the information encoded in each field. Our study represents an initial step toward demystifying AI ocean forecasting systems. By complementing traditional verification with model profiling, we aim to deepen our understanding of the physical knowledge captured by these models, identify potential shortcuts or spurious dependencies, and establish a foundation for more interpretable and trustworthy AI systems for operational ocean forecasting.