https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

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

Ziliang Zhang[1,2], Huaming Yu[1,2,3], Xiaotian Dong[4], Jiaqi Dou[1,2], Danqin Ren[5], and Xin Qi[6]

[1]College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao 266100, Shandong,China
[2]State Key Laboratory of Physical Oceanography, Ocean University of China, Qingdao 266100, Shandong, China
[3]Sanya Ocean Institute, Ocean University of China, Sanya 572024, Hainan, China
[4]School of Business, Qingdao University, Qingdao 266071, Shandong, China
[5]Dawning Information Industry Co., Ltd, Qingdao 266000, Shandong, China
[6]College of Management, Ocean University of China, Qingdao 266100, Shandong, China

Correspondence: Huaming Yu (hmyu@ouc.edu.cn) and Xin Qi (x.qi@ouc.edu.cn)

Abstract.

Data-driven deep learning (DL) models are increasingly powerful tools for Earth system prediction, but their "black box" nature and a perceived lack of physical consistency hinder scientific trust. Validating the physical realism of these models requires new methodologies that can look inside the "black box" and map internal computations to physical processes.

This paper proposes and demonstrates such a "glass box" dissection framework. We apply this framework—which combines architectural analysis and systematic functional ablation experiments—as a case study to the OceanCastNet (OCN) v1.0 model.

The dissection demonstrates that the v1.0 model's processor autonomously learns an emergent functional partitioning. We statistically identify and validate distinct computational modules analogous to the source terms in third-generation (3G) physical wave models: a foundational propagation and climatology module (Group 4), a non-linear wind-input operator (analogous to S_{in}, Group 3), and a state-dependent balancing operator for dissipation (analogous to S_{ds}, Group 1). Furthermore, the analysis reveals that other higher-order physics are managed by a complex, coupled system of operators.

This methodological dissection provides tangible evidence of emergent physical realism in a DL model. It offers a reproducible blueprint for validating the physical fidelity of future AI-based Earth system models, providing a concrete pathway toward developing and trusting physically-constrained "grey-box" systems.

1 Introduction

Data-driven deep learning (DL) models represent a significant advance in Earth system prediction, with recent examples (Bi et al., 2023; Lam et al., 2023; Pathak et al., 2022) achieving performance comparable to established numerical weather prediction (NWP) systems. This trend extends to ocean wave forecasting; our recent work introduced OceanCastNet (OCN) (Zhang et al., 2025a), a global DL model demonstrating forecast skill competitive with the operational ECWAM system.

Despite this success, a persistent challenge regarding the credibility and adoption of these models remains (Adadi and Berrada, 2018). Skepticism within the physical modeling community often stems from their 'black box' nature, hindering

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

scientific trust (McGovern et al., 2019; Carvalho et al., 2019). A core concern is whether models trained on reanalysis data are learning generalizable physical processes or relying on non-physical statistical correlations (Clare et al., 2022). This concern is validated by documented failures where models have shown physically inconsistent behavior, particularly when extrapolating to events outside their training distribution (Sun et al., 2025).

To bridge this gap, a new focus on trust, interpretability, and physical consistency is required. A DL model's predictions should be "right for the right reasons." In wave modeling, the "right reasons" are defined by third-generation (3G) spectral models (The Wamdi Group, 1988; Tolman, 1992). These models are built upon the action balance equation (Komen et al., 1994), which balances physical source terms (S_{tot}). This balance includes S_{in} (wind input) (Janssen, 1989; Cavaleri and Rizzoli, 1981; Rogers et al., 2012), S_{ds} (dissipation via white capping and swell attenuation) (Tolman, 1992; Ardhuin et al., 2010; Rogers et al., 2012), S_{nl} (non-linear interactions) (Tolman, 1992), and propagation (Janssen, 1989; Cavaleri and Rizzoli, 1981; Tolman et al., 2009). Any realistic wave model must implicitly or explicitly account for this fundamental energy balance.

While Physics-Informed Neural Networks (PINNs) explicitly enforce these laws (Ehlers et al., 2025), this study investigates a different question: can a purely data-driven model, trained only on observational data, autonomously learn and replicate these physical mechanisms as the most efficient solution pathway?

Verifying this hypothesis, however, requires moving beyond standard performance metrics (e.g., RMSE, correlation), which treat the model as an opaque system. New methodologies are needed to assess the physical realism of a model's internal computations.

This paper proposes and demonstrates a methodological framework—termed a "glass box" dissection—to systematically investigate the internal reasoning of DL-based geoscientific models. The framework combines architectural analysis with a series of functional ablation experiments designed to isolate, identify, and validate the computational pathways within the model (detailed in Sec. 2.2).

We apply this framework as a case study to the OCN model. As a high-performance system, OCN serves as an effective testbed to determine if, and how, emergent physical realism is linked to predictive skill. Applying this dissection framework reveals that the OCN processor has autonomously learned an emergent functional partitioning. We demonstrate that distinct computational modules have been formed that are analogous to the source terms in 3G models. The analysis identifies a foundational propagation and climatology module (Group 4) that provides the model's geographic baseline and swell field. This is complemented by a dynamic, wind-dependent wind-input operator (Group 3), analogous to S_{in}, and a state-dependent balancing operator (Group 1), analogous to S_{ds}, which appears crucial for maintaining the model's numerical stability and energy balance.

By translating the internal mechanisms of a DL model into the language of physical oceanography, this work provides a tangible methodological bridge between the data-driven and physics-based communities. This dissection approach serves as a reproducible blueprint for validating the physical fidelity of future AI Earth system models, offering a pathway to move from opaque 'black boxes' to more trusted, physically-grounded 'grey-box' systems.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

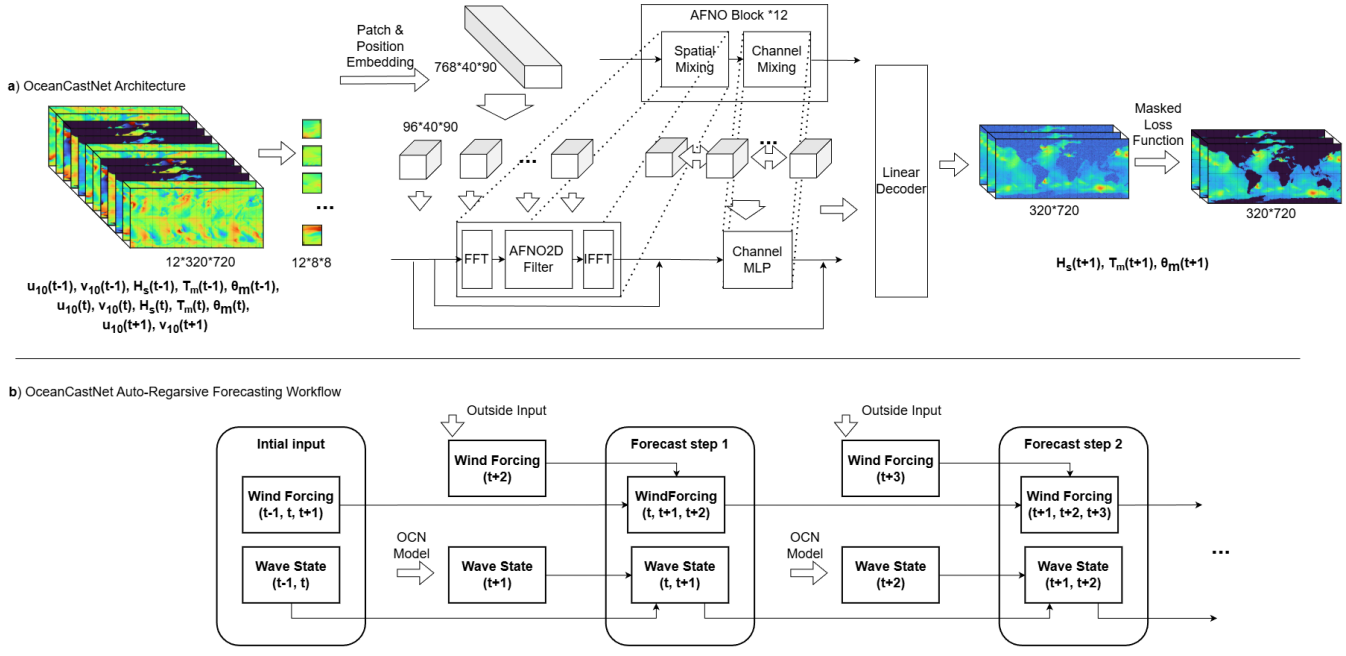

Figure 1. OceanCastNet (OCN) architecture and auto-regressive workflow. (a) The single-step model architecture. Input variables are passed through a patch and position embedding layer into the 12-block processor. Each block contains two distinct sub-modules: a spatial-mixing path (using the AFNO2D Filter for intra-group processing) and a channel-mixing path (using the Channel MLP for inter-group mixing). A linear decoder projects the latent features back to the physical output. (b) The auto-regressive workflow. The model uses the wave state from the previous two steps (t-1, t) and wind forcing from three steps (t-1, t, t+1) to predict the wave state at the next 6-hour step (t+1).

2 Methodology: Framework Development and Case Study

This section details the "glass box" dissection framework. We first describe the architecture of the OceanCastNet (OCN) model (Sec 2.1), which serves as the case study and testbed for the framework's development. We then detail the components of the analytical framework itself, as they are applied to the OCN architecture (Sec 2.2).

2.1 The OceanCastNet (OCN) Architecture: A Case Study Testbed

To dissect the internal mechanisms of OceanCastNet (OCN), we must first briefly review its core architecture. OCN (Zhang et al., 2025a) is a data-driven global wave forecast model developed from the FourCastNet and Adaptive Fourier Neural Operator (AFNO) frameworks. The model operates auto-regressively, predicting the subsequent 6-hour wave state (significant wave height, mean wave period, and mean wave direction) by integrating wave fields from the previous two time steps (t-1,t) and wind forcing data from three time steps (t-1,t,t+1).

As illustrated in Figure 1, a single forecast step is executed through three principal stages: an Encoder, a Processor, and a Decoder.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

The Encoder, or Embedding layer, first receives the multi-step, 12-channel input tensor (shape 12 ×320 ×720). It partitions this input into non-overlapping 8x8 patches and, via a linear projection, embeds the information from each patch (12 ×8 ×8) into a 768-dimensional feature vector, which defines the model's latent space. This process transforms the 12 discrete physical variables into a high-dimensional, mixed-feature representation.

This latent space representation is then summed with a learnable Positional Embedding vector, which supplies the model with static geospatial information. The resulting vectors are fed into the core Processor, which consists of 12 sequential AFNO Blocks (Depth=12). A critical aspect of this Processor, and the primary focus of our "glass box" analysis, is its functional organization. The 768-dimensional feature space is functionally divided into 8 distinct Feature Groups, each comprising 96 features. As shown in Figure 1, each AFNO Block is composed of two sub-modules that treat these groups differently. The first sub-module, AFNO2D Filter, performs global spatial mixing via FFT and utilizes a block-diagonal sparse MLP for intra-group processing, effectively strengthening the specialized features within each of the 8 groups. The second sub-module is a standard MLP that operates across the full 768-feature dimension. Its primary function is to facilitate inter-group interaction, allowing the 8 feature groups to globally exchange and mix information.

