



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

COMET-QR: Conformal Multi-fidelity Transformer with Quantile Regression for Antarctic Sea-Ice Concentration Forecasting

TESI DI LAUREA MAGISTRALE IN
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Giorgio Aurina, 10715220

Advisor:
Prof. Andrea Manzoni

Co-advisors:
Dr. Paolo Conti

Academic year:
2024-2025

Abstract: Antarctic Sea-Ice plays a critical role in regulating the global climate system through its influence on ocean circulation, albedo, and atmosphere-ocean coupling. Despite decades of satellite and model analyses, the underlying dynamics governing Antarctic Sea-Ice variability remain poorly understood due to strong regional contrasts, high natural variability, and complex feedbacks between atmospheric, oceanic, and cryospheric processes. Traditional physical and statistical models have struggled to integrate these nonlinear interactions, particularly when combining heterogeneous datasets with different spatial and temporal resolutions. This study develops an innovative two-stage approach to improve Sea-Ice forecasting, combining (i) a data-driven physics-based algorithm, consisting of a Bagging optimised Dynamic Mode Decomposition, and (ii) a deep learning framework based on Transformer architectures to investigate Antarctic Sea-Ice dynamics across multiple data sources and levels of fidelity, overall providing a unified approach to learn from diverse observational and simulated records. By leveraging the attention mechanism, the model captures spatiotemporal dependencies and cross-modality relationships between satellite observations, reanalysis products, and climate model outputs. Moreover, quantile regression is employed to produce calibrated confidence intervals for the predicted Sea-Ice Concentration. Preliminary analyses indicate that Transformer-based approaches can represent large-scale variability and regional anomalies more flexibly than conventional methods. Overall, the proposed COMET-QR (COnformal Multi-fidELity Transformer with Quantile Regression) strategy enhances our understanding of Antarctic Sea-Ice behaviour, by taking into account also extreme events, and bridging the gap between data-driven inference and physically based modelling.

Key-words: Sea-Ice Concentration Forecasting, Progressive Multi-Fidelity Learning, Dynamic Mode Decomposition, Transformers, Conformal Quantile Regression, Spatiotemporal Predictions

1. Introduction

The Antarctic cryosphere stands as one of the critical cores of the Earth's climate engine. Far from being a static indicator, Antarctic ice acts as an active thermodynamic modulator, regulating the interfacial exchange of heat, momentum, and humidity between the Southern Ocean and the atmosphere [28]. Its seasonal cycle of

growth and retreat drives the global thermohaline circulation, which fundamentally alters ocean stratification, as shown in [34].

The urgency of accurate Sea-Ice forecasting is underscored by recent observational anomalies in the Southern Hemisphere. As reported by The New York Times [11], Antarctic Sea-Ice reached a record low annual maximum in September 2023, peaking at just under 17 million square kilometres, as we can see in Figure 1. This deviation is statistically significant: the deficit compared to the 1981–2010 average represents an area roughly the size of Alaska. This event marks a potential "regime shift" in polar dynamics. While Antarctic Sea-Ice remained relatively stable until 2016, it has since entered a period of volatility characterised by frequent record lows.

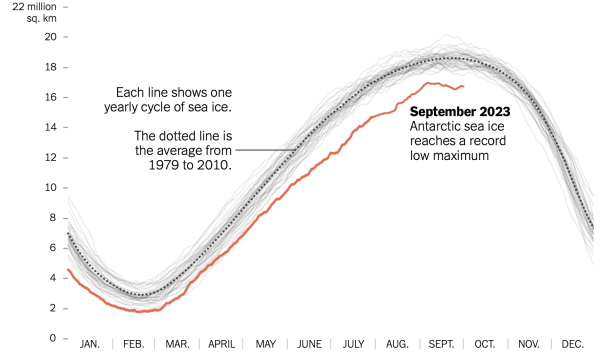


Figure 1: Rapid Decline Scenario, from [11].

Consequently, the capacity to accurately forecast the spatiotemporal evolution of Sea-Ice Concentration (SIC) constitutes a strategic imperative, serving both as a diagnostic for climatological stability and as an operational necessity for polar navigation. However, the Antarctic system is governed by high-dimensional spatiotemporal chaos, characterised by unbounded interactions between atmospheric forcing and oceanic currents. This complexity is mirrored in the observational record: the dataset utilised in this work, comprising 33 years of daily observations (1991–2023), reveals a system in transition. Unlike the Arctic, which exhibits a relatively linear decline, the extent of Antarctic Sea-Ice has demonstrated complex behaviours and extreme volatility in recent decades, challenging traditional linear extrapolations and pushing the predictive limits of deterministic physical models.



Figure 2: Antarctica is surrounded by a belt of frozen ocean called Sea-Ice [2].

1.1. Methodological Framework

1.1.1 Cost of acquiring and simulating data

A fundamental barrier to effective forecasting lies in the prohibitive cost of acquiring and simulating accurate state representations. Most theoretical models are heavily dependent on where the data are collected, and the Antarctic region is one of the least accessible for humans. For this reason, the cost of collecting measurements from physical sensors is very high, so we need to use simulated data instead or, in the context of our application, convert raw satellite radiance into reliable probability maps of Sea-Ice presence, which is a complex inverse problem, troubled by retrieval ambiguities caused by atmospheric attenuation and land contamination. Simulated data are obtained by solving complex Partial Differential Equations (PDEs) governing fluid dynamics and thermodynamics, typically using Finite Element Methods (FEM). Resolving the fine-scale interactions needed

for accurate local forecasting requires mesh densities that render FEM computationally intractable for real-time operational scenarios or large-scale ensemble forecasting [32].

1.1.2 Physics-Based vs Data-Driven: exploring possibilities for SIC forecasting

Conventional approaches to Sea-Ice forecasting rely primarily on either physics-based numerical models or purely data-driven statistical methods. Physics-based models, while providing interpretable representations of underlying processes, often struggle with computational efficiency and parameter uncertainty. Conversely, statistical approaches may lack physical consistency and fail to capture long-term dynamical structures. Given these computational and observational constraints, there is a pressing need for a data-driven framework capable of emulating the physical dynamics at a fraction of the cost while robustly handling the inherent uncertainty of the observations. Antarctic sea-ice dynamics are notoriously difficult to model due to their complex, non-stationary nature. Recent deep learning approaches have shown promise in addressing this challenge. For instance, [13] demonstrated that Convolutional Neural Networks (CNNs) could effectively correct systematic biases in sea-ice models by learning from data assimilation increments. Similarly, [37] utilised Convolutional Long Short-Term Memory (ConvLSTM) networks to capture spatiotemporal dependencies, successfully predicting the seasonal evolution of Antarctic Sea-Ice during the rapid decline scenarios we have been facing in the last decade. However, despite these advances, accurately modelling the full system remains an immense challenge, also because most existing approaches utilise single-fidelity data sources, missing opportunities to leverage complementary information across different scales and data types.

Recent advances in computer vision have been marked by a paradigm shift with the introduction of Vision Transformers (ViT) [10]. Unlike traditional CNNs, which rely on local receptive fields and translation invariance, Transformers utilise Self-Attention mechanisms to capture global dependencies across the entire input domain and they have a much lower inductive bias. However, the standard ViT approach, which partitions an image into a fixed grid of rectangular patches, is ill-suited for the Antarctic environment. To address this geometric constraint, in [section 1.2](#) we present how to treat the entire spatial field of each single data source as a distinct semantic unit.

1.1.3 Hybrid approach, enhancing with surrogate modelling

To overcome the cost of acquiring and simulating data, the engineering community often turns to approximation techniques such as surrogate models, e.g., Neural Networks and Gaussian Processes, to make fast predictions. Within surrogate models, Reduced Order Models (ROMs) aim to reconstruct the full solution field (e.g., the entire map of Sea-Ice) by combining dimensionality reduction with a model that runs in a compressed, or reduced, latent space. Furthermore, when working with multiple sources of data, Multi-Fidelity (MF) surrogate models have emerged as a critical strategy to avoid the prohibitive computational costs associated with high-fidelity solvers, by leveraging low-fidelity datasets to speed-up computations in many applications such as uncertainty quantification and sensitivity analysis. In the following, we present some recent MF architectures that offer robust strategies for complex forecasting domains where high-fidelity data are sparse and computational resources are constrained. For example, in [14] four distinct neural network architectures have been designed to fuse data from varying fidelity levels, categorising them into "all-in-one" and "multilevel" models. The obtained results indicate that the accuracy of the multi-fidelity prediction is driven primarily by the quality of the low-fidelity ROM rather than the quantity of expensive HF data. This suggests that investing computational resources into improving the ROM is a more efficient strategy than generating additional high-fidelity snapshots. Addressing the scarcity of high-fidelity data for training, in [9], the authors propose a Multi-Fidelity Reduced-Order Surrogate modelling strategy. This method synergises dimensionality reduction with multi-fidelity neural network regression to efficiently approximate time-parameter-dependent solutions of partial differential equations (PDEs). The method first extracts a spatial basis from a limited set of high-fidelity snapshots using Proper Orthogonal Decomposition (POD); then, to model the temporal evolution of the reduced states, [9] employs Long Short-Term Memory (LSTM) neural networks. The LSTM is trained to map low-fidelity (LF) POD coefficients, which were derived from inexpensive, coarse simulations, to their high-fidelity counterparts; this mapping allows the surrogate to correct physical inaccuracies inherent in the LF model while leveraging its low computational cost. This work highlights a critical advancement in non-intrusive surrogate modelling: the ability to fuse data from disparate fidelity levels without requiring modification to the underlying solvers. Unlike purely data-driven methods that may lack physical consistency, this approach incorporates physically meaningful low-fidelity data, enhancing the interpretability and reliability of the surrogate even when high-fidelity data coverage is sparse. As we have mentioned, Antarctic Sea-Ice dynamics are notoriously difficult to model due to their complex, non-stationary nature. The available data differ significantly in origin, ranging from real observations to synthetic simulations, as well as in their domain coverage and inherent noise levels. Thus, this scenario is well-suited for a multi-fidelity (MF) modelling framework, as data sources of varying accuracy and cost are combined to

optimise the prediction task. It is worth emphasising once again that these components are categorized into high-fidelity (precise but expensive) and low-fidelity (less accurate but computationally cheap) sources, relying on a combination of both to balance computational cost with performance.

1.2. The Proposed Method

In this work, we provide a data-driven framework for Sea-Ice prediction that combines a physics-based surrogate model to replace expensive simulations with a multi-fidelity strategy. This framework assimilates data from multiple sources to correct biases and improve overall predictive accuracy. To tackle the problem, this work introduces a hybrid framework that addresses these challenges through an Agnostic Multi-modal Learning paradigm. We re-conceptualise the forecasting problem not as a direct regression, but as a combination between an Offline Surrogate Model and a Residual Learning task. We decompose the system dynamics into two distinct regimes: (i) a coherent linear structure captured by a data-driven physics-based surrogate model, and (ii) a nonlinear stochastic residual modelled with deep learning.

The core surrogate model is based on Dynamic Mode Decomposition (DMD). DMD offers a computationally efficient and physics-grounded approach, creating a reduced-order surrogate model directly from snapshot data. DMD combines linear dimensionality reduction with a linear model to describe the dynamics, where the evolution of the system is determined by the identified dynamical spectrum (modes and eigenvalues). In particular, we employ Bagging optimised Dynamic Mode Decomposition (BOPDMD) [21] to extract the dominant coherent structures, thus enhancing modes' robustness and providing an uncertainty quantification framework. A critical strategic advantage of this spectral approach over traditional numerical solvers lies in the distinct decoupling of computational costs. While the decomposition of massive historical datasets is computationally intensive, it is entirely confined to an offline training phase, thus the online inference is reduced to a simple analytical evaluation of linear combinations. This guarantees $\mathcal{O}(1)$ runtime complexity, providing instantaneous, physics-consistent baselines in real-time operational settings without the need for expensive time-stepping integration.

With the stable dynamics handled efficiently offline, we employ a Multi-Fidelity-Transformer (MELTER) architecture [25] to model the complex, high-frequency residuals between our DMD surrogate model and the ground truth. This Transformer assimilates online data streams from diverse sources, including both simulations and measurements across different modalities, to enhance the surrogate's predictions by leveraging the assimilated information. Inspired by the philosophy of recent scientific foundation models [27], we adopt an Omnimodal approach: we treat all available information as equal tokens in a shared latent space. We employ tokenization to render all data modalities comparable, allowing the model to learn how to best exploit available information through attention mechanisms. We argue that a priori hierarchies are sub-optimal in dynamic environments where the utility of a data source is context-dependent. To encode these high-dimensional spatial fields into compact latent tokens, we combine a spatial encoder, using Proper Orthogonal Decomposition (POD), with a temporal encoder, using Gated Recurrent Units (GRU). We flatten the irregular spatial domain into a dense vector and feed it directly into the GRU. Furthermore, a masking strategy is implemented to handle missing modalities robustly, ensuring the model can provide reliable predictions (potentially falling back to the baseline DMD surrogate), regardless of the specific combination of available inputs. Thus, the Transformer receives the sequence of fidelity tokens (the GRU outputs) along with a binary mask indicating which sources are currently available and, through the Self-Attention mechanisms, it creates a dynamic "information marketplace", learning *end-to-end* which source to trust for every specific pixel and time-step. To preserve signal integrity and enhance gradient flow across these deep representations, we incorporate skip connections around the attention blocks. The resulting sequence of re-weighted fidelity tokens is then aggregated via a global pooling layer, distilling the diverse multi-source information into a unified latent context vector. To explicitly account for the non-stationary nature of the system, the current time coordinate is concatenated with this context vector before reconstruction in the decoding phase. The latter is initially handled by a GRU which unrolls the temporal evolution of the latent state, and then passed through a Multi-Layer Perceptron (MLP) to project the dense features back into the high-dimensional spatial domain, effectively recovering the complex residual field.

Finally, acknowledging the stochastic nature of the target, a probabilistic field of ice presence, we move beyond deterministic point forecasts [29]. We formulate the learning objective as Conformalized Quantile Regression (CQR). This ensures that our forecasts are not merely accurate in the mean squared sense, but are statistically calibrated, providing rigorous confidence intervals that explicitly account for the uncertainty inherent in both the satellite retrieval process and the chaotic dynamics of the system. This probabilistic formulation stands in contrast to recent deterministic state-of-the-art architectures such as SICFormer [18]. While SICFormer successfully leverages a 3D-Swin Transformer to capture spatiotemporal dependencies for end-to-end daily Sea-Ice prediction, it relies on minimising Mean Squared Error (MSE). By treating the problem as a quantile regression rather than a direct point-wise reconstruction, our framework acknowledges that minimising MSE is insufficient for risk-aware operational settings in the polar regions.

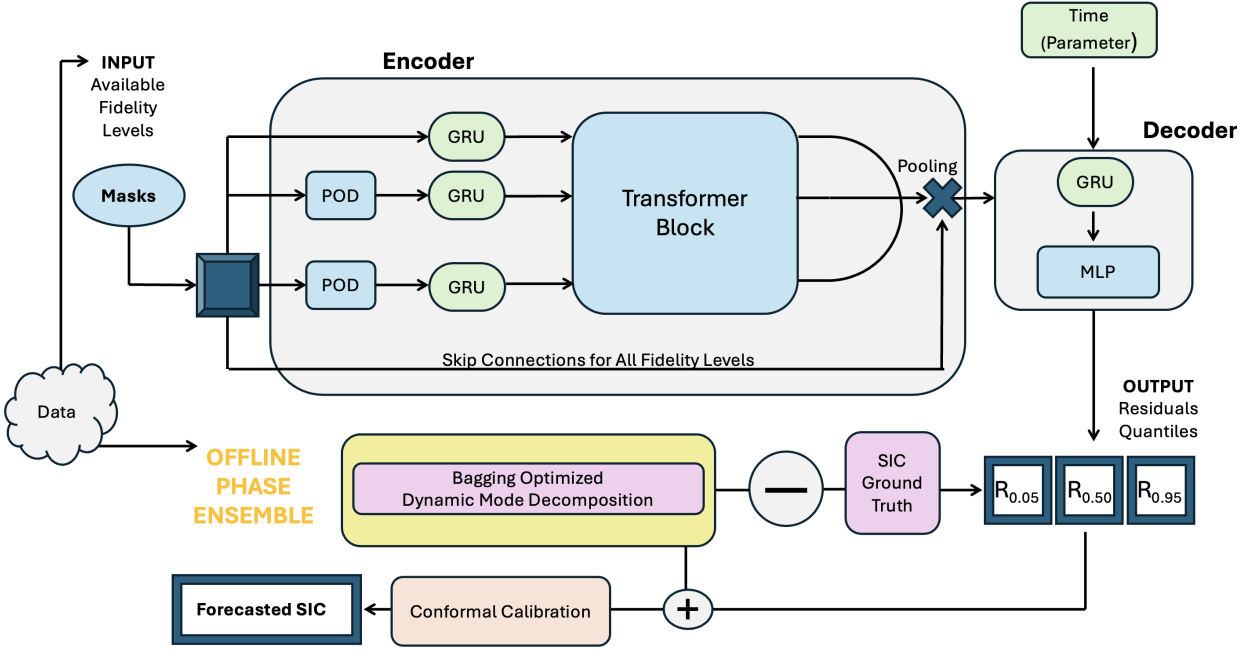


Figure 3: COMET-QR Architecture

Figure 3 illustrates the COMET-QR architecture, detailing how the Multi-fidelity Transformer refines the initial BOPDMD point estimation, subsequently processed through a Conformal calibration stage to produce reliable SIC quantile predictions.