Finally, after passing through all 12 Blocks, the resultant 768-dimensional vector is passed to the Decoder, a simple linear Head MLP. This layer projects the latent features back into the physical domain, decoding them into the 3-channel output (significant wave height, mean period, and mean direction) at the original grid resolution.

Our central hypothesis is that the OCN processor has autonomously organized its computational workflow, assigning distinct physical processes (e.g., wind input, dissipation, propagation) to specific computational pathways within this 8-group structure. The following sections will validate this "emergent physical partitioning" by systematically analyzing and ablating these key architectural components.

2.2 An Analytical Framework for Internal Mechanism Dissection

To dissect OCN's internal reasoning, we employed an analytical framework to map latent-space computations to physical processes, using four complementary techniques.

First, we analyze the Channel MLP Group Interaction by calculating an "effective linear weight" matrix ($W_{eff} = W_2 \times W_1$). This approach, common in "weight-based circuit analysis" (Dunefsky et al., 2024), approximates the MLP as a linear layer, ignoring the GELU non-linearity. This is justified by research (Pearce et al., 2024) suggesting that the core function of such modules is heavily encoded in their linear transformation paths.

Second, we analyze the AFNO2D Filter Weight Strength . As OCN is based on the Fourier Neural Operator (FNO) (Guibas et al., 2021; Li et al., 2020), which operates in the frequency domain, we compute the magnitude of its complex-valued weights. This directly measures the "energy" or "gain" applied to each feature group.

Third, we implement an Average Physical Contribution Evolution analysis. This method isolates the activation of each group g at each block b and propagates it through the decoder to compute its "physical energy". Normalizing these energies yields an 8 ×12 evolution matrix, revealing the dynamic computational handoff between groups.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

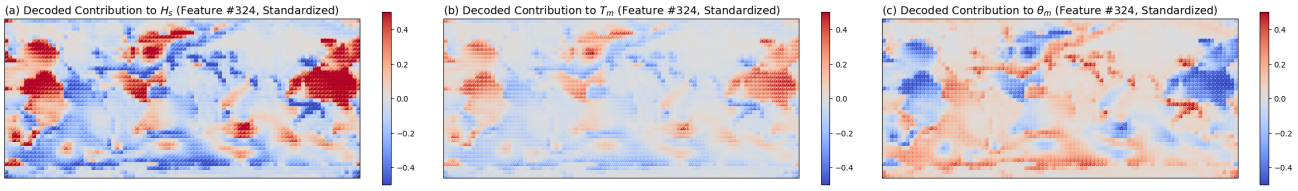

Figure 2. Physical-space projection of a single latent feature. The standardized, decoded contribution of a single, representative latent feature (feature #324) to the three physical output variables: (a) H_s, (b) T_m, and (c) θ_m. The activations of this feature were isolated from the final processor block and passed through the linear decoder head.

Fourth, we conduct an Average Physical Ablation, a classic "Ablation Study" (Meyes et al., 2019) to verify functional roles. We use hooks to "zero-out" a target group across all 12 layers—a standard technique (Li and Janson, 2024)—and observe the impact on the final averaged physical output.

For several of these analyses, we leverage the model's final linear Decoder Head as a consistent projection from the latent space back to the physical domain. This linearity makes it a valid tool for visualizing the physical information encoded in latent vectors, such as the Position Embedding and the results of our ablation experiments.

3 Results: Identifying Emergent Physical Analogs in OCN v1.0

Applying the analytical framework defined in Section 2.2, this section details the results of the "glass box" dissection of the OCN model. The findings provide quantitative evidence that the model, despite being purely data-driven, has learned to approximate fundamental physical processes by partitioning its computational workflow.

3.1 Emergent Physical Intuition: What the Model Attends To?

The OCN model processes information in a 768-dimensional latent space. A naive hypothesis would be that individual features within this space (1 through 768) might learn to represent specific, isolated physical concepts. We tested this hypothesis by attempting to visualize the physical-space meaning of a single feature. This was done by isolating one feature (e.g., feature #324) from a model's output activation, passing it through the linear decoder, and plotting the resulting field.

As shown in Figure 2, the decoded image from a single feature is spatially sparse and physically uninterpretable. The field consists largely of low-information, near-zero values. This result, combined with the lack of any discernible pattern in the 768 individual decoder weights (Figure 3), strongly suggests that individual features are not the fundamental unit of physical meaning in this model.

The model's architecture itself suggests the appropriate analytical unit. The AFNO2D Filter processor, the core of each block, explicitly partitions the 768 features into 8 parallel Feature Groups (of 96 features each) and performs its primary computations within these groups. This architectural design implies that the group, not the individual feature, is the likely

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

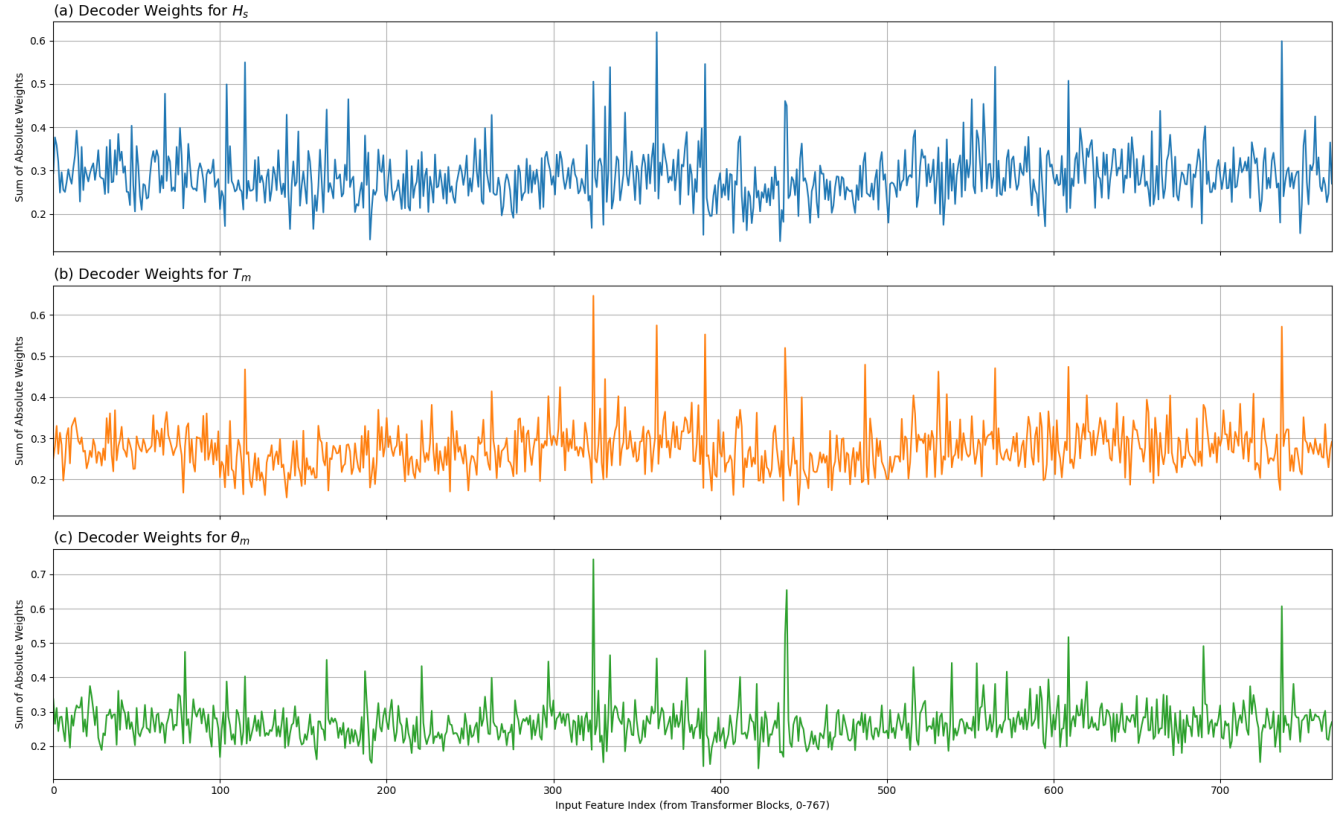

Figure 3. Decoder weight magnitude for individual latent features. The sum of absolute decoder weights connecting each of the 768 latent features (x-axis) to the three physical output variables: (a) H_s, (b) T_m, and (c) θ_m.

functional unit of physical representation. Therefore, all subsequent analysis in this paper is conducted at the level of these 8 Feature Groups.

As a first test of this group-based framework, we re-analyze the initial Input Embedding. Instead of averaging all 768 features, we aggregate the embedding weights by group. Figure 4 visualizes the total connection strength from each of the 12 Input Channels (x-axis) to each of the 8 Feature Groups (y-axis).

The result is (Figure 4) is notable. A clear, physically meaningful hierarchy emerges, and it is universally shared by all 8 groups. The entire embedding is primarily dominated by Input Channel 7 (current Significant Wave Height), as indicated by the bright vertical band. This is followed by strong, model-wide preferences for Channel 8 (current Mean Wave Period), Channel 10 (future Zonal Wind), and Channel 9 (current Mean Wave Direction).

This analysis confirms the model's emergent physical intuition, but adds a crucial insight: the initial "focus" on the current wave state and future wind is a fundamental, model-wide bias. All feature groups start with this same physical-based preference.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

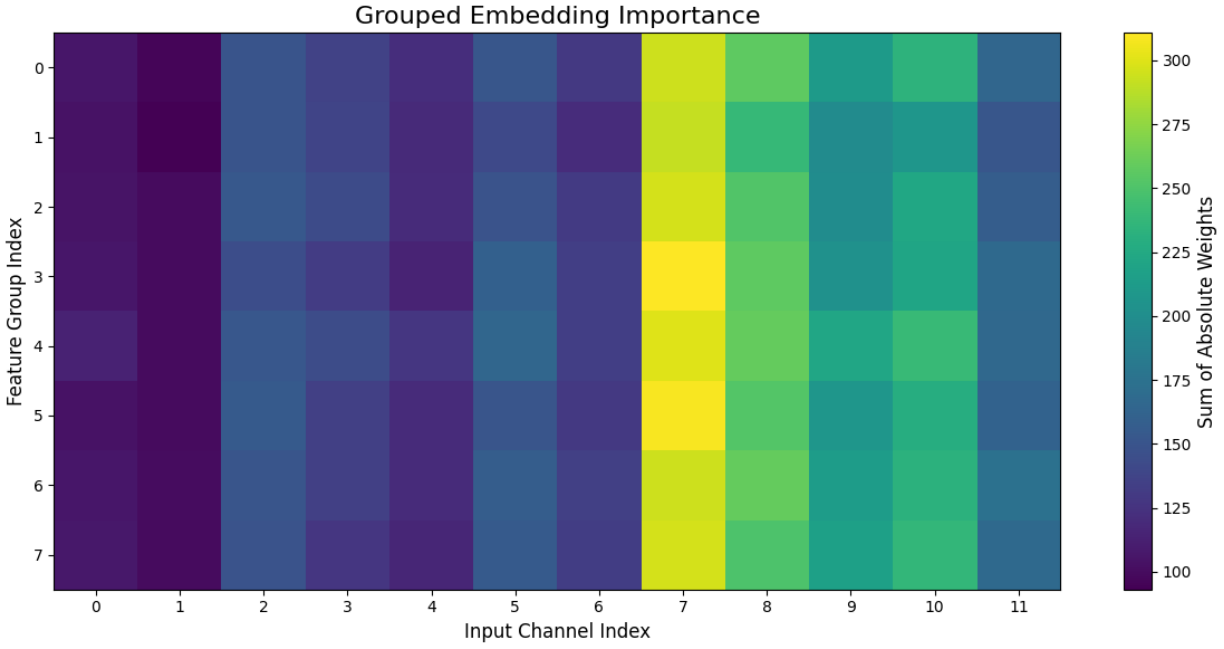

Figure 4. Grouped analysis of input embedding importance. Heatmap showing the sum of absolute weights from the initial embedding layer, connecting each of the 12 input channels (x-axis) to each of the 8 latent feature groups (y-axis). The plot reveals a strong, uniform preference across all groups for Input Channel 7 (H_s at time t).