This work pursues two primary research objectives culminating in the COMET-QR architecture:

- **Multi-Fidelity Fusion Strategy:** Demonstrating that relying on single-source data is insufficient for capturing non-linear climate dynamics. Through targeted ablation studies, we investigate the relative importance of global features and localized error proxies, seeking to achieve a significant predictive improvement over the physical BOPDMD surrogate.
- **Probabilistic Forecasting Framework:** Recognising the limitations of deterministic point estimates in chaotic climate systems, we couple Quantile Regression with Conformal Prediction to produce statistically calibrated confidence bounds.

The remainder of this work is organized as follows:

- **Section 2** presents the observational dataset (target SIC and predictors SIT, SST), detailing acquisition protocols and spatiotemporal distributions. These features are treated as distinct input modalities to model Sea-Ice Concentration.
- **Section 3** focuses on the offline surrogate model: Bagging optimised Dynamic Mode Decomposition (BOPDMD). Rooted in Koopman Operator Theory for linearizing non-linear dynamics, we derive the DMD algorithm and explore enhancements (time embeddings, Fast Fourier Transform). A key contribution of this section is the implementation of Bagging (Bootstrap + Aggregation), which we employ to stabilize the spectral decomposition against noise and initial condition sensitivity, ensuring a robust, single-fidelity baseline that effectively isolates the dominant coherent linear structures of the system.
- **Section 5** establishes the framework for Uncertainty Quantification (UQ) via Conformalized Quantile Regression (CQR), shifting focus from deterministic accuracy to probabilistic calibration. For Sea-Ice forecasting, calibration must be dual-faceted, holding spatially (pixel-wise coverage) and temporally (across varying lead times), ensuring statistically valid confidence intervals under distributional shifts.
- **Section 6** evaluates the empirical performance of COMET-QR across varying data availability scenarios. It demonstrates the predictive superiority of multi-fidelity fusion and visually maps the spatiotemporal reliability of the conformalized prediction intervals.
- **Section 7** synthesises these findings, concluding on the hybrid approach’s efficacy. It discusses broader implications, highlighting how integrating physics-based priors with deep learning and rigorous uncertainty quantification advances risk-aware operational climate modelling.

2. Data Provenance and Exploration

2.1. Sea-Ice Concentration and Ice Mask

The primary variable of interest is the Sea-Ice Concentration (SIC), represented as a scalar field $n_t(x, y)$ describing the fraction of sea surface covered by ice at a given time t and spatial location (x, y) . The values represent probabilities, therefore they are strictly bounded in the unit interval $n \in [0, 1]$, where 0 corresponds to open ocean and 1 indicates a completely frozen surface. A critical feature of the dataset is the topological constraint imposed by the Antarctic continent. Since Sea-Ice dynamics are physically restricted to the oceanic domain, the value of the concentration field on land is null by definition ($n = 0$). To handle this, the dataset includes a binary continent mask (as a boolean tensor). This mask identifies the grid points corresponding to landmasses, allowing the model to distinguish between valid zero-values (open ocean) and invalid structural zeros (land), effectively excluding the latter from the learning process. While the Antarctic continent mask excludes landmasses, a significant portion of the original 432×432 grid corresponds to the open ocean at lower latitudes (typically north of -60°), where Sea-Ice never forms regardless of the season. Including these static regions ($SIC \equiv 0$) in the training process would introduce unnecessary computational overhead and potentially bias the network towards predicting trivial zeros. To address this, we restricted our analysis to the ice mask, shown in Figure 4. This mask identifies the spatial subset where Sea-Ice dynamics actually occur. A grid point (x, y) is included in the domain only if it exhibits a non-zero Sea-Ice concentration variance across the historical dataset of daily observations (1991–2023). This spatial filtering drastically reduces the dimensionality of the problem. Instead of processing the full rectangular grid ($\approx 186,000$ pixels), the model operates on a flattened vector of $N_{active} = 43,736$ valid pixels.

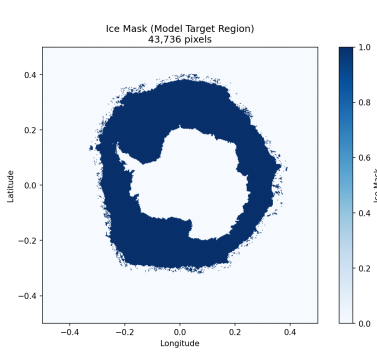


Figure 4: Ice Mask

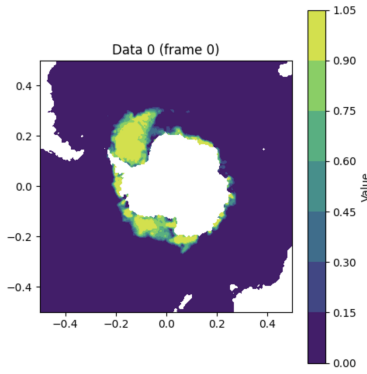


Figure 5: SIC Year 1, Day 1.

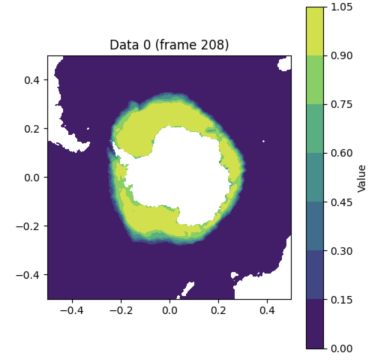


Figure 6: SIC Year 1, Day 209.

The Antarctic Sea-Ice exhibits a clear but approximate periodicity, driven by the seasonal cycle of freezing and melting. Consequently, a Climatology approach, that is, simply predicting using the historical average for a given day of the year, can achieve deceptively low error metrics. However, such an approach is scientifically insufficient as it lacks a generative model of the system. Climatology is a static statistical baseline that ignores the current state of the cryosphere, rendering it incapable of predicting inter-annual variability, anomalies, or regime shifts. A true predictive model must go beyond reproducing this mean seasonal cycle; it must demonstrate the ability to capture the causal mechanisms and high-frequency patterns that drive the system away from its climatological norm.

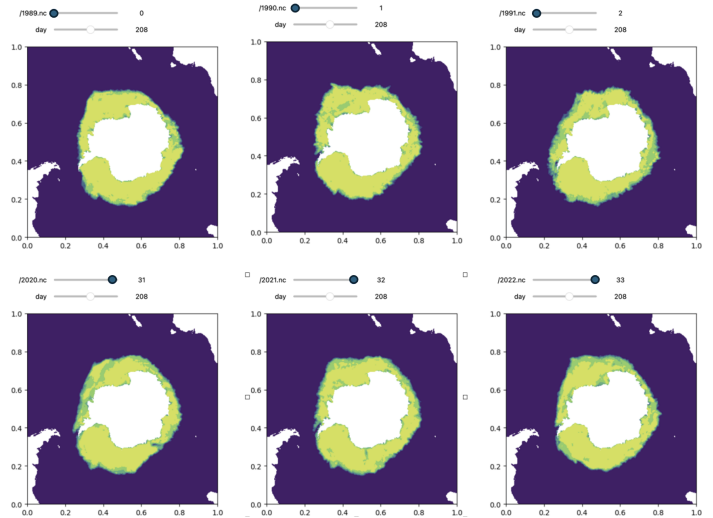


Figure 7: Day 209 over years. 1st row: 1989, 1990, 1991. 2nd row: 2020, 2021, 2022. SIC is clearly periodic only "approximately"

2.1.1 Data Provenance and Usage

The SIC data were obtained in a pre-processed format from *The Alan Turing Institute (London)*, where raw satellite telemetry was converted into consistent probabilistic concentration fields. This primary dataset was subsequently augmented with Sea-Ice Thickness (SIT) and Sea Surface Temperature (SST) records from the *Copernicus Marine Service* to provide a more comprehensive physical description of the domain. Within our proposed two-phase framework, these auxiliary features are reserved exclusively for the online phase. Rather than being incorporated into the initial offline surrogate construction, they serve as distinct input modalities for the Multi-Fidelity Transformer. This design allows the architecture to assimilate real-time environmental data streams independently, leveraging the attention mechanism to dynamically weigh their relevance in correcting the baseline DMD predictions. By treating these variables as transient online inputs, we maintain a computationally efficient offline stage while providing the Transformer with the high-resolution information necessary to resolve complex, nonlinear residuals.

2.2. Justification for the Use of GLORYS12 Reanalysis

The environmental input features exploited in this study are derived from the GLORYS12V1 Global Ocean Physics Reanalysis, a state-of-the-art product distributed by the Copernicus Marine Environment Monitoring Service (CMEMS) [12]. This dataset represents a high-resolution, multi-decadal global ocean simulation generated using the NEMO (Nucleus for European Modelling of the Ocean) platform, coupled with the LIM3 (Louvain-la-Neuve Sea-Ice Model) Sea-Ice component. The choice of GLORYS12 is predicated on its high eddy-resolving spatial resolution ($1/12^\circ$, approximately 9 km at the equator, and significantly finer at high Antarctic latitudes), which is essential for capturing mesoscale processes that drive Sea-Ice dynamics. Unlike raw satellite observations, which are frequently hindered by sensor limitations, such as cloud cover obscuring optical sensors or orbital gaps, GLORYS12 provides a spatially complete and temporally continuous dataset. This continuity is achieved through Data Assimilation (DA) [22]. The system employs a Reduced Order Kalman Filter (SEIK) to strictly constrain the model physics with millions of daily in-situ observations (e.g., Argo floats, shipboard measurements). For a machine learning framework, this "gap-free" setting avoids the need for artificial interpolation of missing values, ensuring that the training data preserves consistency and physical validity across the entire Antarctic domain. To enhance the predictive capability of the model regarding Sea-Ice Concentration (SIC), two critical auxiliary variables were extracted and processed:

- Sea Water Potential Temperature (**thetao**), specifically sampled at the surface layer (0.49 m depth); this variable represents the temperature of the liquid ocean interface. In polar regions, this metric acts as a critical thermodynamic boundary condition. While atmospheric temperatures may collapse drastically, the temperature under Sea-Ice remains thermodynamically constrained near the freezing point of saline water ($\approx -1.8^\circ\text{C}$). Fluctuations above this threshold are the primary driver of basal melting and Sea-Ice retreat, serving as a high-sensitivity precursor to variations in SIC. Although SIC indicates the extent of coverage (0-100%), it lacks information regarding thermal inertia.
- Sea-Ice Thickness (**sithick**) provides the system volume information. A thick multi-year ice pack exhibits significantly higher resilience to thermal forcing compared to thin, newly formed ice, despite both potentially showing 100% concentration. Integrating thickness allows the Multi-Fidelity Transformer model to distinguish between fragile and robust ice cover, thereby improving the accuracy of temporal predictions.

By leveraging the GLORYS12 reanalysis, this study grounds its data-driven approach in rigorous physical modelling. The dataset provides a verified ground truth that synthesises decades of observational data with fluid dynamic laws, offering the multi-fidelity spatiotemporal context required for progressive learning tasks.

2.3. Data Preprocessing

Both SIT and SST cover the temporal range from January 1, 1993, to December 31, 2023, with a daily temporal resolution. The raw data is originally distributed in a standard geographical coordinate system (WGS84). A critical preprocessing phase involved the spatial harmonisation of the input datasets. A distinct discrepancy existed between the target and feature grids: the Sea-Ice Concentration (SIC) is defined on a projected 432×432 grid with a spatial resolution of 25×25 km, whereas the auxiliary Sea Surface Temperature (SST) and Sea-Ice Thickness (SIT) data were provided in a geographic latitude-longitude format with a finer resolution of approximately 5×5 km. To align these modalities, we performed a downsampling operation to map the high-resolution features onto the coarser SIC grid. Specifically, we computed the local arithmetic mean of the 25 source pixels (corresponding to a 5×5 sub-grid block) that fall within the footprint of a single target pixel. This averaging strategy ensures that the regridded values physically represent the integrated thermal and structural properties of the 25km cell, rather than relying on point-wise interpolation.

1. Coordinate Transformation and Regridding

Given the polar nature of the region of interest, the original latitude-longitude grid is unsuitable due to the severe distortion near the pole. We reprojected the data into the Antarctic Polar Stereographic coordinate system (EPSG:3031). We defined a fixed target grid of size 432×432 pixels centred on the South Pole. Through linear interpolation, the data was regridded to a standardised spatial resolution of 25 km per pixel. This specific spatial configuration was adopted to ensure a perfect alignment with the existing Sea-Ice Concentration (SIC) dataset, thereby guaranteeing pixel-wise coherence across all the different physical variables fed into the model.

2. Land and Missing Value Handling

The raw CMEMS product utilises NaN (Not a Number) values to mask landmasses and ice shelves where oceanic data is unavailable. During preprocessing, these missing values were imputed with zeros (0.0). This operation ensures that the input tensors are dense and numerical, a strict requirement for the layers of the network. Consequently, a value of 0 indicates the absence of water (land) or data outside the model domain.

3. Serialization

After processing, the daily snapshots were aggregated into annual tensors and verified for spatial consistency. The final dataset was serialized into binary files (Pickle format) to optimise Input/Output operations during the training phase.

Table 1: Summary of dataset specifications.

Parameter	Specification
Source	CMEMS Global Ocean Reanalysis
Variables	SST, Sea-Ice Thickness (SIT)
Time Period	1993 – 2023 (31 Years)
Temporal Res.	Daily
Coord. System	Antarctic Polar Stereographic
Spatial Res.	25 km
Grid Dims	432×432

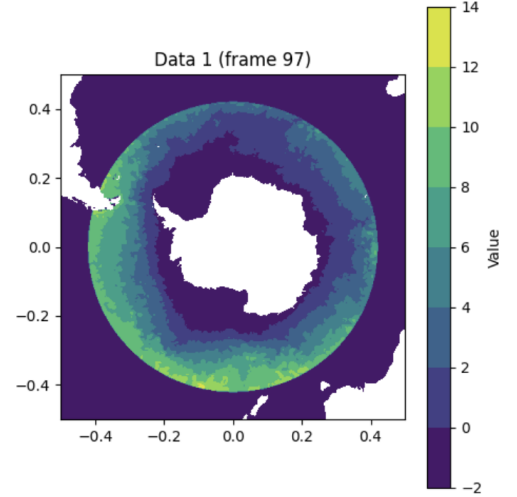


Figure 8: Sea Surface Temperature (SST) [°C]. Range: $[-2.992, 17.041]$.

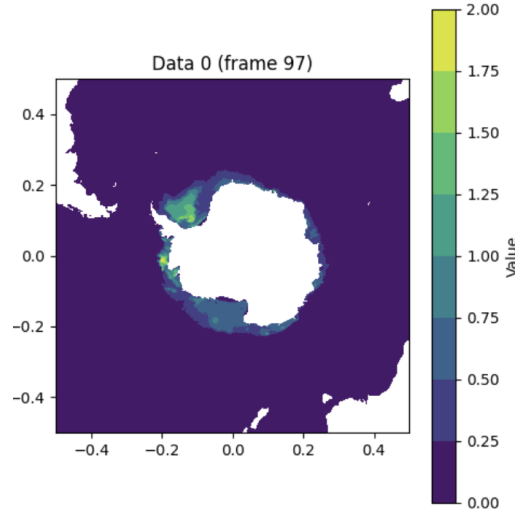


Figure 9: Sea-Ice Thickness (SIT) [m]. Range: $[0, 9.53]$.

2.3.1 Preliminary Spatiotemporal Exploratory Analysis

Before implementing the multi-fidelity Transformer framework, we carried out a preliminary spatiotemporal exploration of the dataset to establish a baseline of the long-term climate signal. This diagnostic phase involves quantifying secular trends across the 1993–2023 period using both parametric and non-parametric statistical methods. By aggregating daily observations into annual means, we isolate inter-annual variability from seasonal noise, allowing for a pixel-wise assessment of climatic shifts in Sea-Ice Concentration (SIC), Thickness (SIT), and Surface Temperature (SST). A comprehensive description of these exploratory protocols, including the Ordinary Least Squares (OLS) regression metrics and the results of the Mann-Kendall (MK) trend significance tests, is provided in Appendix A. To model the temporal evolution of Sea-Ice Concentration without relying on auxiliary variables, one might consider classical univariate time-series forecasting methods. For example, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model provides a well-established statistical framework capable of capturing both secular linear trends and periodic seasonal cycles, as explained in Appendix B. Transitioning from classical statistical methods to deep learning approaches, recurrent architectures such as Long Short-Term Memory (LSTM) networks are naturally well-suited to capture these complex, non-linear temporal dependencies.

3. BOPDMD: Derivation and Offline Phase

In this section, we establish the mathematical framework governing the offline surrogate component of our hybrid architecture. We rely on the hypothesis that the chaotic, high-dimensional evolution of Sea-Ice Concentration, $n_t(x, y)$, can be synthesised in a lower-dimensional coherent structure governed by physical laws. To extract this structure without relying on computationally expensive numerical solvers, we turn to Operator Theory.

3.1. Koopman Operator Theory: Linearizing the Nonlinear

The analysis of dynamical systems has traditionally relied on the geometric perspective introduced by Poincaré, which studies trajectories in the state space $\mathcal{M}$. While intuitive, this approach struggles with high-dimensional, highly nonlinear systems like Sea-Ice dynamics, where analytical linearization (e.g., via Jacobian analysis) is only valid locally and for short time horizons.

In 1931, B.O. Koopman introduced an alternative, operator-theoretic perspective [20]. Instead of analysing the evolution of the state vector $x \in \mathcal{M}$ directly, Koopman proposed analysing the evolution of observable functions defined on the state space.

3.1.1 The Operator Definition

Consider a discrete-time dynamical system on a manifold $\mathcal{M} \subseteq \mathbb{R}^n$:

$$x_{k+1} = F(x_k) \quad (1)$$

where F is a nonlinear map representing the physical evolution (e.g., Navier-Stokes or thermodynamic equations). Let $g : \mathcal{M} \rightarrow \mathbb{C}$ be a scalar-valued observable function (e.g., a measurement of Sea-Ice Concentration at a specific coordinate).

The Koopman operator $\mathcal{K}$ is an infinite-dimensional linear operator that acts on the space of observable functions $\mathcal{F}$ (typically a Hilbert space such as $L^2(\mathcal{M})$):

$$\mathcal{K}g(x_k) := g(F(x_k)) = g(x_{k+1}) \quad (2)$$

The fundamental insight here is the trade-off between complexity and dimensionality. While the underlying dynamics F are finite-dimensional but highly nonlinear, the Koopman operator $\mathcal{K}$ is linear but infinite-dimensional [7]. This allows us to apply powerful tools from linear spectral theory to analyze nonlinear systems globally, avoiding the errors introduced by local linearization.

3.1.2 Spectral Decomposition

Since $\mathcal{K}$ is a linear operator, its behaviour is fully characterised by its eigenvalues $\lambda_j \in \mathbb{C}$ and eigenfunctions $\varphi_j \in \mathcal{F}$:

$$\mathcal{K}\varphi_j(x) = \lambda_j\varphi_j(x) \quad (3)$$

If we consider a vector of p observables $g(x) = [g_1(x), \dots, g_p(x)]^T$, and we assume it lies within the span of the eigenfunctions, we can expand the state evolution as:

$$g(x_k) = \sum_{j=1}^{\infty} v_j \lambda_j^k \varphi_j(x_0) \quad (4)$$

where v_j are the Koopman modes [24]. This expansion implies that the dynamics of an arbitrary nonlinear system can be represented as a superposition of linear oscillations (characterised by the phase of λ_j) and exponential growth/decay (characterised by the magnitude $|\lambda_j|$), provided the correct basis functions φ_j are identified.

3.2. Dynamic Mode Decomposition (DMD)

While Koopman theory provides the rigorous theoretical framework, finding the infinite-dimensional eigenfunctions φ_j analytically is computationally intractable for complex geophysical systems. Dynamic Mode Decomposition (DMD), introduced by Schmid (2010) [30], serves as a robust data-driven algorithm to approximate the spectral properties of the Koopman operator directly from finite data measurements.

3.2.1 The DMD Algorithm

Given a sequence of m snapshot vectors, we arrange them into two matrices: X (the state at time t) and X' (the state at time $t + 1$):

$$X = [x_1, x_2, \dots, x_{m-1}], \quad X' = [x_2, x_3, \dots, x_m] \quad (5)$$

DMD seeks the best-fit linear operator A that maps X to X' in a least-squares sense:

$$A = \underset{A}{\operatorname{argmin}} \|X' - AX\|_F = X'X^\dagger \quad (6)$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse.

In high-dimensional systems, such as our 432×432 Sea-Ice grids, the matrix A is too large to construct explicitly. DMD circumvents this by projecting the dynamics onto a low-rank subspace:

1. SVD Reduction: Compute the Singular Value Decomposition of $X \approx U_r \Sigma_r V_r^*$, retaining rank r .
2. Projection: Compute the reduced-order operator $\tilde{A} \in \mathbb{C}^{r \times r}$:

$$\tilde{A} = U_r^* X' V_r \Sigma_r^{-1} \quad (7)$$

3. Eigendecomposition: Solve the eigenvalue problem $\tilde{A} \tilde{w}_j = \lambda_j \tilde{w}_j$.
4. Mode Reconstruction: Recover the high-dimensional DMD modes $\phi_j = U_r \tilde{w}_j$.

3.2.2 Connection to Koopman Theory

Under specific conditions, the eigenvalues λ_j and modes ϕ_j extracted by DMD converge to the eigenvalues and modes of the Koopman operator ($\phi_j \approx v_j$) [21]. Specifically, if the observables $g(x)$ span a finite-dimensional Koopman invariant subspace, DMD recovers the underlying linear dynamics exactly. For general nonlinear systems, variants such as Extended DMD (EDMD) or Hankel-DMD utilise nonlinear dictionary functions or time-delay embeddings to better approximate the infinite-dimensional nature of $\mathcal{K}$.

In this work, we utilise Bagging Optimised DMD (BOPDMD). By aggregating multiple DMD models over bootstrapped temporal windows, we estimate the stable, physical eigenvalues of the Koopman operator while filtering out spurious modes induced by observational noise. This provides a robust and consistent baseline for our Transformer architecture.

3.3. Model Selection Strategy

3.3.1 State Definition and Rolling Window Strategy

Let $n_t \in \mathbb{R}^{N_p}$, denote the flattened state vector representing the Sea-Ice Concentration (SIC) at time t , where N_p is the number of active pixels in the region of interest. The complete historical dataset spans a period of 33 years with daily observations of $N_p = 432 * 432$ points:

$$\mathcal{D} = \{n_t\}_{t=1}^{33*365}$$

To capture the nonstationary nature of Antarctic Sea-Ice, we reject the assumption of a static dynamical operator over the entire 33-year period. Instead, we employ a 2-year sliding window strategy to generate the physics-based baseline. For every target year y_k (where $k \geq 3$), the BOPDMD model is fitted exclusively on the daily snapshots of the two preceding years, y_{k-2} and y_{k-1} (a training horizon of $m = 730$ days). Once the operator is learned from this local window, it is used to forecast the full evolution of the subsequent year y_k in a closed loop. The window then slides forward by one year, using y_{k-1} and y_k to forecast y_{k+1} , and so on. This iterative process generates a continuous sequence of 1-year out-of-sample forecasts starting from 1993 (year 3), in agreement with the time range of SST and SIT. These concatenated forecasts form the final surrogate prediction that serves as the input prior for our residual learning framework. This ensures that the baseline adapts to inter-annual climate shifts while strictly preventing information leakage from the future.

3.3.2 Time-Delay Embedding (Hankel Transform)

Standard DMD assumes the system state is fully observable. However, Sea-Ice dynamics are governed by latent variables (e.g., subsurface ocean temperature) not present in the SIC snapshots. Relying on Takens' embedding theorem [33], we augment the state space by stacking $d = 2$ time-shifted copies of the data into a Hankel matrix $\mathcal{H}$, with $m = 730$ (2 years sliding window):

$$\mathcal{H} = \begin{bmatrix} n_1 & n_2 & \cdots & n_{m-d+1} \\ n_2 & n_3 & \cdots & n_{m-d+2} \\ \vdots & \vdots & \ddots & \vdots \\ n_d & n_{d+1} & \cdots & n_m \end{bmatrix} \in \mathbb{R}^{(d \cdot N_p) \times (m-d+1)}.$$

This transformation linearizes trajectories that appear nonlinear in the original lower-dimensional space, allowing the spectral decomposition to capture standing waves and long-term memory effects.

3.3.3 Physics-Informed Constraints

The decomposition seeks eigenvalues λ_j and modes ϕ_j that approximate the operator A such that $n_{t+1} \approx A n_t$. To ensure physical consistency with the bounded nature of Sea-Ice Concentration ($0 \leq \text{SIC} \leq 1$), we enforce two critical constraints during the optimisation:

1. **Stability Constraint:** We require the continuous-time eigenvalues $\omega_j = \ln(\lambda_j)/\Delta t$ to have a non-positive real part, ensuring non-divergent forecasts, i.e., $\Re(\omega_j) \leq 0 \iff |\lambda_j| \leq 1$.
2. **Conjugate Pairs:** Since the SIC field lives in a subset of $\mathbb{R}$ ($n \in [0, 1]$), complex modes must appear in conjugate pairs. This ensures that the imaginary components cancel out in the reconstruction: $\lambda \in \sigma(A) \implies \bar{\lambda} \in \sigma(A)$.

3.4. Hyperparameter Tuning

The performance of the physics-based prior relies heavily on the optimal selection of model complexity (truncation rank) and the quality of the input signal. To determine the most robust configuration, we conducted a systematic tuning campaign using a standard Hankel-DMD formulation. The primary objective was to validate two architectural hypotheses:

1. whether spectral preprocessing (FFT low-pass filtering) enhances the stability of dynamical modes against high-frequency observational noise, and
2. identifying the optimal truncation rank r that balances reconstruction accuracy with predictive capability.

For the former, the selection process is grounded in a comparative evaluation framework where the model's forecasting skill is benchmarked against a Climatology baseline, defined as the historical day-of-year average. We quantify performance using the relative improvement in the Mean Squared Error (MSE) for the predicted spatial ice-integral, over a 1-year forecast horizon. We recall that a prevalent challenge in satellite-derived datasets is the presence of high-frequency noise and atmospheric artifacts. Thus, we hypothesized that such noise could lead to spurious high-frequency modes and, to address this, we introduced a temporal preconditioning step based on spectral filtering. For each spatial degree of freedom, we computed the Discrete Fourier Transform (DFT) of the training time series and applied a symmetric ideal low-pass mask defined by a cutoff ratio ρ . We performed a grid search over bandwidths $\rho \in \{0.02, 0.03, 0.05, 0.08, 0.09\}$.

Empirical Findings: Contrary to the initial hypothesis, the experimental results indicated that spectral filtering does not yield a statistically significant improvement in forecasting skill compared to the raw data as we can see in Figure 10. The dynamical modes extracted from the filtered series did not exhibit enhanced stability, and in some instances, aggressive filtering ($\rho < 0.05$) removed genuine high-frequency physical dynamics, degrading the model's responsiveness to rapid onset events. Consequently, to minimize pipeline complexity and avoid introducing arbitrary frequency cutoffs, we discarded the FFT preprocessing step, feeding the raw satellite observations directly into the time-delay embedding.

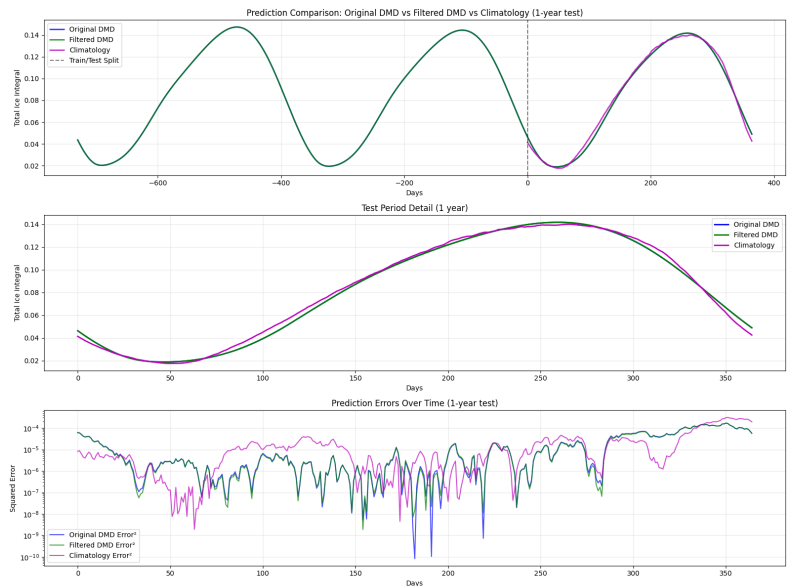


Figure 10: FFT pre-processing vs original DMD vs CLIMATOLOGY predictions comparison

Rank Selection

The second critical hyperparameter is the truncation rank r , which determines the dimension of the subspace onto which the dynamics are projected. This parameter governs the bias-variance trade-off: a low rank captures only the most energetic coherent structures (high bias), while a high rank risks overfitting spectral noise (high variance). We evaluated the forecasting performance of Hankel-DMD for ranks $r \in \{5, 6, 7\}$ over a sliding validation window and using a temporal thinning of 3 days. The quantitative analysis, as shown in Figure 11, revealed the following:

- Rank $r = 7$: Achieved the lowest absolute error, providing a marginally richer representation of the dynamics.
- Rank $r = 5$: Demonstrated performance nearly identical to $r = 7$, with the difference in the improvement of the MSE being negligible for operational purposes.

Adhering to the principle of parsimony (Occam’s Razor), we selected rank $r = 5$ as the optimal configuration. A lower-rank model is inherently more robust to instabilities and computationally lighter. Physically, this rank is sufficient to capture the fundamental coherent structures of the system (the mean flow, the seasonal cycle, and the primary modes of the Antarctic dynamics), while deliberately delegating the modelling of finer-scale higher-rank complexities to the subsequent deep learning correction stage.

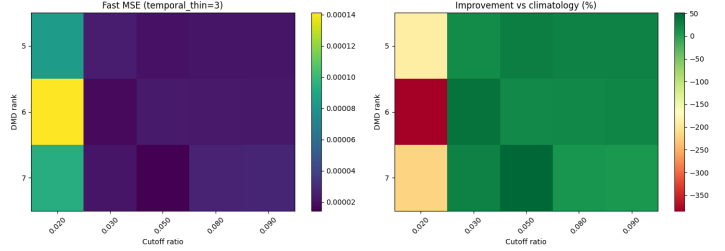


Figure 11: Error Heatmap for optimal Rank selection

3.5. Final Physics Configuration of Bagging Optimised DMD

Based on these empirical findings, the final configuration for the physics prior module is defined as follows:

- **Input Processing:** Raw SIC data (No FFT filtering).
- **Embedding:** Time-Delay with $d = 2$. To capture the latent non-Markovian dynamics inherent in the Antarctic cryosphere, we augment the input constructing a Hankel matrix representation of the system state.
- **Truncation Rank:** $r = 5$ (Dominant 5 modes).
- **Constraints:** Stability ($\text{Re}(\omega) \leq 0$) and Conjugate Pairs enforced to ensure physical consistency and real-valued forecasts.

To establish a robust dynamical baseline for the hybrid framework, we employ Bagging Optimised Dynamic Mode Decomposition (BOPDMD) [21]. Unlike standard DMD, which provides a single spectral decomposition sensitive to measurement noise and transients, BOPDMD utilises bootstrap aggregating (bagging) to produce a statistical ensemble of models. This allows for the estimation of spectral uncertainty and the filtering of spurious dynamical modes. The state $y(t)$ is reconstructed as a linear superposition of spatiotemporal modes:

$$y^{(k)}(t) = \sum_{j=1}^r b_j^{(k)} \phi_j^{(k)} e^{\omega_j^{(k)} t} \quad (8)$$

where $\phi_j^{(k)}$ represent the spatial DMD modes, $\omega_j^{(k)}$ are the complex eigenvalues determining the frequency and growth/decay rates, and $b_j^{(k)}$ are the mode amplitudes. This local approximation serves as the physics-based prior for the subsequent year y_k .