The question for our subsequent analysis is, therefore, how these groups specialize from this common starting point as they are processed by the model's 12 layers.

3.2 Emergent Physical Parameterization: The Learned Positional Embedding

Before the processor's 12-block evolution begins, the model adds a single, learnable Position Embedding (PE) vector to the input features. This component's sole purpose is to inject static, grid-dependent information. To understand what static knowledge the model learned, we isolate this PE vector and visualize its content by passing it through the linear decoder head, as established in our methodology.

The results, shown in Figure 5, are notable. The decoded PE reveals that the model has learned a detailed and physically relevant map of global geography.

A clear land-sea boundary is visible across all three output channels. This is expected, as the model's masked loss function explicitly ignores land regions . However, the PE has learned finer-scale detail than the 0.5° land-sea mask it was trained with. In the "Pattern for Output Channel 0" (Significant Wave Height), fine-scale geographic features are clearly resolved. These include the Malé island chain, islands east of Madagascar, and even small archipelagos in the Southern Ocean near Antarctica. These features are often smaller than the 0.5° grid resolution of the training mask, yet the model has learned to represent them.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

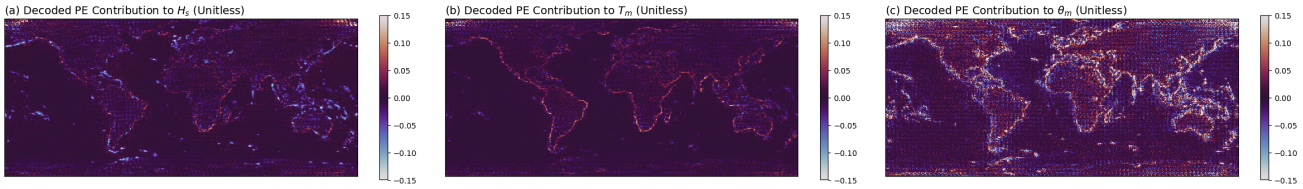

Figure 5. Physical-space projection of the learned Positional Embedding. The static, learnable Positional Embedding (PE) vector is isolated and passed through the linear decoder head to visualize its encoded geographic information. The maps show the PE's unitless contribution to (a) H_s, (b) T_m, and (c) θ_m.

This is not simple image memorization; it is an emergent physical parameterization. The model learned these sub-grid-scale features because these islands have a real-world physical impact in the ERA5 training data—they create wave-shadowing effects that consistently lower the significant wave height in their respective grid cells. The PE has captured this static, physical property of the sub-grid geography.

Interestingly, the decoded PE reveals a sophisticated, variable-dependent parameterization of global geography. The "Pattern for Output Channel 0" (Significant Wave Height), contrary to the other channels, appears to have learned particularly fine-grained, open-ocean details. While the continental outlines are sparse, this channel resolves small-scale geographic features, such as the archipelagos in the Indian, Pacific, and Southern Oceans. This suggests the model has learned an emergent parameterization for the physical wave-shadowing effect of these sub-grid-scale islands, which has a direct, localized impact on H_s in the training data. Conversely, the "Pattern for Output Channel 1" (Mean Wave Period) appears more diffuse in the open ocean, lacking these island details. Its information is largely concentrated on sharp, well-defined continental boundaries, suggesting the model learned that the primary static influence on T_m is the coastline itself, which acts as a critical boundary for wave reflection and shoaling. The "Pattern for Output Channel 2" (Mean Wave Direction) remains a hybrid, capturing both the sharp coastal boundaries of Channel 1 (critical for refraction) and other static, open-ocean features, likely related to climatological wind and swell steering patterns.

3.3 Results: Emergent Physical Realism in OCN

The core results of the dissection stem from a multi-stage analysis of the 8 feature groups within the OCN v1.0 weights, as outlined in our methodological framework (Sec 2.2). The analysis demonstrates that the model's processor has partitioned its workflow into distinct, specialized pathways that function as analogs to the physical source terms governing traditional 3G wave models . We follow a "hypothesis-and-verify" approach, using the static analyses (Sec 2.2) to form a hypothesis about each group's function, which we then test using the functional ablation experiments.

To establish a clear baseline for all functional tests, we first computed the model's average physical output over 146 distinct samples covering the full 2020 year. This "Control Run" (Figure 6), which shows the model's mean-state prediction with no

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

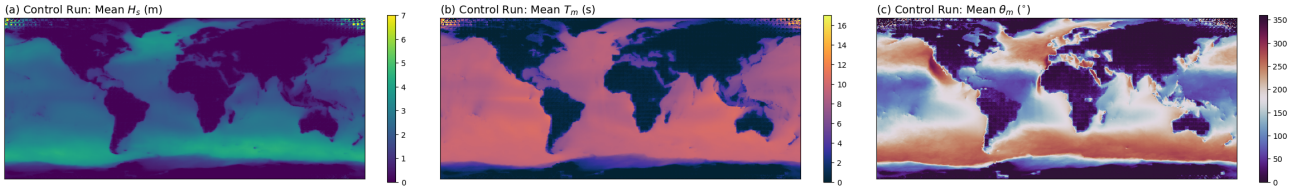

Figure 6. Mean physical state of the Control Run. The ensemble-averaged physical output of the OCN model with all components active (no groups removed), calculated over 146 distinct samples. The maps show: (a) Mean Significant Wave Height (H_s) in meters, (b) Mean Wave Period (T_m) in seconds, and (c) Mean Wave Direction (θ_m) in degrees.

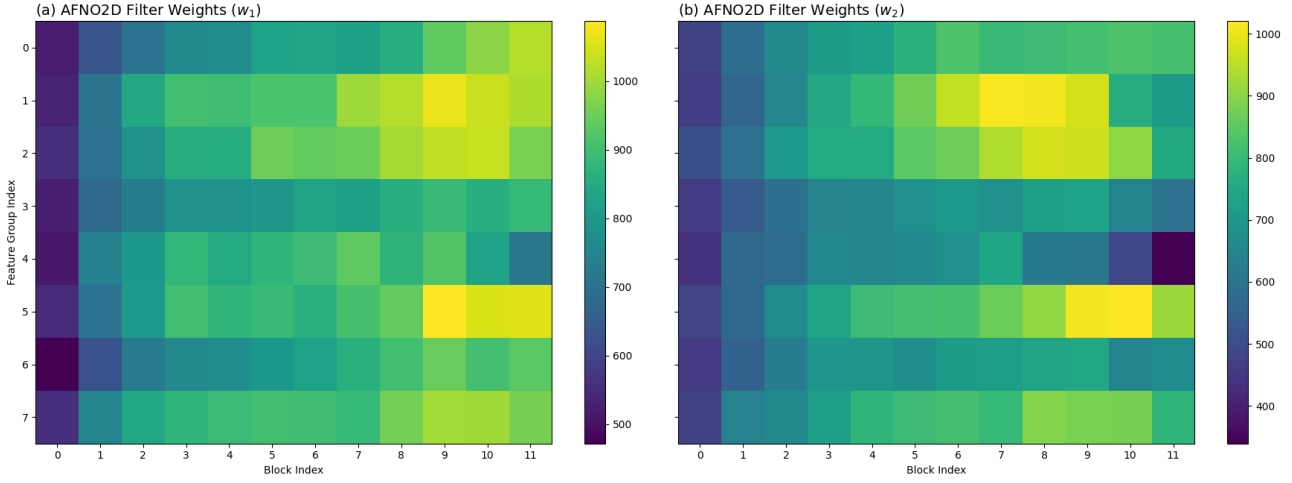

Figure 7. Evolution of AFNO2D filter weight strength. This heatmap shows the sum of absolute weights for the AFNO2D filter, which performs frequency-domain, intra-group processing. The strength is shown for each feature group (y-axis) across all 12 processor blocks (x-axis). The two panels correspond to the learnable parameters within the AFNO2D class: (a) the first-layer weights (w_1) and (b) the second-layer weights (w_2).

groups removed, serves as the ground truth for all subsequent ablation experiments. It displays the expected climatological patterns of global wave heights, periods, and directions.

3.3.1 Group 4: The Emergent Propagation and Climatology Module

Our static analysis, performed on the OCN v1.0 weights, first identified Group 4 (Index 4) as a foundational component based on three distinct computational signatures.

First, the Evolution of AFNO2D Filter Weight Strength heatmap (Figure 7) reveals the intra-group processing intensity. Group 4 (row 4) is notable, exhibiting a distinct signature of low processing intensity (darker colors) that becomes most