3.5.1 Algorithms

The complete procedure for extracting the robust physics prior is detailed in Algorithm 1. The algorithm iteratively samples subsets of the temporal history, computes a low-rank approximation via optimised DMD, and aggregates only the statistically stable modes. Operating as a purely data-driven dynamical systems approach, the BOPDMD pipeline reshapes the spatiotemporal data into snapshots and employs Singular Value Decomposition (SVD) to extract the dominant linear modes of the system. To ensure robustness, a rank-5 bootstrap ensemble is utilised, where 100 distinct DMD models are trained on resampled data, and their mean prediction is taken as the baseline forecast.

Algorithm 1 Physics-Constrained Bagging Optimised DMD (BOPDMD)

Require: Snapshot matrix $\mathbf{X}$, Delay embedding d , Window length m , Number of bags B , Target Rank r .

- 1: **Step 1: Hankel Augmentation**
 - 2: Construct Hankel matrix $\mathcal{H}$ from $\mathbf{X}$ with embedding dimension d .
 - 3: Split $\mathcal{H}$ into input $\mathcal{H}_1$ (time $1 \rightarrow T-1$) and output $\mathcal{H}_2$ (time $2 \rightarrow T$).
 - 4: Initialise empty ensemble sets: $\Lambda_{ens} \leftarrow \emptyset, \Phi_{ens} \leftarrow \emptyset$.
 - 5: **for** $b = 1$ to B **do**
 - 6: *// Bootstrap Sampling*
 - 7: Sample random time indices $\mathcal{I}_b \subset \{1, \dots, m-d\}$ with replacement.
 - 8: Construct subsampled matrices $\mathbf{Y}_1 = \mathcal{H}_1[:, \mathcal{I}_b]$ and $\mathbf{Y}_2 = \mathcal{H}_2[:, \mathcal{I}_b]$.
 - 9: *// POD Projection (SVD)*
 - 10: Compute SVD: $\mathbf{Y}_1 \approx \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^*$.
 - 11: *// Optimised Operator Approximation*
 - 12: Compute reduced operator: $\tilde{\mathbf{A}} = \mathbf{U}_r^* \mathbf{Y}_2 \mathbf{V}_r \mathbf{\Sigma}_r^{-1}$.
 - 13: Solve eigenproblem: $\tilde{\mathbf{A}} \mathbf{w}_j = \lambda_j \mathbf{w}_j$.
 - 14: *// Physics-Informed Filtering*
 - 15: **for** each eigenvalue λ_j in spectrum **do**
 - 16: **if** $|\lambda_j| > 1 + \epsilon$ **then**
 - 17: Discard λ_j (Unstable mode).
 - 18: **else**
 - 19: Project to unit circle if necessary: $\lambda_j \leftarrow \lambda_j / |\lambda_j|$.
 - 20: Reconstruct High-Dim Mode: $\phi_j = \mathbf{U}_r \mathbf{w}_j$.
 - 21: Add (λ_j, ϕ_j) to ensemble $\{\Lambda_{ens}, \Phi_{ens}\}$.
 - 22: **end if**
 - 23: **end for**
 - 24: **end for**
 - 25: **Step 3: Ensemble Aggregation**
 - 26: Cluster eigenvalues in Λ_{ens} to identify r dominant centroids $\bar{\lambda}_k$.
 - 27: Compute mean modes $\bar{\phi}_k$ corresponding to centroids.
 - 28: **Step 4: Amplitude Optimisation**
 - 29: Solve for amplitudes $\mathbf{b}$ to minimise reconstruction error at $t = 0$:
 - 30: $\mathbf{b} = \arg\min_{\mathbf{b}} \|\mathbf{x}_0 - \sum_{k=1}^r b_k \bar{\phi}_k\|_2$.
 - 31: **return** Stable Eigenvalues $\bar{\Lambda}$, Modes $\bar{\Phi}$, Amplitudes $\mathbf{b}$.
-

The output of Algorithm 1 defines the linear predictor $y_{k,t} = \sum b_k \bar{\phi}_k e^{\omega_k t}$. Because the model employs a two-year sliding window for its initial training, forecasting begins in the third year. This yields a 31-year prediction period spanning from 1993 to 2023, perfectly aligning with the temporal availability of the auxiliary SIT and SST datasets discussed later in this work. To properly quantify the uncertainty inherent in the spectral decomposition, we employ a bootstrap ensemble forecasting strategy with $B = 100$ realizations. The resulting set of daily ensemble forecasts is denoted as:

$$\{y_{k,t}\}_{k=1}^B \quad \forall t$$

The ensemble mean provides the robust baseline forecast used for the subsequent multi-fidelity integration:

$$y_t^{DMD} = \frac{1}{B} \sum_{k=1}^B y_{k,t} \quad \forall t \quad (9)$$

Additionally, the ensemble standard deviation is computed to capture the variability of the DMD approximation:

$$\sigma_t = \sqrt{\frac{1}{B-1} \sum_{k=1}^B (y_{k,t} - y_t^{DMD})^2} \quad \forall t \quad (10)$$

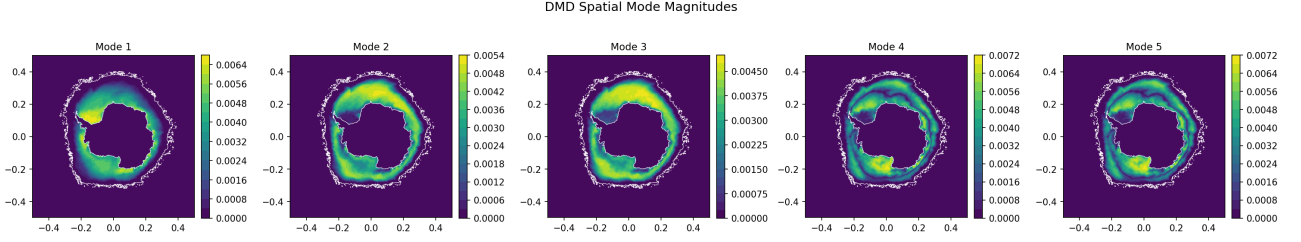


Figure 12: Identified BOPDMD Spatial Modes for the window 2021-2022

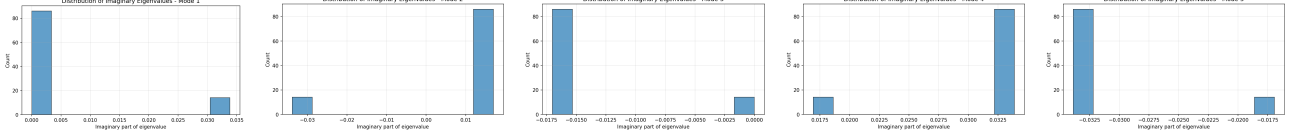


Figure 13: Eigenvalue distribution of the extracted modes following the BOPDMD algorithm. Across almost all trials, Mode 1 is associated with a purely real eigenvalue, thus representing non-oscillatory dynamics such as pure exponential decay. The remaining four modes strictly adhere to the complex conjugate pair constraint, yielding oscillatory behaviour. Specifically, Modes 2 and 3 capture the annual cycle ($T = 1$ year), while Modes 4 and 5 isolate the semi-annual harmonic ($T = 0.5$ years).

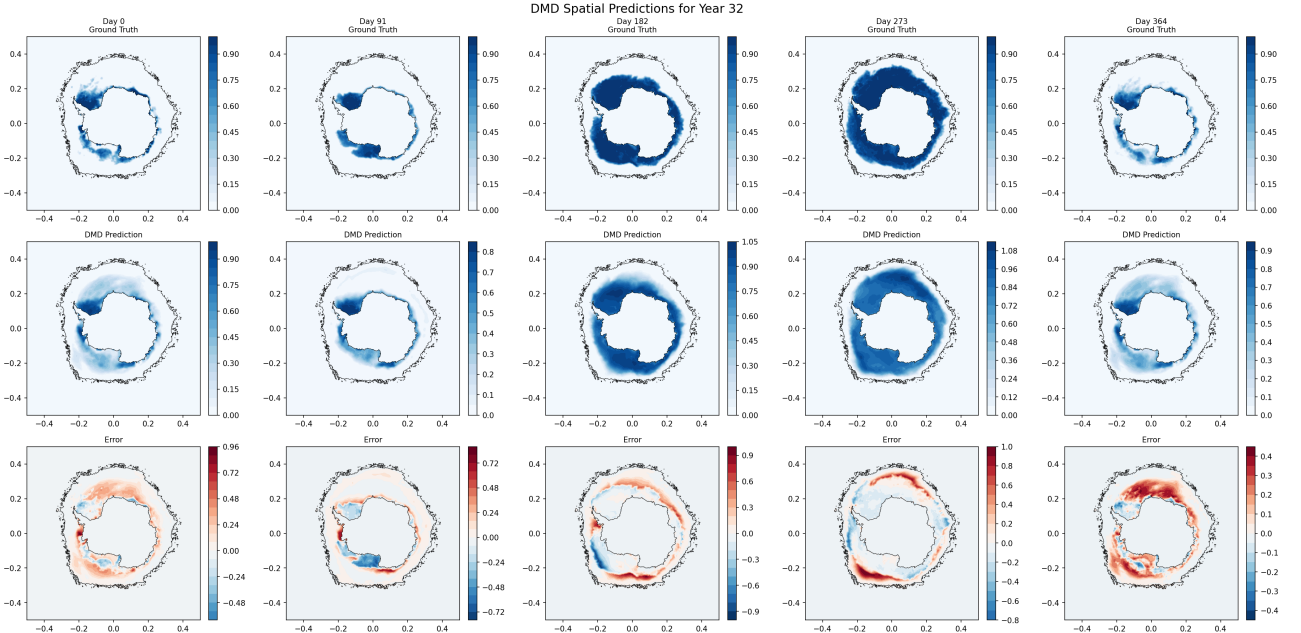


Figure 14: BOPDMD predictions in 5 days of 2023: 1st row Ground Truth. 2nd row DMD Forecasted SIC. 3rd row Prediction Error

3.6. Toward the Online Residual Learning Phase

Having obtained the SIC predictions from the BOPDMD model, we computed the daily spatial residuals with respect to the ground truth. To maintain consistency with their physical interpretation as probabilities, we bounded the predicted values within the $[0, 1]$ interval. Since the first two years of the dataset are consumed to initialise the first rolling window, the effective dataset for the residual learning phase spans 31 years. The residual field is computed as the point-wise difference between the SIC ground truth and the DMD forecast: $r_{i,t} = n_{i,t} - y_{i,t}^{DMD} \quad \forall i, \quad \forall t$. The final hybrid model combines the zero-cost inference of the spectral prior with the probabilistic correction of the deep learning module:

$$\hat{n}_t = \underbrace{y_t^{DMD}}_{\text{DMD Prior}} + \underbrace{f_\theta(\text{Inputs})}_{\text{MF-Transformer Correction}} \quad (11)$$

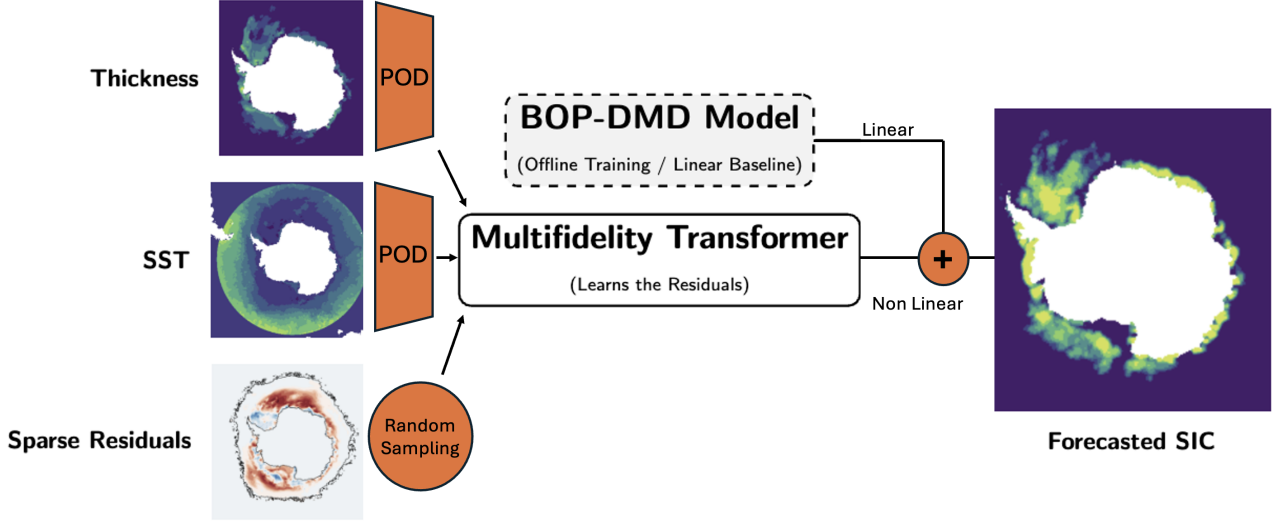


Figure 15: The Multi-Fidelity Transformer is trained to correct the surrogate prediction, by learning the quantiles of the residual, using different data sources which are encoded as low-fidelity inputs.

3.7. Low-Fidelity Input Modalities Encoding

In our multi-fidelity (MF) framework, the high-fidelity (HF) target is the Sea-Ice Concentration (SIC) residual. To achieve accurate predictions, we define three distinct low-fidelity (LF) input modalities, denoted as $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$. These LF data provide the necessary physical context and localised information required for the Transformer to correct the baseline surrogate. Each modality is extracted from raw observations or preliminary models and processed into temporal sequences of tokens. To address the high dimensionality of the spatial fields, specifically Sea-Ice Thickness (SIT) and Sea Surface Temperature (SST), we employ Proper Orthogonal Decomposition (POD) [31] to project the dynamics onto a low-dimensional space. This approach identifies the most energetic coherent structures, as it captures the long-term structural evolution of the ice pack and the cross-correlation between oceanic thermal forcing, allowing us to represent complex dynamics using a compact set of temporal coefficients. For each input feature, we compute the number of modes that capture more than 90% of the total system variance. Operating on the flattened grid of 432×432 pixels, the POD algorithm successfully reduces the spatial fields into highly compact latent representations (for a detailed derivation of the POD basis and the Method of Snapshots, the reader is referred to Appendix B). Specifically, the Sea-Ice Thickness (SIT) field was reduced to just 95 modes, while the Sea Surface Temperature (SST) field was compressed into 91 modes. This choice provides a significant dimensionality reduction (approximately $2000\times$). The third modality, $x^{(3)} \in \mathbb{R}^{80}$, represents the Sparse Residual Sensors. Unlike the global POD features, $x^{(3)}$ is obtained by randomly sampling the BOPDMD residual field, defined as the discrepancy between the high-fidelity ground truth and the BOPDMD prediction, at 80 random locations. These samples are strictly restricted to the active ice mask $\mathcal{M}_{\text{ice}}$ to ensure the model focuses on physically active regions. This modality serves as a sparse real-time error map, providing the high-frequency information necessary to learn the proper corrections. It is worth mentioning that, to optimise memory usage during the Transformer’s attention and decoding phases, the spatial data is cropped to a 300×308 bounding box after the POD dimensionality reduction has been performed.

In the end, given that the modalities possess different physical units and variances, we apply a collective Robust Scaling. Normalisation is performed on the combined set of inputs $x^{(1)}, x^{(2)}, x^{(3)}$ after extraction but prior to the GRU layers. This approach ensures that the Transformer’s attention mechanism can effectively weigh the distinct tokens ($d_{\text{emb}} = 64$), without being biased by numerical scale.

To bridge the gap between these heterogeneous data types, each modality $x^{(i)}$ is passed through a dedicated Gated Recurrent Unit (GRU) encoder. The GRU processes a continuous temporal lookback sequence of length 365 days. These sequences are extracted using a sliding window with a stride of 30 days, generating a sequence of latent tokens that encapsulate the historical evolution of that specific source before being passed to the attention mechanism. For the k -th token, the input sequence is defined as:

$$X_{\text{seq}}^{(k)} = \{x_{30k}^{(i)}, x_{30k+1}^{(i)}, \dots, x_{30k+365}^{(i)}\}.$$

With the input modalities defined and organised into temporal sequences, we now present the inference engine of our framework: the MELTER architecture [25].