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

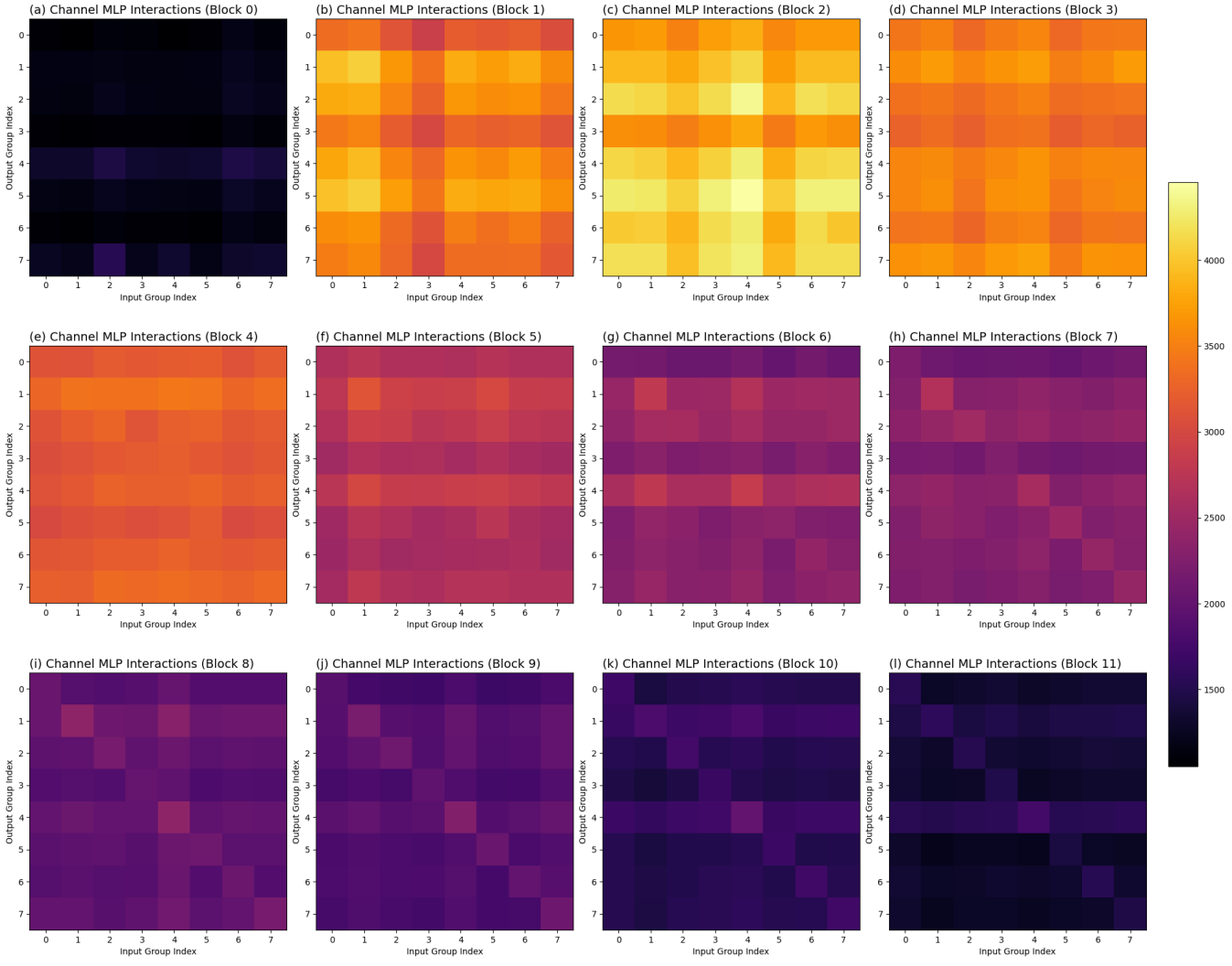

Figure 8. Evolution of inter-group interactions within the Channel MLP. Each panel shows the effective weight strength (calculated as $W_{eff} = W_2 \times W_1$) of the Channel MLP, which governs the inter-group mixing. The heatmaps illustrate the interaction strength from input feature groups (x-axis) to output feature groups (y-axis) for each processor block. Panels are indexed by block: (a) Block 0, (b) Block 1, and so on, up to (l) Block 11.

pronounced in the final processor blocks (Block Order 10-11). TThis suggests a high degree of "inertia," where information is preserved rather than heavily re-computed, a characteristic of a foundational module.

Second, the Channel MLP Group-to-Group Interactions (Figure 8) reveal the inter-group connectivity. While the v1.0 weights show complex interactions across all blocks, Group 4 (row 4) displays a strong, persistent "horizontal bar" pattern

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

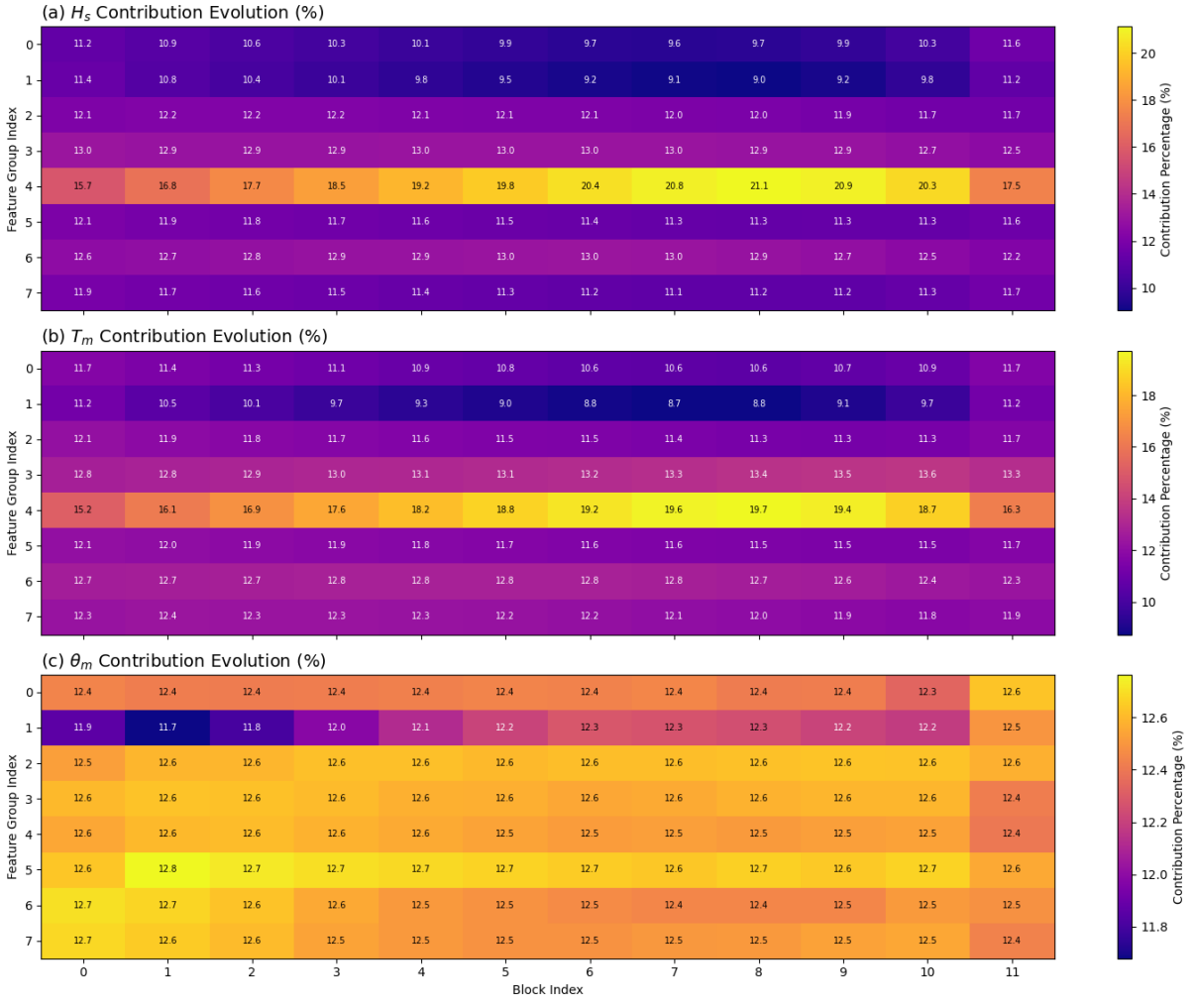

Figure 9. Evolution of relative physical contribution by feature group. The contribution percentage (colorbar) for each feature group (y-axis) is plotted as it evolves across the 12 processor blocks (x-axis). This analysis, detailed in Section 2.2, is shown separately for the three physical output variables: (a) H_s, (b) T_m, and (c) θ_m. The contribution from θ_m (c) is calculated cyclically.

(especially in Blocks 10 and 11), indicating that it receives significant input from (or "listens to") all other feature groups. This "global awareness" is a key characteristic of a foundational module.

Third, the critical importance of this group is confirmed by the Average Physical Contribution Evolution (Figure 9). Here, Group 4's contribution (row 4) is not only high but dominant. For H_s (Channel 0), its contribution starts at 15.7% and grows to 17.5%, and for T_m (Channel 1), it starts at 15.2% and grows to 16.3%, making it the single most important contributor to both wave height and period.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

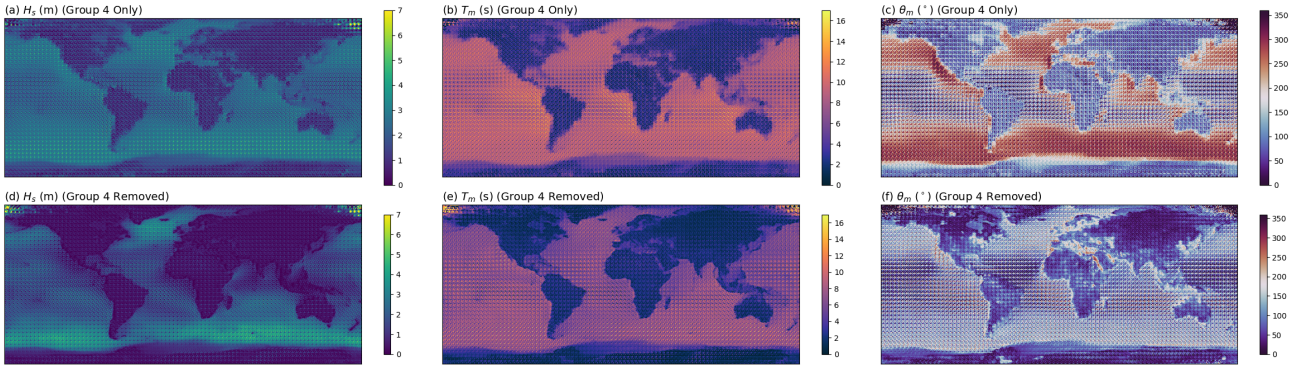

Figure 10. Functional analysis of Group 4 (Climatology) by isolation and ablation. Ensemble-averaged physical outputs (over 146 samples) from two complementary experiments. (a-c) The "isolation" experiment, showing the physical-space output when only the contribution from Group 4 is kept. (d-f) The "ablation" experiment, showing the output when the contribution from Group 1 is removed (zeroed-out). Columns correspond to H_s (m), T_m (s), and θ_m (°).

These three static observations paint a clear and consistent picture: Group 4 functions as a foundational module that is internally inert (low AFNO2D Filter weights), globally connected (high Channel MLP interaction), and critically important (highest contribution %) to the final result. This strongly supports the hypothesis that Group 4 represents the model's learned climatology and geographic baseline.

We tested this hypothesis using our average ablation analysis (Figure 10). The results are conclusive. The "isolation" panel (a, b, c), showing the output from only Group 4, is not a full wave field but rather a static, climatological map. It sharply defines the land-sea boundaries and reproduces the well-known global swell patterns. Conversely, the "ablation" panel (d, e, f), showing the model without Group 4, demonstrates a systemic model collapse. The model output collapses, losing all defined swell features and reverting to a spatially uniform, low-energy state.

The 10-day (40-timestep) forecast ablation provides strong quantitative evidence (Figure 11). This experiment confirms the failure is systemic and cumulative. The "Masked" spatial maps (Group 4 removed) fail to produce any meaningful wave patterns. The statistical distributions on the right quantify this breakdown: the Significant Wave Height histogram (orange) narrows significantly, forming a single, sharp peak near 1m, with the entire tail of high-energy waves disappearing. Similarly, the Mean Wave Period histogram (orange) shows a pronounced shift to lower periods, forming a new peak around 4-6s, while the climatological swell (periods > 8s) is largely eliminated .

The Mean Wave Direction analysis reveals a particularly complex finding. While the contribution heatmap (Figure 9) suggests a uniform contribution from all groups, the ablation forecast in Figure 11 shows a systemic, non-random failure. The removal of Group 4 does not cause the directional distribution to revert to uniform noise; instead, it causes a pronounced systemic bias. The rose plot (orange) shows a non-physical accumulation of energy in the 0-90 degree quadrant, while simultaneously showing a significant loss of energy in the 180-360 degree range, including a near-absence of energy between 370

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

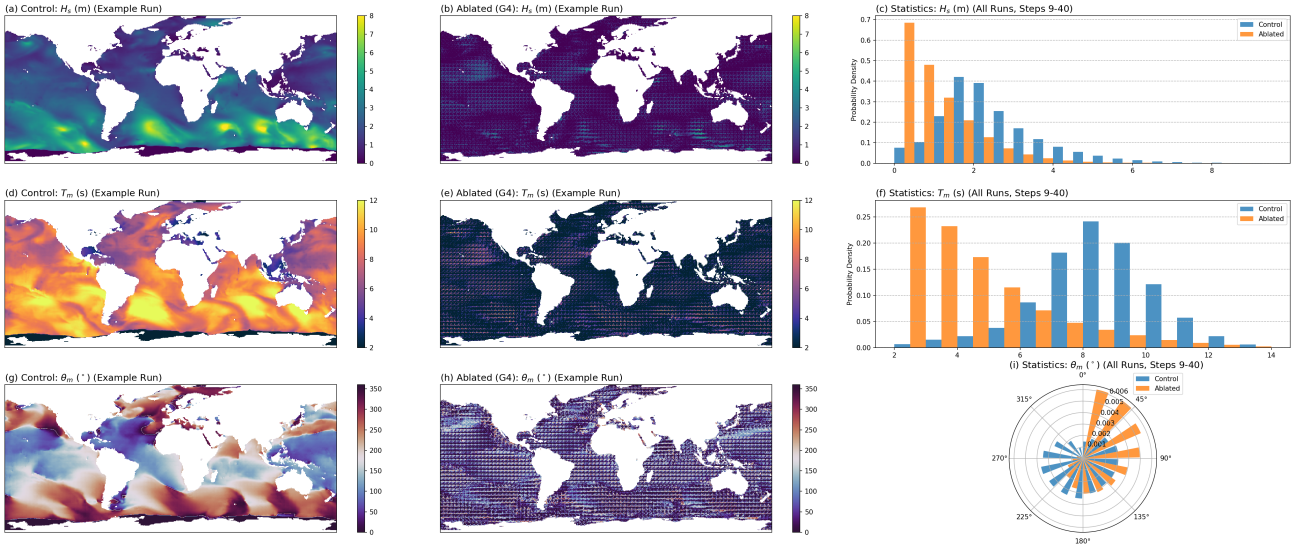

Figure 11. 10-day forecast ablation experiment for Group 4 (Climatology). The figure compares a Control run (all groups) and an Ablated run (Group 4 removed). The grid is organized by physical variable (rows) and analysis type (columns). Left Column (a, d, g): A snapshot of the final forecast state (Day 10, Step 40) from a single, representative run (Run #14). Middle Column (b, e, h): The corresponding final-state snapshot from the Ablated run (Group 4 removed). Right Column (c, f, i): Aggregate statistical distributions compiled from all 32 forecast runs, using data from forecast steps 9 through 40. Distributions show the Control (blue) vs. Ablated (orange) runs. Rows correspond to H_s (m), T_m (s), and θ_m (°).

and 0 degrees. This complex failure mode, which is not yet fully understood physically, underscores the intricate nature of the wave direction computation and confirms that Group 1 provides an essential, foundational steering pattern, without which the entire directional system degrades into a biased, erroneous state.

3.3.2 Group 3: The Emergent Wind-Input Operator (Analog to S_{in})

In contrast to the foundational nature of Group 4, our analysis identifies Group 3 (Index 3) as the model's primary dynamic operator, analogous to the wind-input source term (S_{in}) in 3G models.

The static analysis for Group 3 presents an intriguing puzzle. The Evolution of AFNO2D Filter Weight Strength heatmap (Figure 7) shows that Group 3 (row 3) is a computationally active group. Its processing intensity shows a slight increasing trend in the deeper layers (peaking around block 9-11), though this signature is less pronounced than the foundational Group 4. This indicates it is a highly dynamic, non-inert operator. However, the Average Physical Contribution heatmap (Figure 9) shows that Group 3's direct contribution to the final H_s output (Channel 0, row 3) is stable and modest, hovering around 12.9-13.0%. This apparent contradiction—dynamic internal computation but modest direct contribution—makes its function difficult to interpret

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

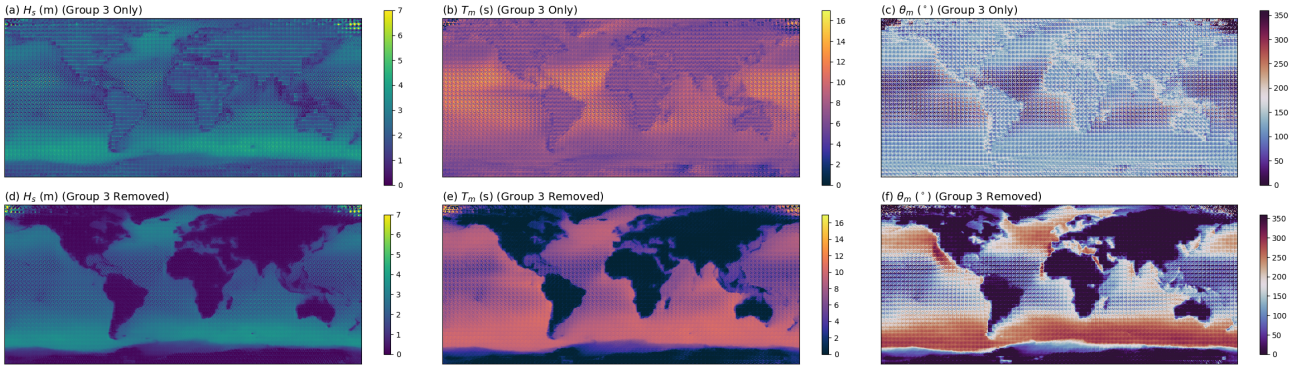

Figure 12. Functional analysis of Group 3 (Wind-Input) by isolation and ablation. Ensemble-averaged physical outputs (over 146 samples) from two complementary experiments. (a-c) The "isolation" experiment, showing the physical-space output when only the contribution from Group 3 is kept. (d-f) The "ablation" experiment, showing the output when the contribution from Group 3 is removed (zeroed-out). Columns correspond to H_s (m), T_m (s), and θ_m (°).

from static analysis alone. This led to the hypothesis that Group 3's function is not to be the final output, but to perform a critical calculation (wind-sea generation) whose results are then integrated with other groups.

Our functional analyses confirm this hypothesis unequivocally. The Average Ablation analysis (Figure 12) provides the first clear evidence. The "isolation" panel (a, b, c), showing the output of only Group 3, is a map of pure wind-sea. The signal is strongest almost exclusively in the major storm tracks—the Southern Ocean, the North Pacific, and the North Atlantic—while the swell-dominated regions (governed by Group 4) are nearly empty. Conversely, the "ablation" panel (d) shows the model without Group 3. This field retains the global swell patterns from Group 4 but is conspicuously "calm" in the storm tracks, where high-energy wind-sea has been erased.

The 10-day (40-timestep) forecast ablation provides definitive, quantitative proof of this function (Figure 13).When Group 3 is removed from the 10-day run, the model's ability to generate any significant wave events is eliminated. The "Masked" H_s histogram (orange, panel c) demonstrates this: the entire high-energy tail of the distribution (waves > 4m) is completely cut off, leaving only the low-energy swell baseline. Critically, the removal of this wind-sea component also causes a clear shift in the T_m histogram (panel f). The ablated distribution (orange) shifts toward lower periods, with its peak clustering around 6-8s, compared to the control run's peak at 8-10s. This indicates that the wind-sea generated by Group 3 (in this v1.0 model) contributes significantly to the mean wave period, and its removal causes the period distribution to revert to a lower-period baseline.

Finally, to substantiate the causal link and quantify the functional response of Group 3 to wind forcing, we conducted a quantitative experiment (Figure 14). Instead of merely removing the wind, we ran 145 simulations where the model was forced with a range of uniform U-Wind speeds, from 0 m/s (calm) to 30 m/s (extreme storm). The results are illustrative of the model's emergent physical realism.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

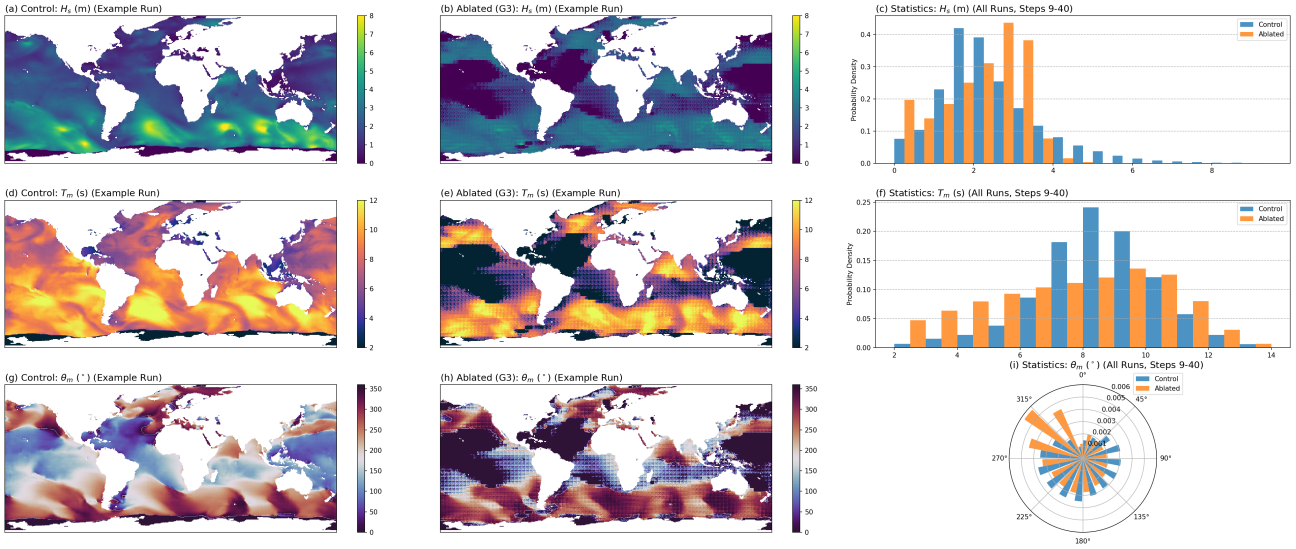

Figure 13. 10-day forecast ablation experiment for Group 3 (Wind-Input). The figure compares a Control run (all groups) and an Ablated run (Group 3 removed). The grid is organized by physical variable (rows) and analysis type (columns). Left Column (a, d, g): A snapshot of the final forecast state (Day 10, Step 40) from a single, representative run (Run #14). Middle Column (b, e, h): The corresponding final-state snapshot from the Ablated run (Group 3 removed). Right Column (c, f, i): Aggregate statistical distributions compiled from all 32 forecast runs, using data from forecast steps 9 through 40. Distributions show the Control (blue) vs. Ablated (orange) runs. Rows correspond to H_s (m), T_m (s), and θ_m (°).

First, the model's physical output (blue line, left y-axis) demonstrates a correct physical response, with the average H_s increasing with wind speed. This response curve closely follows the theoretical Pierson-Moskowitz (PM) spectrum for a fully developed sea (black dotted line), suggesting the model has learned a physically realistic, fetch-unlimited growth curve.