4. MELTER Building Blocks

This section presents the MELTER architecture [25], the first Transformer-based approach tailored for Multi-Fidelity environments. This novel Transformer-based model is explicitly designed to address the challenges of multi-fidelity data fusion, moving beyond standard sequence-to-sequence tasks to handle heterogeneous physical inputs. We elaborate on its key features, including the self-attention mechanism, masking strategy, and fidelity embedding. Previously benchmarked on synthetic Navier-Stokes datasets, the model is now tested against real-world forecasting scenarios. This transition introduces significant complexity, characterised by observational noise and unpredictable physical interactions not fully captured by theoretical models.

4.1. POD Spatial Encoding and Gated Recurrent Units Temporal Encoding

A critical design choice in our architecture concerns the handling of spatial dimensionality. Although all input data originate from a structured 2D grid ($\mathcal{G} \subset \mathbb{R}^{H \times W}$), we deliberately avoid the use of Convolutional Neural Networks (CNNs) for spatial feature extraction in favour of a flattened vector representation.

This decision is motivated by three factors:

1. **Domain Irregularity:** The region of interest is defined by a dynamic "active" Ice Mask that follows the irregular and fractal geometry of the Antarctic coastline and ice edge. Standard CNNs, which operate on rectangular grids, would require extensive zero-padding to process such sparse domains, leading to significant computational inefficiency and potential boundary artifacts.
2. **Global Receptive Field:** By flattening the spatial domain into a dense vector $x_t \in \mathbb{R}^{N_p}$ and feeding it directly into the Gated Recurrent Unit (GRU) encoders [8], we enable a global receptive field. Unlike CNNs, which are inductively biased towards local connectivity, the dense weight matrix of the GRU (W_r) can empirically learn long-range spatial correlations across the entire domain from the outset, effectively treating the spatial field as a fully connected graph.
3. **Optimal Variance Preservation:** While CNNs reduce dimensionality through localised, spatial downsampling, we opt for Proper Orthogonal Decomposition (POD) on the flattened state vector. POD evaluates the global covariance of the data to extract an optimal set of orthogonal basis functions. This data-driven, spectral truncation dramatically reduces the dimensionality of the system while ensuring that the most critical, large-scale physical dynamics are preserved for the subsequent recurrent layers.

The Gated Recurrent Unit is a simpler variant of the Long-Short Term Memory architecture, able to process time series and create a hidden state that acts like a memory. In our case, the hidden state is the embedding of a single fidelity level, before sending it into the Transformer Block.

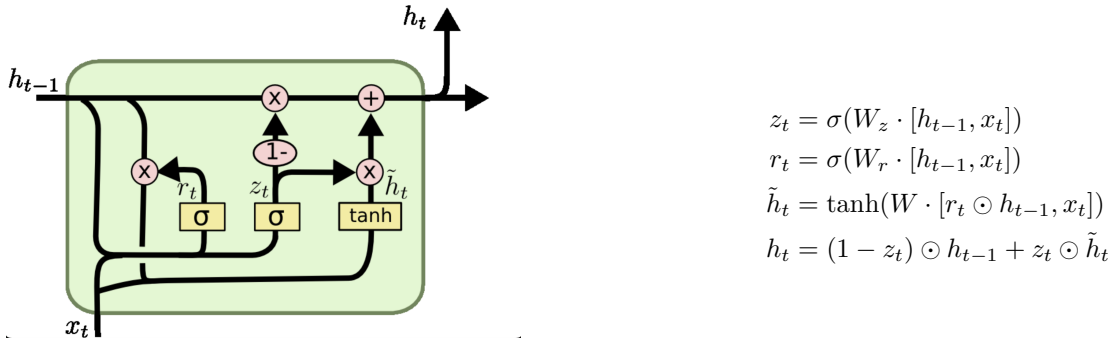


Figure 16: The Gated Recurrent Unit (left) and its update equations (right). σ denotes the Sigmoid function, and $\odot$ denotes the Hadamard product (element-wise multiplication).

From the above image taken from [26], we can understand that at each timestep t we have:

- z_t Update Gate: Determines how much of the past state (h_{t-1}) to retain and how much of the new content ($\tilde{h}_t$) to accept.
- r_t Reset Gate: Determines how much of the past information to forget before calculating the new candidate state.
- $\tilde{h}_t$ Candidate Hidden State: The temporary hidden state, calculated using the reset gate.
- h_t Output/Hidden State: The linear interpolation between the previous state and the candidate state, regulated by the update gate.

We utilise a dedicated GRU for each of the three fidelity levels. The generated hidden states h_t are fed into a Transformer, which aggregates the multi-fidelity data using an attention mechanism.

4.2. Transformer Blocks

The architectural backbone of this work relies on the original Transformer block, as first proposed in [35]. This module is fundamentally composed of two sequential operations: an Attention layer, designed to mix and contextualise information carried by each token across the sequence, followed by a Feed-Forward layer, which serves to inject non-linearities into the model representation. A distinguishing feature of the Transformer Block is its property of dimensionality preservation. The architecture is isometric regarding the input sequence structure: given an input sequence comprising L tokens of dimension D , the block produces an output sequence maintaining the identical $L \times D$ dimensionality. Internally, each sub-layer—both the Multi-Head Attention and the Feed-Forward Network—is encapsulated within a residual structure. The input to each sub-module is added to its output (a skip connection), followed by a Layer Normalisation step. This design, mathematically expressed as $\text{LayerNorm}(x + \text{Sublayer}(x))$, facilitates the flow of gradients during backpropagation, mitigating the vanishing gradient problem and ensuring stable convergence even in deeper configurations. In our application, we use one Transformer Block with two heads of attention. In [25] they scale both these hyper-parameters and compare the results. Regarding our case, we leave this for further investigation.

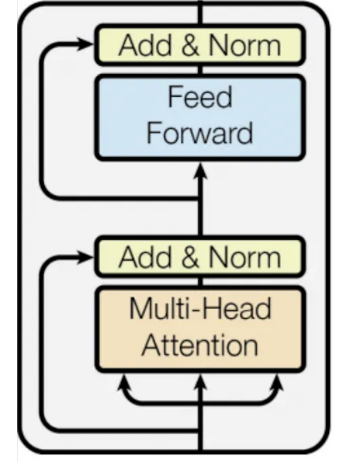


Figure 17: Transformer block from [35].

4.2.1 Multi-Head Attention

The Multi-Head Attention mechanism enhances the model’s expressivity by applying the attention operation purely in parallel across H independent heads.

To formalise the mechanism, let us consider a single head, denoted by the index h . The layer receives as input a sequence of feature vectors (tokens) x_i , where $i \in \{1, \dots, L\}$. Through linear projections, each token x_i is mapped into three distinct vector subspaces, namely Query, Key, and Value, via the learned weight matrices $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{D \times d_H}$, where d_H represents the dimension of the subspace.

These projected tokens are organised row-wise into three matrices: Q^h (Queries), K^h (Keys), and V^h (Values) $\in \mathbb{R}^{L \times d_H}$. Specifically, the i -th row of Q^h , denoted as q_i^h , contains the projection of the i -th token onto the Query subspace.

For every token i in the sequence, the mechanism computes a set of L attention coefficients c_{ij}^h (for $j \in \{1, \dots, L\}$), resulting in a total of L^2 interactions. These coefficients quantify the relevance of token j in updating the representation of token i . It is crucial to observe that this relationship is asymmetric, meaning generally $c_{ij}^h \neq c_{ji}^h$.

Mathematically, these coefficients are derived from the scalar product between the query projection q_i^h and the key projection k_j^h . To ensure numerical stability, these products are rescaled by a factor related to the subspace dimension d_H before being normalised via a Softmax function to produce the final attention scores.

The attention coefficients are computed using the Softmax function (Eq. 12) to ensure they are normalised and sum to one. Once these scores are obtained, the representation of the i -th token is updated as the linear combination of the value vectors v_j^h , weighted by the corresponding importance scores c_{ij}^h (Eq. 13).

$$c_{ij}^h = \frac{\exp(\frac{q_i^h \cdot k_j^h}{\sqrt{d_H}})}{\sum_{j=1}^L \exp(\frac{q_i^h \cdot k_j^h}{\sqrt{d_H}})}, \quad \forall i, j \in \{1, \dots, L\} \quad (12)$$

$$\text{Attention}^h(x_i) = \sum_{j=1}^L c_{ij}^h v_j^h \quad \forall i \in \{1, \dots, L\}. \quad (13)$$

These operations can be efficiently computed in matrix form as shown in Eq. 14. It is worth noting that the computational complexity of this mechanism is quadratic with respect to the sequence length ($O(L^2)$), as it requires computing L attention scores for each of the L tokens.

$$\text{Attention}(Q^h, K^h, V^h) = \text{softmax}\left(\frac{Q^h(K^h)^T}{\sqrt{d_H}}\right) V^h. \quad (14)$$

This process is repeated H times in parallel. To ensure that the computational cost does not increase with the number of heads, the dimension of the projection subspaces is set to $d_H = D/H$, where D is the model

embedding dimension. Finally, the outputs of the individual heads ($head_h \in \mathbb{R}^{L \times d_H}$) are concatenated along the feature dimension and linearly projected back to the original space $\mathbb{R}^{D \times D}$ using a projection matrix W^O (Eq. 15):

$$MultiheadAttention = Concat(head_1, \dots, head_h)W^O. \quad (15)$$

The output of the Multi-Head Attention block is therefore a sequence of contextualised vector representations, having the same length and embedding dimension as the input.

4.2.2 Feed Forward, Skip Connections and Normalisation

After the Attention mechanism has mixed the information across the sequence, a Feed Forward Network (FFN) is applied to inject non-linearity. This layer processes each token individually and is agnostic to the other tokens in the sequence. It consists of two linear transformations separated by a ReLU activation function, as described in Eq. 16:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2. \quad (16)$$

To facilitate the optimisation process and gradient propagation, Skip Connections are added between the input and output of both the Multi-Head Attention and the FeedForward layers. Additionally, Layer Normalisation [3] is applied to the output of each sub-layer to stabilise training and mitigate covariate shift.

4.3. Masking for Multi-Fidelity modelling

We re-purpose the masking mechanism, typically used in NLP for self-supervised training, to handle the availability of different fidelity levels. We frame the Multi-Fidelity modelling task as a ‘Masked Levels Prediction’ problem: given a set of potential fidelity levels $S_F = \{1, \dots, L\}$, the model must predict the target quantity of interest y utilising only the subset of currently available fidelities $\{x^{l_1}, \dots, x^{l_k}\} \subseteq S_F$.

To achieve this, we associate a binary mask vector $z \in \{0, 1\}^L$ to the input sequence, where $z_j = 1$ indicates that the j -th fidelity level is unavailable (masked), and so adopting a combinatorial strategy.

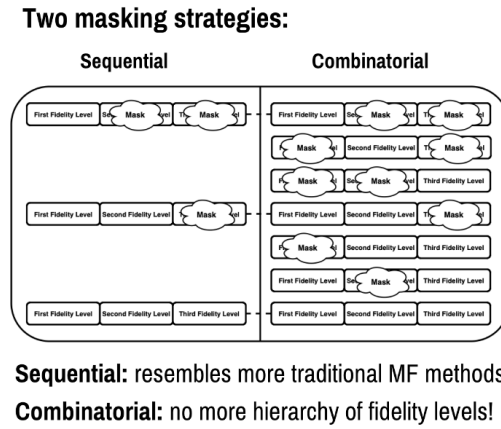


Figure 18: From [25], we use a combinatorial strategy during the training phase

The masking operation is implemented directly within the Multi-Head Attention mechanism by modifying the computation of the attention scores. Specifically, before applying the Softmax function, the scalar product between queries and keys corresponding to masked tokens is substituted with $-\infty$:

$$b_{ij}^h = \begin{cases} \frac{q_i^h \cdot k_j^h}{\sqrt{d_H}} & \text{if } z_j = 0 \quad (\text{available}) \\ -\infty & \text{if } z_j = 1 \quad (\text{missing}). \end{cases} \quad (17)$$

Consequently, when computing the normalised attention weights c_{ij}^h (Eq. 18), the contribution of the masked tokens becomes exactly zero:

$$c_{ij}^h = \frac{\exp(b_{ij}^h)}{\sum_{k=1}^L \exp(b_{ik}^h)} \approx 0 \quad \text{for all } j \text{ where } z_j = 1. \quad (18)$$

This technique ensures that the updated representation of the input sequence relies exclusively on the valid fidelity levels. The Transformer effectively learns to aggregate information dynamically, ignoring missing data sources without requiring architectural changes for different input configurations.

A key advantage of our tokenization strategy is that the Self-Attention mechanism scales quadratically with the number of modalities ($T = 3$), but remains invariant to the internal dimensionality of each token. Consequently, scaling the spatial resolution affects only the input/output projection layers, not the attention mechanism itself.

4.4. Decoder: Direct Mapping to Probabilistic Quantiles

For the decoding stage, we adopt a recurrent architecture designed to model the temporal evolution of forecast residuals rather than reconstructing the full physical state from scratch. Given that our objective is the online correction of the offline BOPDMD baseline, the decoder must capture the dynamical drift of the error terms. From an implementation perspective, the decoding module is explicitly structured into two sequential blocks. The first is a single-layer GRU that ingests a concatenated vector combining the latent embeddings extracted by the Transformer with auxiliary parametric variables. The GRU maintains a hidden state dimension of 128, a size specifically selected to provide sufficient representational capacity for the temporal dynamics while avoiding over-parameterisation. The resulting temporal features are then processed by the MLP projector to map the latent representations back to the required spatial dimension. To achieve this, the projector first expands the 128-dimensional hidden state into a wider 512-dimensional intermediate space. This expansion is stabilised via Layer Normalisation and activated by a non-linear Rectified Linear Unit (ReLU). Finally, following a Dropout layer for regularisation, a terminal linear layer projects the features to the final output dimension. We explicitly favour this recurrent decoding scheme over spatially-generative alternatives, such as Transposed Convolutional Networks or Implicit Neural Representations (INRs). While CNN-based decoders excel at reconstructing static, high-fidelity spatial patterns via upsampling, and INRs offer continuous resolution independence, both approaches incur significant computational overhead when scaling to high-dimensional manifolds. More importantly, the task of online residual correction is fundamentally a problem of modelling temporal drift rather than spatial generation. Since the offline BOPDMD baseline already provides the primary spatial structure of the Sea-Ice Concentration, the network’s role is to track how the forecasting error evolves and propagates over time. The GRU architecture is naturally aligned with this objective, capturing the sequential dependencies and decay rates of the residuals more effectively than static spatial decoders. Furthermore, it allows us to quantify the uncertainty of the error correction without the complexity of training generative convolutional layers to produce deterministic pixel grids.

Standard regression models typically output a single deterministic value (the mean), which fails to capture the stochastic nature of Sea-Ice dynamics. By training our decoder to predict multiple quantiles (e.g., $\tau = \{0.1, 0.5, 0.9\}$), we construct a probabilistic envelope around the forecast. The median quantile ($\tau = 0.5$) serves as the primary correction term to rectify the BOPDMD baseline. Meanwhile, a 90% confidence interval (e.g., $Q_{0.95} - Q_{0.05}$) provides a dynamic measure of aleatoric uncertainty [17].

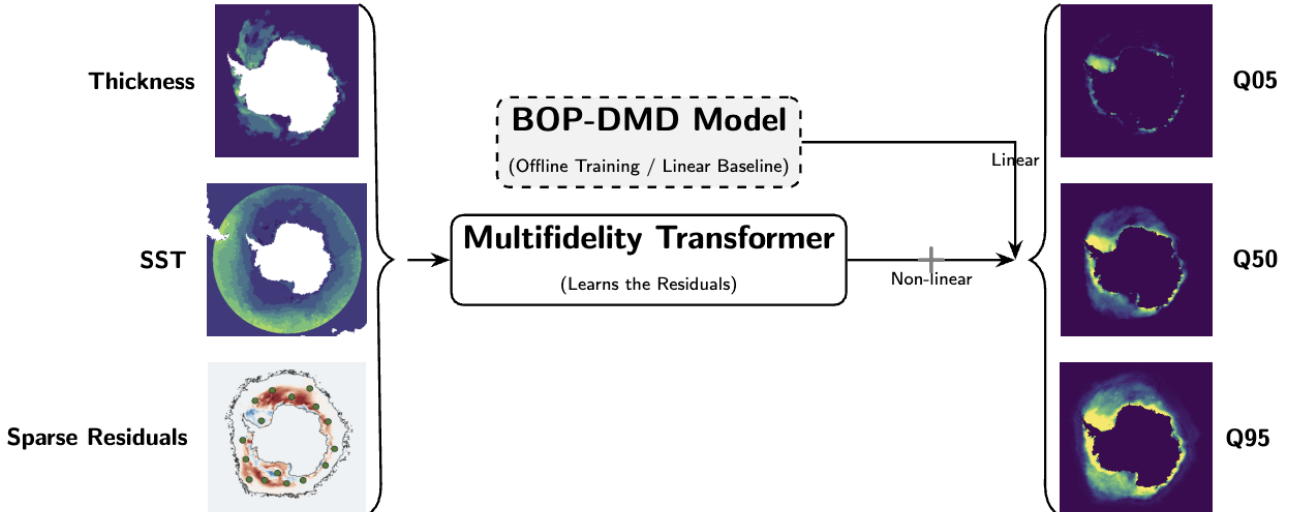


Figure 19: End-to-end inference pipeline of the COMET-QR framework. Three distinct low-fidelity input modalities (Thickness, SST, and Sparse Residual Sensors) are processed by the Multi-Fidelity Transformer to learn the non-linear residual dynamics. This data-driven correction is additively fused with the physics-based linear baseline generated offline by the BOPDMD model. The final output is a probabilistic envelope defined by the 5th, 50th, and 95th quantiles (Q05, Q50, Q95) of the predicted Sea-Ice Concentration, providing dynamic uncertainty quantification.