Second, and more critically, the analysis of the internal mechanism (red line, right y-axis) reveals how the model achieves this. The contribution of Group 3 to the H_s channel is not linear; it functions as a non-linear "throttle." At low wind speeds (0-7.5 m/s), Group 3 is dormant (with a near-zero or even negative contribution). As the wind speed increases past this physical threshold, the model activates Group 3, and its contribution increases sharply and non-linearly, peaking at over 35% and driving the corresponding physical growth in H_s. This experiment provides strong quantitative evidence that Group 3 is the model's learned, quantitative, and non-linear analog for the wind-input source term, S_{in}.

3.3.3 Group 1: The Emergent Dissipation Operator (Analog to S_{ds})

Our analysis identifies Group 1 (Index 1) as a critical balancing operator within the OCN v1.0 model, performing functions analogous to the dissipative source terms (S_{ds}) in 3G wave models.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

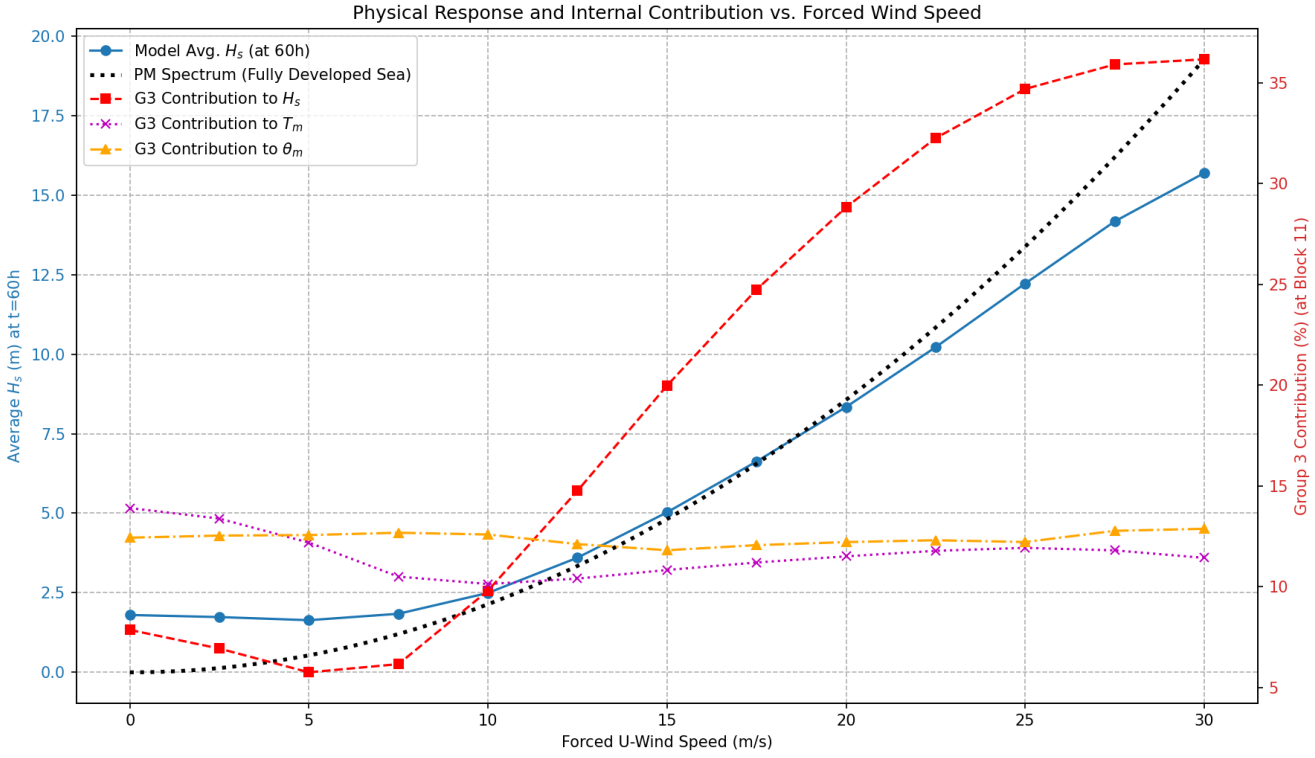

Figure 14. Model response to uniform, constant wind forcing. The plot shows ensemble-averaged results from 145 simulations where the model was forced with a range of uniform U-Wind speeds (x-axis), from 0 to 30 m/s. The final, globally-averaged H_s (m) at t=60h (10 steps) is shown by the blue line (left y-axis), compared against the theoretical Pierson-Moskowitz spectrum (black dotted line). The colored lines (right y-axis) show the internal contribution percentage of Group 3 (at the final processor block) to the three output variables: H_s (red), T_m (magenta), and θ_m (orange).

The static analysis for Group 1 presents a different signature from the foundational Group 4. In the AFNO2D Filter heatmap (Figure 7) , Group 1 (row 1) is computationally active, showing bright (high-weight) patterns in the deeper layers. In the Channel MLP interactions (Figure 8), it shows strong "horizontal bar" interactions (especially in Blocks 10 and 11), indicating it "listens" to all other groups. However, its direct contribution to the final output (Figure 9, row 1) is modest and stable (around 11.2% for H_s and 11.0% for T_m). This combination—high internal activity but modest direct contribution—suggests a role in modulating or balancing the system.

Functional analyses confirm this role. The Average Ablation analysis (Figure 15) clearly reveals Group 1's opposing function to energy input. The "ablation" panels (d-f), showing the model without Group 1, demonstrate a critical failure in the model's energy balance. The H_s field (panel d) is slightly higher, but the T_m field (panel e) undergoes a pronounced, non-physical shift, with mean periods across the global ocean increasing significantly into a "hotter" state.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

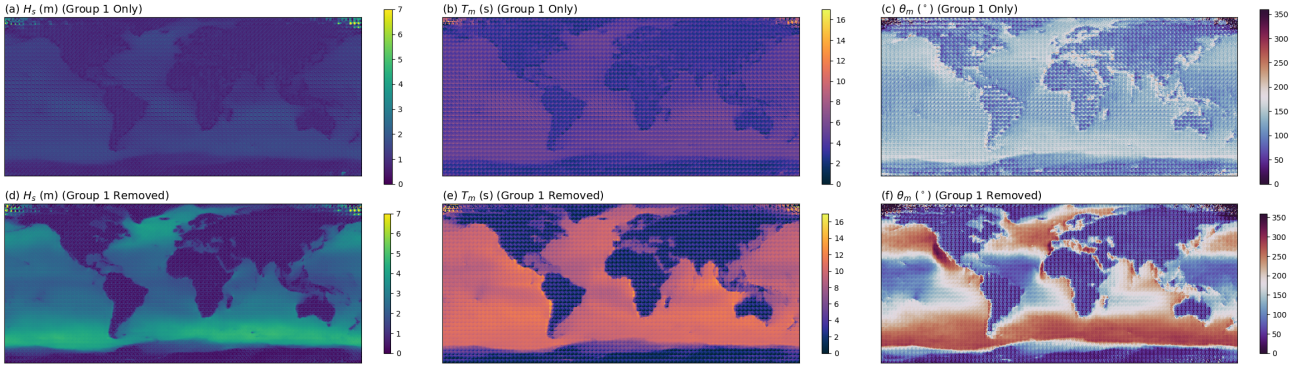

Figure 15. Functional analysis of Group 1 (Balancing) by isolation and ablation. Ensemble-averaged physical outputs (over 146 samples) from two complementary experiments. (a-c) The "isolation" experiment, showing the physical-space output when only the contribution from Group 1 is kept. (d-f) The "ablation" experiment, showing the output when the contribution from Group 4 is removed (zeroed-out). Columns correspond to H_s (m), T_m (s), and θ_m (°).

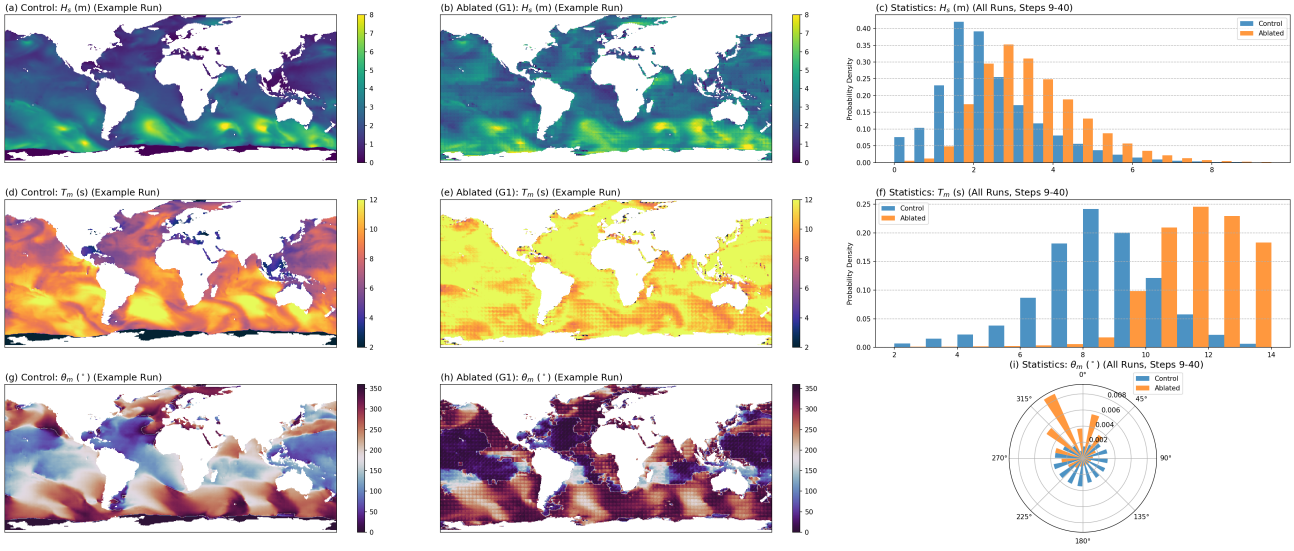

Figure 16. 10-day forecast ablation experiment for Group 1 (Balancing). The figure compares a Control run (all groups) and an Ablated run (Group 1 removed). The grid is organized by physical variable (rows) and analysis type (columns). Left Column (a, d, g): A snapshot of the final forecast state (Day 10, Step 40) from a single, representative run (Run #14). Middle Column (b, e, h): The corresponding final-state snapshot from the Ablated run (Group 1 removed). Right Column (c, f, i): Aggregate statistical distributions compiled from all 32 forecast runs, using data from forecast steps 9 through 40. Distributions show the Control (blue) vs. Ablated (orange) runs. Rows correspond to H_s (m), T_m (s), and θ_m (°).

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

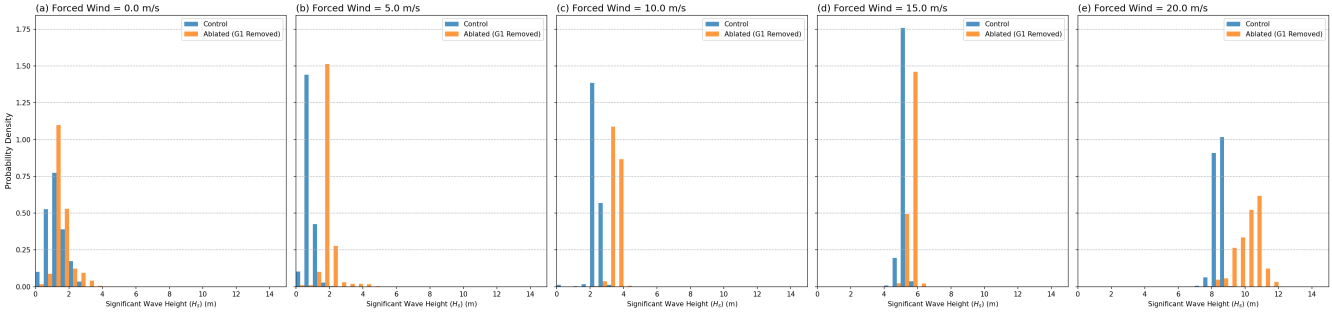

Figure 17. H_s distribution comparison for the Group 1 ablation under uniform, constant wind forcing. The figure shows the final probability density distributions of H_s (x-axis) after a 10-day (40-step) forecast. This experiment compares the Controlrun (blue) against the Ablated (G1 Removed) run (orange). Each panel (a-e) corresponds to a different global, uniform, and constant U-Wind forcing scenario: (a) 0 m/s, (b) 5 m/s, (c) 10 m/s, (d) 15 m/s, and (e) 20 m/s.

The 10-day (40-timestep) forecast ablation (Figure 16) quantifies this systemic breakdown and provides definitive evidence for Group 1's necessity in maintaining global energy balance. Visually, the "Masked" spatial maps (panels b, e, h) show H_s values that are globally higher and T_m values that are significantly "hotter" (higher period) than the Control run. The statistical distributions on the right reveal a systemic breakdown of the model's energy balance. The Mean Wave Period histogram (panel f, orange) undergoes a pronounced, non-physical drift to the right; without this balancing operator, the model's ability to maintain the wave spectrum is lost, with the distribution peak shifting to 12s and higher. This lack of energy dissipation also causes a significant shift in the Significant Wave Height histogram (panel c, orange), which is biased towards higher energy states.

A more detailed quantitative analysis under controlled forcing (Figure 17) reveals a crucial aspect of Group 1's function. We conducted an experiment forcing the model with uniform wind speeds (0 m/s to 20 m/s) for 10 days, comparing the Control run (blue) to a run with Group 1 removed (orange). The results are unambiguous: at every wind speed, the removal of Group 1 results in a non-physical shift to higher wave heights.

This finding is significant. It demonstrates that Group 1 (in the v1.0 model) acts as a constant, pervasive balancing operator, or a global "brake," whose dissipative function is essential regardless of the wind forcing. This contrasts with more complex, state-dependent behaviors, suggesting this model's implementation of "dissipation" is a simpler, more globally applied mechanism. This comprehensive analysis confirms that Group 1 is crucial for preventing excessive energy accumulation and maintaining the physical coherence of the wave spectrum.

3.4 Emergent Systemic Balance and Inter-dependency

A physical wave system is not a simple summation of independent source terms; it is a complex, non-linear system where terms are balanced and interdependent . Having identified the OCN v1.0 model's analogs for propagation (Group 4) and dissipation

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

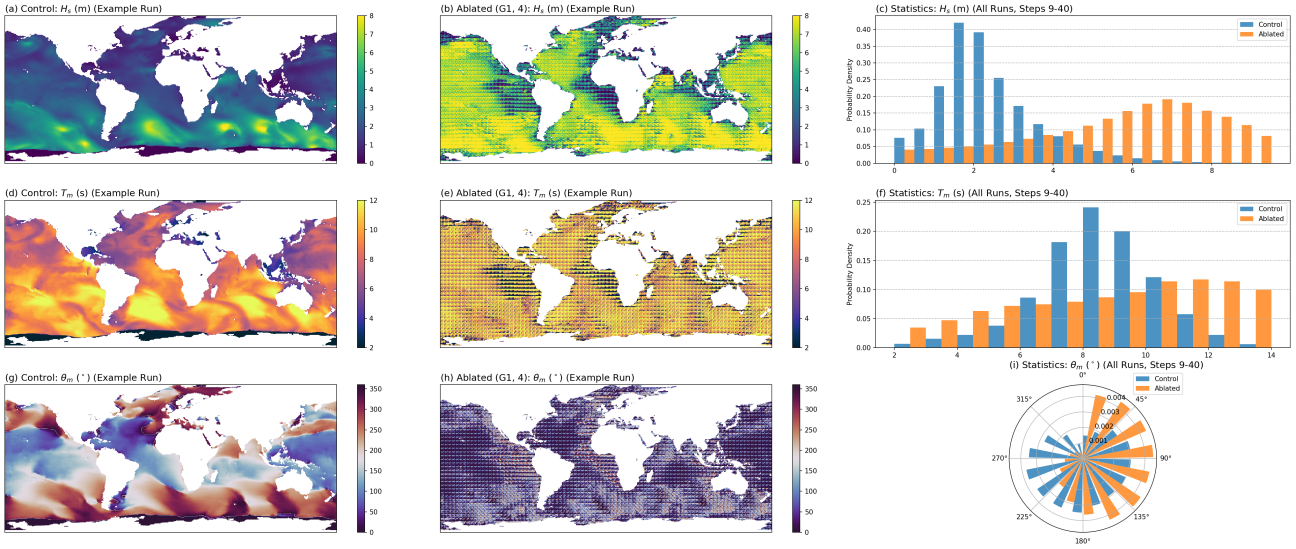

Figure 18. 10-day forecast ablation experiment for Group 1 and 4. The figure compares a Control run (all groups) and an Ablated run (Group 1 and 4 removed). The grid is organized by physical variable (rows) and analysis type (columns). Left Column (a, d, g): A snapshot of the final forecast state (Day 10, Step 40) from a single, representative run (Run #14). Middle Column (b, e, h): The corresponding final-state snapshot from the Ablated run (Group 1 and 4 removed). Right Column (c, f, i): Aggregate statistical distributions compiled from all 32 forecast runs, using data from forecast steps 9 through 40. Distributions show the Control (blue) vs. Ablated (orange) runs. Rows correspond to H_s (m), T_m (s), and θ_m (°).

(Group 1), we designed a final experiment to test this emergent systemic balance. We conducted a 10-day (40-timestep) forecast ablating both Group 4 (the foundation) and Group 1 (the "brake") simultaneously (Figure 18).

The results of this experiment are unambiguous. Unlike the "linear cancellation" illusion observed in other model iterations, the OCN v1.0 model, when stripped of both its primary climatology and balancing modules, undergoes a **total systemic collapse.

The "Masked" spatial maps (Figure 18, panels b, e, h) are physically meaningless, reverting to a uniform, low-energy state. This is quantified by the statistical histograms (panels c and f), which show the H_s and T_m distributions collapsing to sharp, low-value peaks, much like the G4-only ablation (Figure 11).

The Mean Wave Direction rose plot (panel i) provides the definitive evidence of a non-linear, hierarchical collapse. Here, the effects do not cancel or combine. The rose plot for the (G1+G4)-Masked run collapses into the exact same biased, non-physical distribution seen in the Group 4-only ablation (Figure 11).

This demonstrates that the OCN v1.0 model has learned a clear hierarchical dependency: the wave direction computation is fundamentally gated by the Group 4 (propagation/climatology) module. If that foundational layer is removed, the entire directional system fails in a specific, repeatable way, regardless of what happens to the dissipation operator (Group 1). This

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

complex interaction demonstrates that OCN has learned more than just source-term analogs; it has learned a complex, non-linear systemic balance where different physical variables are governed by different dependency structures .

4 Discussion

The opacity of data-driven "black box" models, and the associated persistent challenge regarding their credibility, remains a central issue in the physical sciences . This study addressed this challenge by applying the "glass box" dissection framework (Sec 2.2) to the OCN v1.0 model. The results, as detailed in Section 3, provide a direct and substantive counter-argument to the assumption that such models only learn non-physical statistical correlations. Our findings demonstrate that the OCN v1.0 model, to achieve high-fidelity predictions, has autonomously converged on a computational solution that is functionally analogous to the source-term balance at the heart of third-generation (3G) physical wave models .

Our findings show the model is not an inscrutable statistical correlator; it is a system that has learned to perform emergent functional partitioning. We statistically isolated and verified a foundational propagation and climatology module (Group 4), a wind-dependent wind-input operator (Group 3) analogous to S_{in}, and a dissipative balancing operator (Group 1) analogous to S_{ds}.

However, our analysis shows that the OCN v1.0 model's logic is more complex than just the summation of these primary components. The inability to isolate a single, discrete component for the non-linear wave-wave interaction term (S_{nl}) led us to investigate the roles of the remaining groups.

The experiment on Group 5 (Index 5) provides a compelling example (Figure 19). The 10-day forecast ablation shows that removing Group 5 does not cause a catastrophic failure (like removing G4 or G1), but rather a highly specific, complex change to the wave field.

The statistical plots quantify this. The H_s histogram (panel c) shows only a minor shift, with a decrease in 2m waves and an increase in sub-2m waves. The most significant impact is on the Mean Wave Period (T_m) (panel f): the distribution does not flatten, but instead specifically loses its long-period tail (>10s). Critically, the Mean Wave Direction (θ_m) rose plot (panel i) is almost identical to the control run, showing no systemic bias or collapse.

This complex failure mode—where removing one group only eliminates long-period swell while leaving the primary energy and direction intact—strongly suggests that Group 5 is not a simple operator. Rather, its function is critical to the model's learned long-period swell propagation physics and the coupling between swell and wind-sea. This functionally mirrors the role of S_{nl}, which manages the complex transfer of energy that underpins this swell-sea coupling.

The practical implications of this "glass box" dissection are direct. This functional map provides a pathway toward "Grey-box" model development. The failure of "black box" models to extrapolate to unseen extreme events is a primary limitation. Our work suggests a methodology to address this: now that we can statistically identify Group 3 as the S_{in} operator and Group 5 as a critical swell-coupling operator, it becomes feasible to perform targeted fine-tuning of only these components on extreme storm data, or apply physics-informed (PINN-like) constraints specifically to these groups to enforce known physical