5. Residual Learning by Conformal Quantile Regression

Sea-Ice Concentration forecasting presents a unique challenge when combining model-driven predictions (such as Dynamic Mode Decomposition, DMD) with data-driven residual learning. While neural networks can learn to predict residuals that capture the discrepancy between baseline forecasts and observed values, ensuring rigorous uncertainty quantification requires careful consideration of how predictions are constructed.

In the following, we present a theoretically sound conformal prediction framework for residual-based forecasting. The central insight is subtle but critical: conformity scores must be computed on the final SIC predictions (Sea-Ice Concentration values in $[0, 1]$) rather than on intermediate residuals. This distinction matters because the reconstruction process involves nonlinear operations (adding the DMD baseline and clipping to probabilistic bounds) that invalidate naive approaches to uncertainty quantification.

Our forecasting system operates over a spatial domain consisting of p pixels in an active Antarctic region, observed over T daily time steps. We use $i \in \{1, \dots, p\}$ to index spatial locations and $t \in \{1, \dots, T\}$ to index time (measured in days). This framework naturally accommodates the high-dimensional nature of Sea-Ice forecasting, where thousands of spatial locations must be predicted simultaneously while respecting temporal dependencies. We underline again that our approach builds upon a baseline forecasting system (DMD) by learning to predict residuals, the difference between the offline forecast and the true observed value. Let us formally define the key quantities.

For each pixel i and time t , we work with four fundamental quantities:

$$\begin{cases} y_{i,t}^{\text{DMD}} \in [0, 1], & \text{(DMD baseline prediction)} \\ n_{i,t} \in [0, 1], & \text{(True SIC from observations)} \\ r_{i,t} = n_{i,t} - y_{i,t}^{\text{DMD}}, & \text{(True residual)} \\ \hat{r}_{i,t} \in [-1, 1]. & \text{(Predicted residual from model)} \end{cases} \quad (19)$$

The baseline $y_{i,t}^{\text{DMD}}$ represents our physics-informed prior knowledge, constrained to the probabilistic range $[0, 1]$. The true Sea-Ice Concentration $n_{i,t}$ is what we ultimately wish to predict. The true residual $r_{i,t}$ captures the systematic error of the baseline, and our neural network learns to predict this residual, denoted $\hat{r}_{i,t}$. Following standard machine learning practice, after a robust scaler preprocessing on all input levels, we partition our 31 years of Antarctic Sea-Ice data into three disjoint temporal blocks. This temporal splitting respects the time-series nature of our data and avoids information leakage:

$$\begin{aligned} \mathcal{D}_{\text{train}} &= \{(x_t, r_{i,t}) : t \in \mathcal{T}_{\text{train}}, \forall i\} & \mathcal{T}_{\text{train}}: & 8,395 \text{ days (Model learning)} \\ \mathcal{D}_{\text{val}} &= \{(x_t, r_{i,t}) : t \in \mathcal{T}_{\text{val}}, \forall i\} & \mathcal{T}_{\text{val}}: & 1,460 \text{ days (Conformal calibration)} \\ \mathcal{D}_{\text{test}} &= \{(x_t, r_{i,t}) : t \in \mathcal{T}_{\text{test}}, \forall i\} & \mathcal{T}_{\text{test}}: & 1,460 \text{ days (Final evaluation)} \end{aligned}$$

Here, x_t represents the multi-modal input features (Sea-Ice Thickness POD, Sea Surface Temperature POD, and sensor measurements) available at time t . The validation set plays a special role in our framework: it is never used for training the neural network, but instead provides the conformity scores that calibrate our prediction intervals. This separation is essential for the theoretical guarantees of conformal prediction.

5.1. Learning Uncertainty Through Quantiles

Rather than learning a single point prediction for each residual, we train our model to predict multiple quantiles of the residual distribution. This distributional approach naturally captures uncertainty: wider gaps between quantiles indicate greater uncertainty, while narrow gaps suggest confident predictions.

5.1.1 Residual Normalisation for Numerical Stability

Neural networks train more effectively when inputs and outputs have similar scales. The residuals are theoretically bounded in $[-1, 1]$ (since $r = n - y^{\text{DMD}}$ where both terms are in $[0, 1]$). For numerical stability, residuals are normalised during training:

$$\mu_r = \frac{1}{|\mathcal{T}_{\text{train}}|p} \sum_{t,i} r_{i,t}, \quad \sigma_r = \sqrt{\frac{1}{|\mathcal{T}_{\text{train}}|p} \sum_{t,i} (r_{i,t} - \mu_r)^2}, \quad \tilde{r}_{i,t} = \frac{r_{i,t} - \mu_r}{\sigma_r}. \quad (20)$$

The normalisation parameters (μ_r and σ_r) are computed once on the training set and then held fixed for validation and testing. The model works in this normalised space, predicting:

$$f_{\theta}(x_t) \rightarrow (\tilde{r}_{i,t}^{(0.05)}, \tilde{r}_{i,t}^{(0.50)}, \tilde{r}_{i,t}^{(0.95)}). \quad (21)$$

The input x_t aggregates information from three fidelity levels allowing the model to adaptively weight different information sources. The output predicts the 5th percentile (lower bound), median (central tendency), and 95th percentile (upper bound) of the residual distribution at each of the p spatial locations. To recover physical residuals, we simply invert the normalisation, that is, $\hat{r}_{i,t}^{(q)} = \tilde{r}_{i,t}^{(q)} \cdot \sigma_r + \mu_r$.

5.1.2 The Pinball Loss Function

To train the model for quantile prediction, we employ the pinball loss (also called quantile loss), which asymmetrically penalises over- and under-predictions according to the target quantile level:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{T}_{\text{train}}|} \sum_{t \in \mathcal{T}_{\text{train}}} \sum_{i=1}^p \sum_{q \in \{0.05, 0.50, 0.95\}} \rho_q(\tilde{r}_{i,t} - \tilde{r}_{i,t}^{(q)}) \quad , \quad \text{where} \quad \rho_q(u) = \begin{cases} q \cdot u & u \geq 0 \\ (q-1) \cdot u & u < 0. \end{cases} \quad (22)$$

5.2. Conformal Calibration: The Coverage Guarantee Problem

After training, our model produces quantile predictions, but these quantiles carry no formal coverage guarantees. For example, the 90th percentile is merely what the model learned to call "90th percentile" by looking only at the training data, and so it may systematically over- or under-cover the true values. Conformal prediction provides a rigorous, distribution-free method to correct these predictions and obtain valid coverage guarantees. The theoretical foundation of our approach comes from conformal prediction theory, with no assumptions on the residual distribution, as we have mentioned before. Assuming that the validation and test samples, $\{(x_t, n_{i,t})\}_{t \in \mathcal{T}_{\text{val}} \cup \mathcal{T}_{\text{test}}}$, are exchangeable, then the calibrated prediction intervals constructed using the method below satisfy:

$$\mathbb{P} \left(n_{i,t} \in [\hat{n}_{i,t}^{\text{low}}, \hat{n}_{i,t}^{\text{high}}] \right) \geq 1 - \alpha \quad \text{for any test sample } t \in \mathcal{T}_{\text{test}}. \quad (23)$$

The calibration process begins by transforming model outputs into physical SIC predictions. First, we denormalise the residual quantiles predicted by the model. For each validation sample $t \in \mathcal{T}_{\text{val}}$ and pixel i :

$$\hat{r}_{i,t}^{(0.05)} = \tilde{r}_{i,t}^{(0.05)} \cdot \sigma_r + \mu_r, \quad \hat{r}_{i,t}^{(0.95)} = \tilde{r}_{i,t}^{(0.95)} \cdot \sigma_r + \mu_r. \quad (24)$$

These denormalised residuals are constrained to $[-1, 1]$. We combine them with the DMD baseline and clip to the valid SIC range $[0, 1]$:

$$\hat{n}_{i,t}^{\text{low,uncal}} = \text{clip} \left(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.05)}, 0, 1 \right), \quad \hat{n}_{i,t}^{\text{high,uncal}} = \text{clip} \left(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.95)}, 0, 1 \right). \quad (25)$$

Here, $\text{clip}(x, a, b) = \max(a, \min(x, b))$ enforces the constraint $x \in [a, b]$. We call these "uncalibrated" predictions because they haven't yet been adjusted by conformal correction.

5.2.1 Measuring Miscoverage: The Conformity Score

The heart of conformal prediction lies in the conformity score, a measure of how badly our uncalibrated intervals fail to cover the true value. The key insight is to compute this score on the final SIC predictions, not on intermediate residuals. For each validation spatiotemporal sample (i, t) , the conformity score quantifies miscoverage:

$$E_{i,t} = \max \left(\hat{n}_{i,t}^{\text{low,uncal}} - n_{i,t}, n_{i,t} - \hat{n}_{i,t}^{\text{high,uncal}} \right). \quad (26)$$

This definition captures both directions of miscoverage. When the interval successfully covers the true value (i.e., $\hat{n}_{i,t}^{\text{low,uncal}} \leq n_{i,t} \leq \hat{n}_{i,t}^{\text{high,uncal}}$), both terms are non-positive, but we still record the fact of being close to the boundary, thus preserving information about how close we came to failure.

In standard conformal regression, the model predicts a single point (the mean), $\hat{y}$, and so the standard conformity score is simply the absolute residual: $E = |y - \hat{y}|$, because one just wants to know how far the true value is from your single prediction. However, in CQR, the model predicts an interval (a lower bound and an upper bound) rather than a single point. The conformity score's job is to measure how far outside the predicted interval the true value falls. By taking the 90th percentile of these max scores across the calibration set, we find the exact constant (the Q-score) needed to expand or shrink the intervals. If the 90th percentile of the max scores is positive, it means the base intervals are too narrow, and we must add that positive buffer to widen them. If the 90th percentile is negative, it means our network is overly conservative, and we can safely shrink the intervals while still guaranteeing 90% coverage.

5.2.2 SIC-Based Calibration is Theoretically Necessary

Why do we compute scores on SIC, rather than residuals? Although residuals are bounded in $[-1, 1]$, the clipping operation $\text{clip}(\cdot, 0, 1)$ is nonlinear, meaning residual coverage does not imply SIC coverage. By computing conformity scores on the actual predicted probabilities in $[0, 1]$, we ensure our coverage guarantees apply to the quantities we ultimately report to users. A natural first attempt at conformal calibration might compute conformity scores on residuals, before reconstruction. This approach is appealingly simple: calibrate in the space where the model makes predictions (residual space), then transform to SIC space afterward. Unfortunately, this does not guarantee coverage of the final SIC predictions, even with residuals bounded in $[-1, 1]$. The problem lies in the $\text{clip}(\cdot, 0, 1)$ operation. Consider a concrete example: Suppose $y_{i,t}^{\text{DMD}} = 0.06$ (small baseline SIC), true residual: $r_{i,t} = -0.01$ (valid in $[-1, 1]$) and predicted residual lower bound: $\hat{r}_{i,t}^{(0.05)} = -0.16$ (valid in $[-1, 1]$). In residual space, the error is $|\hat{r} - r| = |-0.16 - (-0.01)| = 0.15$. However, in SIC space, the predicted SIC is clipped to 0, and the true SIC is 0.05. The error is only $|0 - 0.05| = 0.05$. The conformity score in residual space drastically overestimates the error in SIC space due to the boundary effects. Our method avoids this pitfall by measuring miscoverage in the actual prediction space, the space of SIC values in $[0, 1]$. The resulting framework combines the best of two-phase forecasting: DMD provides a physics-informed baseline, neural networks learn complex residual patterns (within $[-1, 1]$), and conformal calibration provides distribution-free coverage guarantees.

5.2.3 Computing the Q-Scores

It is straightforward to understand that not all locations and times require the same calibration correction. Coastal regions may have higher uncertainty than interior; summer dynamics may differ from winter. To capture this heterogeneity, we stratify our calibration spatially (per pixel) and temporally (per season). We define four Antarctic seasons based on climatological patterns:

Seasonal Stratification

Season (s)	Days (DOY)	Months	Dynamics
s=0 (Summer)	335–365, 1–59	Dec–Feb	Max. melt
s=1 (Autumn)	60–151	Mar–May	Freeze-up
s=2 (Winter)	152–243	Jun–Aug	Max. extent
s=3 (Spring)	244–334	Sep–Nov	Melt onset

Table 2: Definition of the four Antarctic seasons used for temporal stratification.

This stratification allows calibration to adapt to seasonal predictability: if the model is more uncertain during transitional seasons (autumn and spring), the Q-scores for crossed periods will be larger, widening the intervals appropriately. For each combination of pixel i and season s , we collect all conformity scores from validation samples in that stratum:

$$\mathcal{E}_{i,s} = \{E_{i,t} : t \in \mathcal{T}_{\text{val}} \text{ and } \text{season}(t) = s\}. \quad (27)$$

The Q-score is then defined as the $(1 - \alpha)$ -quantile of these scores using the “higher” method for computing quantiles:

$$Q_{i,s} = \text{Quantile}_{1-\alpha}(\mathcal{E}_{i,s}) = \inf\{q : \mathbb{P}(E_{i,t} \leq q) \geq 1 - \alpha, t \in \mathcal{T}_{\text{val}}, \text{season}(t) = s\}. \quad (28)$$

By enabling pixel-wise calibration, the model generates distinct Q-scores for each of the 43,499 valid pixels in the ice mask, effectively accounting for spatial heterogeneity where coastal regions often exhibit higher prediction uncertainty compared to the stable interior. This spatial granularity is coupled with temporal stratification, which divides the calibration process into four distinct Antarctic seasons (Summer, Autumn, Winter, and Spring) to capture cyclic variations in ice dynamics. Consequently, by computing residuals on the validation set for every pixel-season combination at a 90% confidence level, the system derives a total of 173,996 unique Q-scores, ensuring that the resulting confidence intervals are locally adaptive both in space and time. To summarise everything, we detail the whole procedure in Algorithm 2.

Algorithm 2 Conformal Calibration (Validation Phase)

```
1: Input: Model  $f_\theta$ , validation set  $\mathcal{D}_{\text{val}}$ , DMD baseline  $\{y_{i,t}^{\text{DMD}}\}$ , true SIC  $\{n_{i,t}\}$ 
2: Output: Q-scores  $\{Q_{i,s}\}_{i=1,\dots,p;s=0,\dots,3}$ 
3: for  $t \in \mathcal{T}_{\text{val}}$  do
4:   Predict:  $(\tilde{r}_{i,t}^{(0.05)}, \tilde{r}_{i,t}^{(0.95)}) \leftarrow f_\theta(x_t)$ 
5:   Denormalise:  $\hat{r}_{i,t}^{(q)} = \tilde{r}_{i,t}^{(q)} \cdot \sigma_r + \mu_r$  for  $q \in \{0.05, 0.95\}$ 
6:   Reconstruct:  $\hat{n}_{i,t}^{\text{low}} = \text{clip}(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.05)}, 0, 1)$ 
7:                $\hat{n}_{i,t}^{\text{high}} = \text{clip}(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.95)}, 0, 1)$ 
8:   Compute:  $E_{i,t} = \max(\hat{n}_{i,t}^{\text{low}} - n_{i,t}, n_{i,t} - \hat{n}_{i,t}^{\text{high}})$ 
9:   Store:  $E_{i,t}$  in set  $\mathcal{E}_{i,s}$  where  $s = \text{season}(t)$ 
10: end for
11: for  $i = 1$  to  $p$  do
12:   for  $s = 0$  to  $3$  do
13:      $Q_{i,s} = \text{Quantile}_{0.90}(\mathcal{E}_{i,s})$ 
14:   end for
15: end for
16: return  $\{Q_{i,s}\}$ 
```

5.3. Test Prediction (Application of Q-Scores)

Once the calibration is complete, prediction intervals for new test points are constructed by adjusting the model's raw predicted quantiles, specifically the 5th and 95th percentiles—using the computed Q-score specific to the relevant pixel and season. With Q-scores computed from the validation set, we can now produce conformalised predictions on the test set. These conformalised bounds are indeed added to the DMD baseline to derive the final Sea-Ice Concentration (SIC) predictions, which are strictly clipped to the physical range of $[0, 1]$ to ensure realism. Theoretically, provided that the calibration and test data are exchangeable, this process guarantees that the true value falls within the calculated interval with a probability of at least $1 - \alpha$; for an α of 0.1, this ensures a minimum 90% coverage rate on the test set.