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

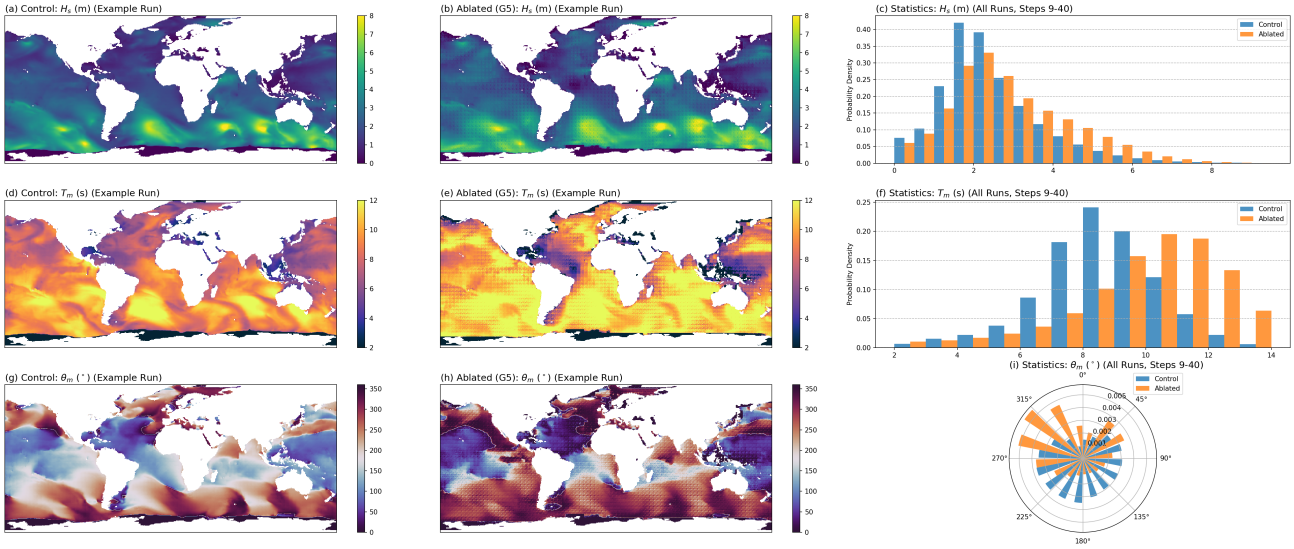

Figure 19. 10-day forecast ablation experiment for Group 5. The figure compares a Control run (all groups) and an Ablated run (Group 5 removed). The grid is organized by physical variable (rows) and analysis type (columns). Left Column (a, d, g): A snapshot of the final forecast state (Day 10, Step 40) from a single, representative run (Run #14). Middle Column (b, e, h): The corresponding final-state snapshot from the Ablated run (Group 5 removed). Right Column (c, f, i): Aggregate statistical distributions compiled from all 32 forecast runs, using data from forecast steps 9 through 40. Distributions show the Control (blue) vs. Ablated (orange) runs. Rows correspond to H_s (m), T_m (s), and θ_m (°)

laws. This work, therefore, provides a practical, methodological bridge from "black boxes" to the next generation of robust, physically-grounded, and trustworthy AI forecasting systems.

5 Conclusions

This study addressed the persistent challenge regarding the credibility of data-driven "black box" models by proposing and applying the "glass box" dissection framework (Sec 2.2). We hypothesized that a high-performance model (OCN v1.0) would, upon dissection, reveal an emergent computational structure that functionally mimics the physical principles of wave forecasting. Our findings provide strong, statistical evidence to support this hypothesis.

Our results demonstrate that the OCN v1.0 model is not an inscrutable black box; instead, its internal architecture has learned to perform an emergent functional partitioning. We successfully identified and validated distinct computational pathways analogous to the core components of third-generation (3G) physical wave models. Specifically, we identified a stable, high-contribution propagation and climatology module (Group 4) that provides the model's foundational geographic and swell field. This was complemented by a dynamic, wind-dependent wind-input operator (Group 3), analogous to S_{in}, which is re-

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

sponsible for generating wind-sea. Finally, we identified a globally-aware, dissipative balancing operator (Group 1), analogous to S_{ds}, which acts as the system's "brake" and prevents the non-physical, runaway accumulation of energy.

Furthermore, we found that the model's learned logic is more complex than a simple linear addition of these components. The discovery of "diffused" physics, such as the complex, non-catastrophic role of Group 5, and the hierarchical dependencies revealed by the G1+G4 ablation, suggests the model has captured a higher-order, non-linear systemic balance.

By translating the internal mechanisms of a DL model into the language of physical oceanography, this work provides a methodological blueprint for validating the physical fidelity of future AI Earth system models. This "glass box" dissection is a critical first step. It provides the necessary functional map to move beyond "black boxes" and begin developing "Grey-box" models. This knowledge of "which component does what" enables targeted, physics-constrained training, providing a concrete path toward solving the critical extrapolation challenges that currently limit purely data-driven forecasting.

Code and data availability. The source code for the "glass box" dissection framework, all evaluation data, and the exact OCN v1.0 model weights and parameters required to reproduce the figures and results in this paper are permanently archived and publicly available on Zenodo at: https://doi.org/10.5281/zenodo.17621476 (Zhang et al., 2025b) A detailed scientific description of the underlying OceanCastNet (OCN) model architecture is available in our previous publication (Zhang et al., 2025a). The evaluation scripts provided in the Zenodo archive are designed to be used with the OCN v1.0 weights included in the same archive.

Author contributions. Z.Z.: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Visualization, Writing – Original Draft. H.Y.: Conceptualization, Supervision, Project Administration, Funding Acquisition, Writing – Review & Editing. X.D.: Writing – Review & Editing. J.D.: Writing – Review & Editing. D.R.: Writing – Review & Editing. X.Q.: Conceptualization, Supervision, Writing – Review & Editing.

Competing interests. The authors declare there are no conflicts of interest for this manuscript.

Acknowledgements. We would like to express our gratitude to the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5 reanalysis data used for model analysis. During the preparation of this work the author(s) used Google's Gemini in order to assist with language polishing, refine the abstract, standardize technical terminology (e.g., 'AFNO2D Filter', 'Channel MLP'), and systematically improve the clarity and consistency of all figure titles and captions. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

References

Amina Adadi and Mohammed Berrada, "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", *IEEE access : practical innovations, open solutions*, volume 6, pages 52138–52160, IEEE, 2018.

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey, "Machine Learning for Precipitation Nowcasting from Radar Images", *arXiv preprint arXiv:1912.12132*, 2019.

Fabrice Ardhuin, Erick Rogers, Alexander V Babanin, Jean-François Filipot, Rudy Magne, Aaron Roland, Andre Van Der Westhuysen, Pierre Queffeulou, Jean-Michel Lefevre, Lotfi Aouf, and others, "Semiempirical Dissipation Source Functions for Ocean Waves. Part I: Definition, Calibration, and Validation", *Journal of Physical Oceanography*, volume 40, number 9, pages 1917–1941, 2010.

Alexander Babanin, *Breaking and Dissipation of Ocean Surface Waves*, Cambridge University Press, 2011.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian, "Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks", *Nature*, volume 619, number 7970, pages 533–538, Nature Publishing Group UK London, 2023.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics", *Electronics*, volume 8, number 8, pages 832, Multidisciplinary Digital Publishing Institute, 2019.

Luigi Cavaleri and Paola Malanotte Rizzoli, "Wind Wave Prediction in Shallow Water: Theory and Applications", *Journal of Geophysical Research: Oceans*, volume 86, number C11, pages 10961–10973, Wiley Online Library, 1981.

Mariana CA Clare, Maike Sonnewald, Redouane Lguensat, Julie Deshayes, and Venkatramani Balaji, "Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics", *Journal of Advances in Modeling Earth Systems*, volume 14, number 11, pages e2022MS003162, Wiley Online Library, 2022.

Abhirup Dikshit and Biswajeet Pradhan, "Explainable AI in Drought Forecasting", *Machine Learning with Applications*, volume 6, pages 100192, Elsevier, 2021.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda, "Transcoders Find Interpretable Llm Feature Circuits", *Advances in Neural Information Processing Systems*, volume 37, pages 24375–24410, 2024.

Svenja Ehlers, Norbert Hoffmann, Tianning Tang, Adrian H Callaghan, Rui Cao, Enrique M Padilla, Yuxin Fang, and Merten Stender, "Physics-Informed Neural Networks for Phase-Resolved Data Assimilation and Prediction of Nonlinear Ocean Waves", *Physical Review Fluids*, volume 10, number 9, pages 094901, APS, 2025.

The Wamdi Group, "The WAM Model—A Third Generation Ocean Wave Prediction Model", *Journal of physical oceanography*, volume 18, number 12, pages 1775–1810, 1988.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro, "Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers", *arXiv preprint arXiv:2111.13587*, 2021.

Peter AEM Janssen, "Wave-Induced Stress and the Drag of Air Flow over Sea Waves", *Journal of Physical Oceanography*, volume 19, number 6, pages 745–754, American Meteorological Society, 1989.

RE Jensen, VJ Cardone, and AT Cox, "Performance of Third Generation Wave Models in Extreme Hurricanes", in *9th International Wind and Wave Workshop, Victoria, BC*, 2006.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, and others, "Neural General Circulation Models for Weather and Climate", *Nature*, volume 632, number 8027, pages 1060–1066, Nature Publishing Group UK London, 2024.

https://doi.org/10.5194/egusphere-2025-5749
Preprint. Discussion started: 8 January 2026

Gerbrand Johan Komen, L Cavaleri, M Donelan, K Hasselmann, S Hasselmann, PAEM Janssen, and others, *Dynamics and Modelling of Ocean Waves*, Cambridge university press UK, volume 532, 1994.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, and others, "Learning Skillful Medium-Range Global Weather Forecasting", *Science*, volume 382, number 6677, pages 1416–1421, American Association for the Advancement of Science, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar, "Fourier Neural Operator for Parametric Partial Differential Equations", *arXiv preprint arXiv:2010.08895*, 2020.

Maximilian Li and Lucas Janson, "Optimal Ablation for Interpretability", *Advances in Neural Information Processing Systems*, volume 37, pages 109233–109282, 2024.

Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith, "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning", *Bulletin of the American Meteorological Society*, volume 100, number 11, pages 2175–2199, 2019.

Richard Meyes, Melanie Lu, Constantin Waubert De Puiseau, and Tobias Meisen, "Ablation Studies in Artificial Neural Networks", *arXiv preprint arXiv:1901.08644*, 2019.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, and others, "Fourcastnet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators", *arXiv preprint arXiv:2202.11214*, 2022.

Michael T Pearce, Thomas Dooms, Alice Rigg, Jose M Oramas, and Lee Sharkey, "Bilinear MLPs Enable Weight-Based Mechanistic Interpretability", *arXiv preprint arXiv:2410.08417*, 2024.

W Erick Rogers, Alexander V Babanin, and David W Wang, "Observation-Consistent Input and Whitecapping Dissipation in a Model for Wind-Generated Surface Waves: Description and Simple Calculations", *Journal of Atmospheric and Oceanic Technology*, volume 29, number 9, pages 1329–1346, 2012.

Y Qiang Sun, Pedram Hassanzadeh, Mohsen Zand, Ashesh Chattopadhyay, Jonathan Weare, and Dorian S Abbot, "Can AI Weather Models Predict Out-of-Distribution Gray Swan Tropical Cyclones?", *Proceedings of the National Academy of Sciences*, volume 122, number 21, pages e2420914122, National Academy of Sciences, 2025.

Hendrik L Tolman, "Effects of Numerics on the Physics in a Third-Generation Wind-Wave Model", *Journal of physical Oceanography*, volume 22, number 10, pages 1095–1111, 1992.

Hendrik L Tolman and others, "User Manual and System Documentation of WAVEWATCH III TM Version 3.14", *Technical note, MMAB Contribution*, volume 276, number 220, 2009.

Matthew D Zeiler and Rob Fergus, "Visualizing and Understanding Convolutional Networks", in *European Conference on Computer Vision*, pages 818–833, Springer, 2014.

Ziliang Zhang, Huaming Yu, Danqin Ren, Chenyu Zhang, Minghua Sun, and Xin Qi, "Ocean Wave Forecasting with Deep Learning as Alternative to Conventional Models", *Journal of Advances in Modeling Earth Systems*, volume 17, number 11, pages e2025MS005285, 2025.

Ziliang Zhang, Huaming Yu, Xiaotian Dong, Jiaqi Dou, Danqin Ren, and Xin Qi, "OceanCastNet (OCN) v1.0: Source Code and Evaluation Data [Data set]", *Zenodo*, https://doi.org/10.5281/zenodo.17621476, 2025.