For each test sample $t \in \mathcal{T}_{\text{test}}$ and pixel i , we follow the following pipeline:

Algorithm 3 Conformal Prediction (Test Phase)

```
1: Input: Model  $f_\theta$ , test sample  $x_t$ , DMD baseline  $y_{i,t}^{\text{DMD}}$ , Q-scores  $\{Q_{i,s}\}$ 
2: Output: Prediction interval  $[\hat{n}_{i,t}^{\text{low}}, \hat{n}_{i,t}^{\text{high}}]$ 
3: Predict:  $(\tilde{r}_{i,t}^{(0.05)}, \tilde{r}_{i,t}^{(0.50)}, \tilde{r}_{i,t}^{(0.95)}) \leftarrow f_\theta(x_t)$ 
4: Denormalise:  $\hat{r}_{i,t}^{(q)} = \tilde{r}_{i,t}^{(q)} \cdot \sigma_r + \mu_r$  for  $q \in \{0.05, 0.50, 0.95\}$ 
5: Reconstruct uncalibrated SIC (ensuring  $y^{\text{DMD}} + r \in [0, 1]$ ):
6:    $\hat{n}_{i,t}^{\text{low,uncal}} = \text{clip}(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.05)}, 0, 1)$ 
7:    $\hat{n}_{i,t}^{\text{med}} = \text{clip}(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.50)}, 0, 1)$ 
8:    $\hat{n}_{i,t}^{\text{high,uncal}} = \text{clip}(y_{i,t}^{\text{DMD}} + \hat{r}_{i,t}^{(0.95)}, 0, 1)$ 
9: Determine season:  $s = \text{season}(t)$ 
10: Apply Q-score correction:
```

$$\hat{n}_{i,t}^{\text{low}} = \text{clip}(\hat{n}_{i,t}^{\text{low,uncal}} - Q_{i,s}, 0, 1) \quad (29)$$

$$\hat{n}_{i,t}^{\text{high}} = \text{clip}(\hat{n}_{i,t}^{\text{high,uncal}} + Q_{i,s}, 0, 1) \quad (30)$$

```
11: return  $[\hat{n}_{i,t}^{\text{low}}, \hat{n}_{i,t}^{\text{high}}]$  with median  $\hat{n}_{i,t}^{\text{med}}$ 
```

Having formally defined the framework and computed the conformal Sea-Ice Concentration (SIC) predictions, the following section presents a comprehensive evaluation of the COMET-QR architecture.

6. Results

In this section, we provide all the important results of our analysis, evaluating COMET-QR for Antarctic Sea-Ice prediction. In particular, we provide a detailed explanation of the test phase, which evaluates model performance in 7 different data availability scenarios, generates conformal prediction intervals, compares against two baseline models (BOPDMD and Climatology), and produces comprehensive visualisations for scientific interpretation. The empirical results, summarised in Table 3 and visualised in Figure 21, validate the core hypothesis of the COMET-QR architecture: the fusion of multi-fidelity data streams significantly enhances predictive accuracy over single-source baselines.

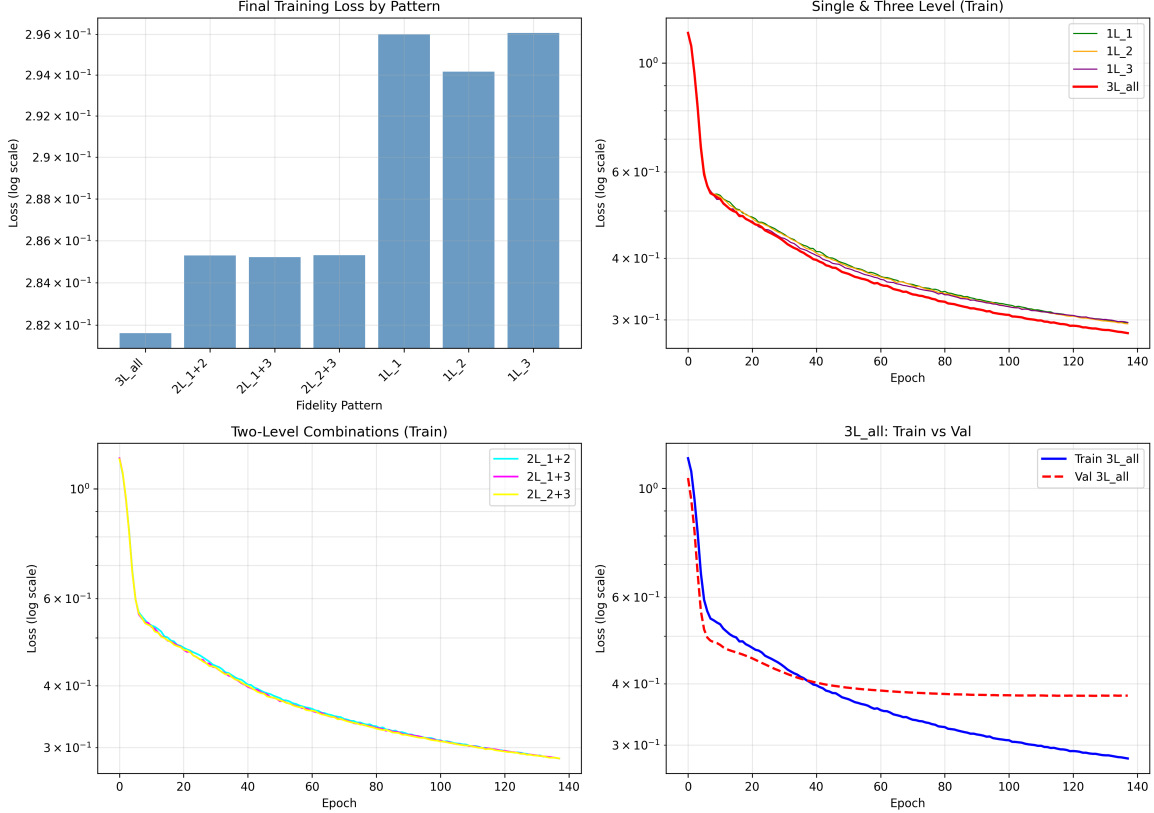


Figure 20: Overview of the individual multi-fidelity input scenarios, $1L_1$ is SIT, $1L_2$ is SST, $1L_3$ is the sparse residual field. During the training phase, specific fidelity levels are randomly masked to promote architectural robustness. A clear decreasing trend in the pinball loss is observed as the number of available input modalities increases

6.1. Predictive Pipeline and Test Metrics

To ensure a rigorous and unbiased assessment of the model’s performance, the predictive pipeline employs a strict chronological data split of 23, 4, and 4 years. Specifically, the model is trained on an extensive 23-year period (8,395 days from 1993 to 2015), validated over four years (1,460 days from 2016 to 2019), and finally evaluated on a distinct four-year test set (1,460 days from 2020 to 2023). This sequential division is critical for preventing data leakage, guaranteeing that the model remains entirely blind to future data during the training phase. The model, comprising approximately 67.5 million parameters, with batch size = 16, converged smoothly over 145 epochs without exhibiting gradient vanishing or exploding. The learning rate scheduler dynamically decayed the learning rate from a peak of 2×10^{-5} down to 5×10^{-6} during the final epochs, allowing the validation loss to steadily decrease from an initial 1.049 to a final, optimised minimum of 0.3827. By reserving four full years for the test set, the evaluation is highly robust and fully capable of capturing natural interannual variability. Furthermore, to maintain strict separation and avoid any look-ahead bias, the climatological baseline is computed exclusively using data from the training years.

To further assess operational robustness, the model is evaluated across seven distinct scenarios that simulate varying levels of data availability, ranging from a comprehensive 3L configuration, which utilises SIT, SST, and Sparse Residual Sensors, to restrictive single modality setups. As we have seen, this is implemented via a strict

mask forcing technique applied prior to the forward pass, which effectively blocks specific input channels to rigorously simulate low fidelity input failures. To rigorously quantify the model’s accuracy and reliability, the evaluation framework employs a comprehensive suite of performance metrics, starting with the Mean Absolute Error (MAE).

$$\text{MAE} = \frac{1}{N_{\text{test}} \cdot |\mathcal{M}_{\text{region}}|} \sum_{t=1}^{N_{\text{test}}} \sum_{i \in \mathcal{M}_{\text{region}}} |n_{i,t} - \hat{n}_{i,t}|. \quad (31)$$

This standard metric calculates the average absolute difference between the median predictions and the ground truth Sea-Ice Concentration (SIC) on the active region $\mathcal{M}_{\text{region}}$, providing a robust and physically interpretable measure of error in fractional units that is less sensitive to outliers than mean squared error. The MAE is computed not only for the model’s predictions but also for the Dynamic Mode Decomposition (DMD) baseline and a climatological reference, which is commonly used to establish the lower bound of forecast skill. This statistical reference is derived by averaging the daily Sea-Ice Concentration across the 23-year training period (1993–2015) for each day of the year. As a result, the climatology baseline captures the mean seasonal cycle but lacks any year-to-year variability, representing the simplest possible prediction one could make without a predictive model. Finally, the value added by the residual learning process is quantified via an Improvement Metric, which expresses the percentage reduction in error relative to the BOPDMD baseline, thereby directly measuring the model’s ability to correct systematic biases in the underlying physical decomposition.

Scenario	MAE model	Imp. DMD	Imp. Clim	Cov.	Width
3L_all	0.0746	21.95%	7.63%	0.895	0.248
2L_noSensors	0.0797	16.55%	1.24%	0.870	0.242
2L_noThickness	0.0751	21.42%	7.00%	0.894	0.248
2L_noSST	0.0750	21.52%	7.12%	0.900	0.251
1L_Sensors	0.0794	16.84%	1.58%	0.886	0.253
1L_SST	0.0822	13.92%	-1.87%	0.859	0.241
1L_Thickness	0.0818	14.36%	-1.36%	0.874	0.246

Table 3: Performance comparison of all data availability scenarios ($MAE_{\text{DMD}} = 0.0955$, $MAE_{\text{CLIM}} = 0.0807$)

Complementing point accuracy, the framework assesses the reliability of uncertainty quantification through empirical Coverage:

$$\text{Coverage} = \frac{1}{N_{\text{test}} \cdot |\mathcal{M}_{\text{region}}|} \sum_{t=1}^{N_{\text{test}}} \sum_{i \in \mathcal{M}_{\text{region}}} \mathbb{I} \left\{ \hat{n}_{i,t}^{\text{low}} \leq n_{i,t} \leq \hat{n}_{i,t}^{\text{high}} \right\}. \quad (32)$$

It measures the fraction of test points where the true SIC value falls within the predicted intervals. Targeted at a 90% confidence level ($\alpha = 0.1$), this metric indicates whether the model is well-calibrated (=90%), under-confident ($\ll 90\%$), or overconservative ($\gg 90\%$), and is analysed both at the aggregate scenario level and across individual Antarctic seasons: To contextualise coverage, the Interval Width is also tracked:

$$\text{Width} = \frac{1}{N_{\text{test}} \cdot |\mathcal{M}_{\text{region}}|} \sum_{t=1}^{N_{\text{test}}} \sum_{i \in \mathcal{M}_{\text{region}}} (\hat{n}_{i,t}^{\text{high}} - \hat{n}_{i,t}^{\text{low}}). \quad (33)$$

This metric represents the average size of the prediction intervals, reflecting the model’s certainty and highlighting the critical trade-off where wider intervals increase coverage but reduce operational utility (narrow intervals with proper coverage indicate efficient uncertainty quantification).

Moving to the spatial evaluation, the model’s output is visualised through high-resolution maps using an Antarctic polar stereographic projection. For individual test days (e.g., day 180), Figure 21 displays the lower bound (5th percentile), median prediction, upper bound (95th percentile), ground truth, BOPDMD baseline, and uncertainty. These maps allow for a direct physical interpretation of the results: narrow intervals indicate high model confidence, while wider intervals, typically found near the dynamic ice edge, reflect increased uncertainty.

Comparing the median prediction against the ground truth and the BOPDMD surrogate provides an immediate visual assessment of prediction accuracy and the specific value added by the residual learning component. A clear hierarchical performance pattern emerges based on data availability, as the complete 3L configuration (incorporating SIT, SST, and Sparse Residual Sensors) achieves the best overall performance with a Mean Absolute Error (MAE) of 0.0746. This represents a substantial 21.95% improvement over the physics-based BOPDMD baseline and a 7.63% improvement over the Climatology baseline. The ablation scenarios reveal the asymmetric importance of the different input modalities.

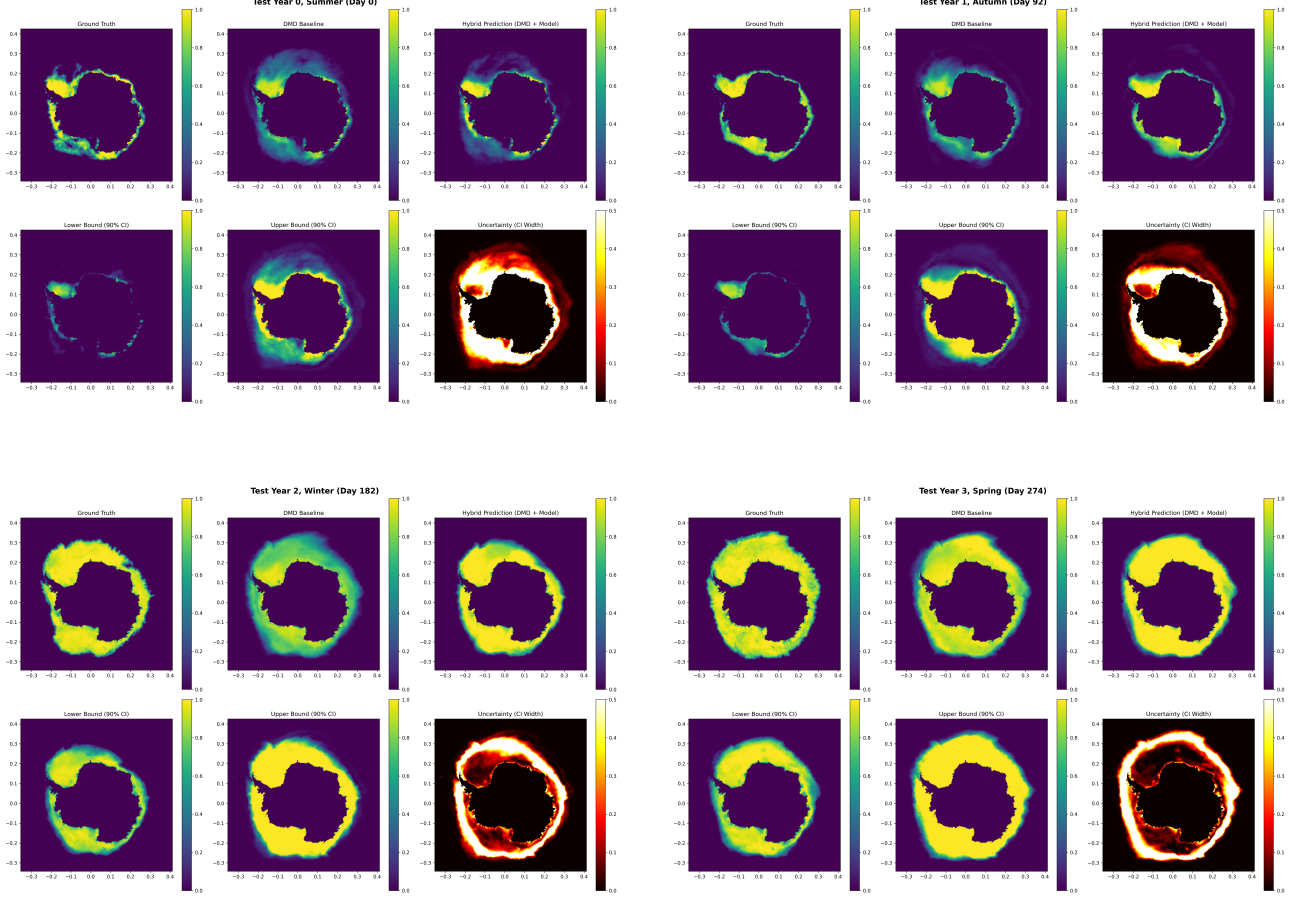


Figure 21: Seasonal predictions comparison over years, using all 3 modalities

When Sparse Residual Sensors are removed, the model’s improvement over Climatology plummets to a mere 1.24%. Furthermore, the single-modality configurations relying solely on global POD features (either SIT or SST) actually perform worse than the Climatology baseline (exhibiting negative improvements of -1.87% and -1.36% , respectively). This stark contrast underscores a fundamental characteristic of the residual learning task: while global thermodynamic proxies (SST, SIT) provide valuable structural context, they are insufficient on their own to correct the high-frequency, localised non-linearities missed by the DMD. The sparse residual field acts as a vital real-time error map, proving indispensable for driving the Transformer’s attention mechanism toward meaningful localised corrections.

Beyond point-estimate accuracy, the primary objective of COMET-QR is to provide reliable Uncertainty Quantification (UQ) for operational risk assessment, and we claim that the Conformal Quantile Regression successfully achieves this goal. Across the test set, the 3L scenario yields an empirical coverage of 89.5%, tightly approximating the theoretical target of 90% ($\alpha = 0.1$). This is achieved with an average Confidence Interval (CI) width of just 0.248 (in fractional SIC units), indicating that the model is not artificially inflating the intervals to achieve coverage, but is instead generating sharp, highly informative bounds. The spatiotemporal distribution of this uncertainty, visualised in Figure 22, perfectly aligns with the physical reality of the Antarctic cryosphere. The Confidence Interval widths are not uniformly distributed; instead, they are heavily concentrated along the Marginal Ice Zone (MIZ), the highly dynamic boundary between the solid ice pack and the open ocean. This is physically consistent, as the MIZ is subject to chaotic oceanic forcing, wave-ice interactions, and rapid thermodynamic melting/freezing. In contrast, the model exhibits extremely high confidence (narrow CI widths) in the deep interior of the Weddell and Ross Seas, where the ice pack remains structurally stable throughout

the winter and spring. Furthermore, the temporal analysis reveals that the model dynamically adjusts its uncertainty bounds in response to the seasonal cycle, as the overall uncertainty (as measured by the average CI width) reaches its minimum during the summer melt season (December–February). This contraction occurs because as the seasonal ice melts away, the remaining ice pack is largely confined to the stable, thick, multi-year ice in the Weddell and Ross Seas, while the highly volatile Marginal Ice Zone shrinks drastically, leading to higher overall predictability. Conversely, the uncertainty bounds steadily expand during the autumn freeze-up (March–May) and peak during the winter and spring months.

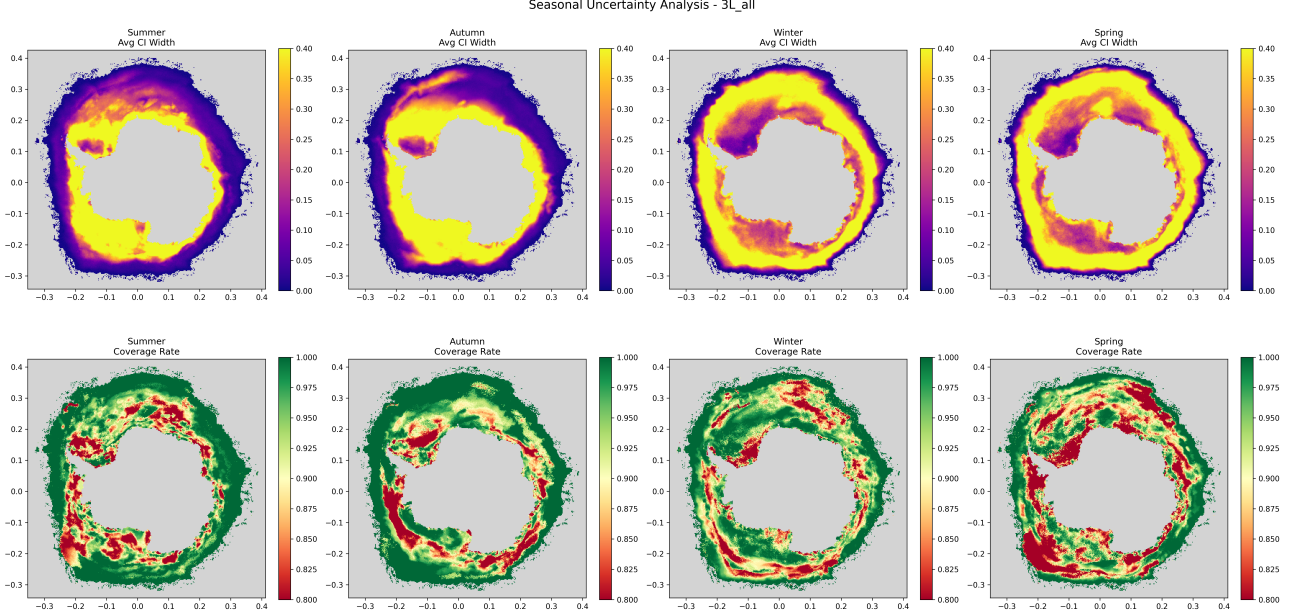


Figure 22: During the colder periods, the sea-ice extent reaches its maximum, creating a vast, highly dynamic perimeter that is continuously subjected to chaotic, large-scale oceanic and atmospheric forcing, thereby inherently increasing the forecasting uncertainty.

6.2. IIEE, Binary Accuracy and SIE

To rigorously evaluate the predictive capabilities of the COMET-QR framework and ensure comparability with state-of-the-art deep learning baselines such as IceNet [1], we adopt a standardised set of deterministic and spatial metrics. We recall again that $n_{i,t} \in [0, 1]$ denote the true Sea-Ice Concentration (SIC) at pixel $i \in \mathcal{M}_{\text{region}}$ and timestep $t \in \{1, \dots, N_{\text{test}}\}$, where $\mathcal{M}_{\text{region}}$ defines the active oceanic domain comprising 43,736 pixels, yielding a total active area of approximately $27.33 \times 10^6 \text{ km}^2$. The test period spans $N_{\text{test}} = 1,460$ days (2020–2023). Our architecture outputs a median prediction $\hat{n}_{i,t}$ alongside the 90% conformalised prediction interval boundaries $\hat{n}_{i,t}^{\text{low}}$ and $\hat{n}_{i,t}^{\text{high}}$.

To specifically quantify the spatial misplacement of the ice edge, we utilise the Integrated Ice-Edge Error (IIEE). The IIEE disaggregates spatial forecasting errors by independently measuring areas of overestimation and underestimation relative to the standard 15% SIC threshold ($\tau = 0.15$):

$$\text{IIEE} = \frac{A}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \left(\sum_{i \in \mathcal{M}_{\text{region}}} \mathbb{I}[\hat{n}_{i,t} > \tau \wedge n_{i,t} \leq \tau] + \sum_{i \in \mathcal{M}_{\text{region}}} \mathbb{I}[\hat{n}_{i,t} \leq \tau \wedge n_{i,t} > \tau] \right) \quad (34)$$

where $A = 625 \text{ km}^2$ represents the area of a single 25 km grid cell. The IIEE is expressed in km^2 per timestep. The first term inside the parenthesis captures overestimation (predicting ice where open water is observed), whereas the second term captures underestimation (failing to predict existing ice).

To provide a highly interpretable measure of spatial skill, the IIEE is normalised against the total active domain to compute the Binary Accuracy. This metric represents the proportion of the Antarctic domain where the binary ice/no-ice classification is correctly predicted:

$$\text{Binary Accuracy} = \left(1 - \frac{\text{IIEE}}{|\mathcal{M}_{\text{region}}| \cdot A} \right) \times 100\% \quad (35)$$

Values approaching 100% indicate negligible spatial misplacement of the ice pack. The full three-fidelity configuration (3L) demonstrates exceptional spatial predictive skill, achieving an average IIEE of $2,165,402 \text{ km}^2$

per timestep across the massive active domain. This corresponds to a remarkable Binary Accuracy of 92.08%, indicating that the model successfully resolves the highly chaotic ice/no-ice boundary for the vast majority of the Southern Ocean. Decomposing the IIEE provides deep physical insight into the model’s behaviour. The total error is driven by an average overestimation of 1,473,678 km² and an underestimation of 691,725 km². This asymmetric error distribution reveals a conservative predictive tendency: the framework slightly favours the retention or formation of ice over aggressive melting. The ablation studies (Figure 24) confirm the necessity of the hybrid architecture. The two-fidelity variants show intermediate performance with IIEEs ranging from 2,192,794 to 2,292,314 km², while single-fidelity baselines exhibit substantially higher spatial errors between 2,355,064 and 2,774,243 km². The progressive reduction in IIEE, from single to three fidelity architectures, confirms that each modality contributes unique, non-redundant information for resolving the ice edge.

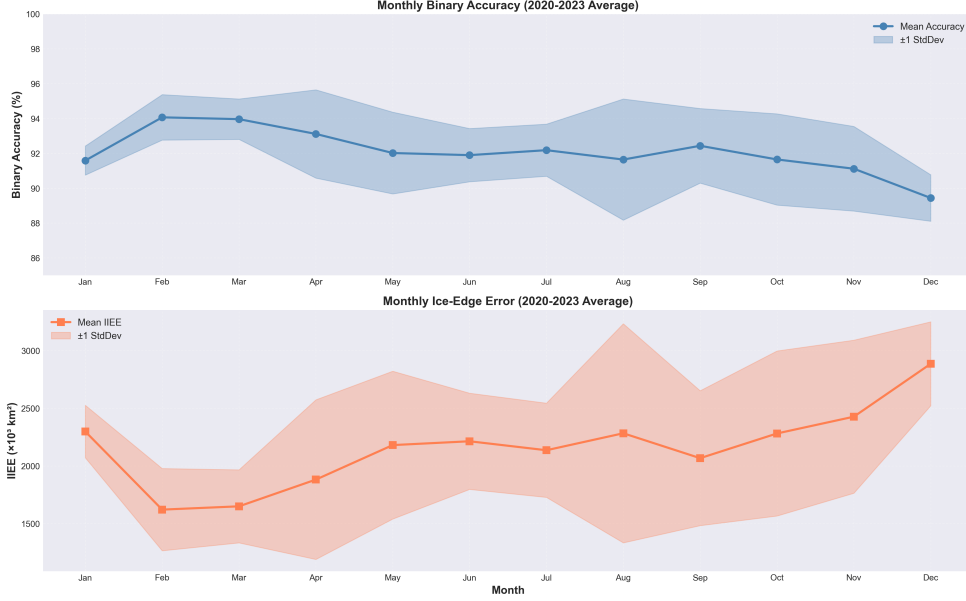


Figure 23: Monthly average Binary Accuracy (top) and Integrated Ice-Edge Error (IIEE) (bottom) for the 3L scenario in the Antarctic region (2020–2023). Shaded areas indicate ± 1 standard deviation. The model demonstrates peak accuracy and lowest error during the summer minimum (February–March), with performance degrading significantly during the early-summer melt phase (December).

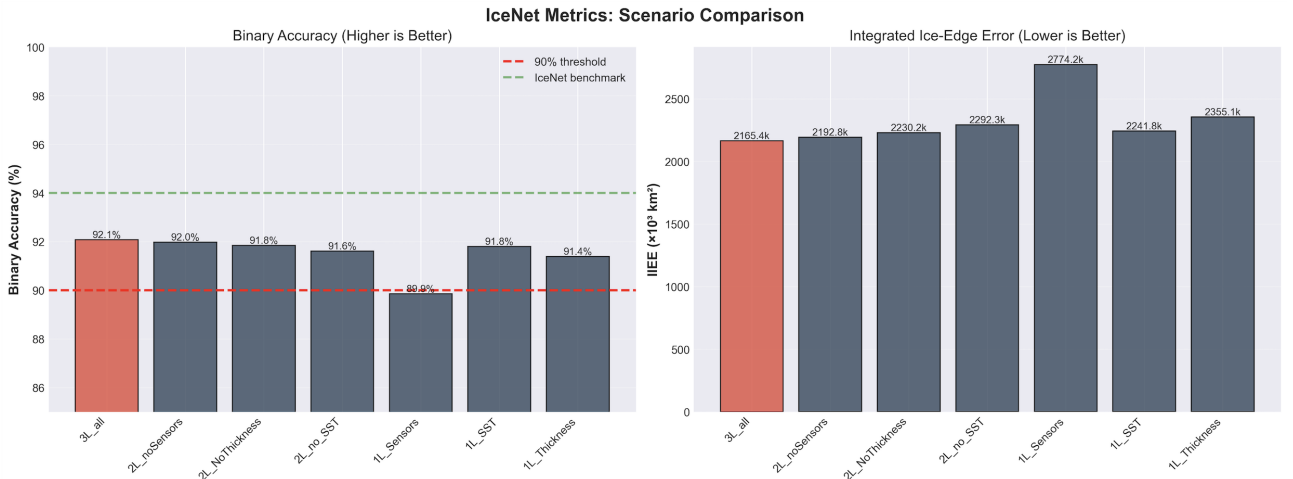


Figure 24: In contrast to the Mean Absolute Error (MAE) results, the model exhibits its poorest performance in terms of Binary Accuracy and Integrated Ice-Edge Error (IIEE) when COMET-QR relies exclusively on the 1L_{Sensors} layer. This particular layer contains only localised residual information. Consequently, while these sensor data are highly informative within their immediate spatial vicinity, they fail to generalise effectively across the entire active domain. This spatial limitation ultimately results in higher misclassification rates during the binary ice/no-ice classification task ($\tau = 0.15$).

While the IIEE evaluates spatial placement, the Sea-Ice Extent (SIE) aggregates this information into a macroscopic scalar representing the total area covered by ice above the threshold τ :

$$\text{SIE}_t = A \sum_{i \in \mathcal{M}_{\text{region}}} \mathbb{I}[n_{i,t} > \tau], \quad \widehat{\text{SIE}}_t = A \sum_{i \in \mathcal{M}_{\text{region}}} \mathbb{I}[\hat{n}_{i,t} > \tau] \quad (36)$$

The temporal accuracy of this macroscopic property is evaluated via the Mean Absolute Error (MAE_{SIE}), measured in millions of square kilometres (10^6 km^2) to align with standard reporting conventions in literature:

$$\text{MAE}_{\text{SIE}} = \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} |\text{SIE}_t - \widehat{\text{SIE}}_t| \quad (37)$$

In Figure 25 we can see how the 3L_all model tracks the massive annual cycle amplitude with high fidelity, achieving a MAE_{SIE} of $1.005 \times 10^6 \text{ km}^2$. Systematic error analysis of the macroscopic extent aligns perfectly with the IIEE decomposition: the model exhibits a net positive bias of approximately $0.782 \times 10^6 \text{ km}^2$ (overestimation); this bias is primarily concentrated during the winter maximum, where even minor fractional errors in predicting the perimeter of the Marginal Ice Zone translate into large absolute area discrepancies due to the sheer length of the winter ice edge. However, the temporal correlation between predictions and observations remains exceptionally high ($r > 0.985$) across all months. In comparison, the scenario (1L_sensors) yields a substantially larger MAE_{SIE} of $1.831 \times 10^6 \text{ km}^2$ and entirely overestimates the transition phases, proving that sparse localised features alone cannot capture abrupt changes in the climatological regime, as relying solely on 80 observation points across a 43,736-pixel grid struggles to resolve the macroscopic ice edge, introducing spatial noise and driving a systematic bias that heavily overestimates the total Sea-Ice Extent. On the other hand, the 2L_noSensors scenario performs almost as well as the complete 3L model, demonstrating that localised sensors offer negligible improvements for macroscopic binary classification. The architecture still exhibits some difficulties in fully capturing severe interannual anomalies. While the models do not simply default to climatological means, their response to extreme events remains constrained. This is particularly evident in Figure 25, during the anomalous retreat in late 2022 and the record-low winter maximum in 2023 ($\sim 17.2 \times 10^6 \text{ km}^2$). During these periods, even the best-performing scenarios struggle to predict the full magnitude of the ice loss, resulting in delayed responses and a notable positive bias that overestimates the actual Sea-Ice Extent.

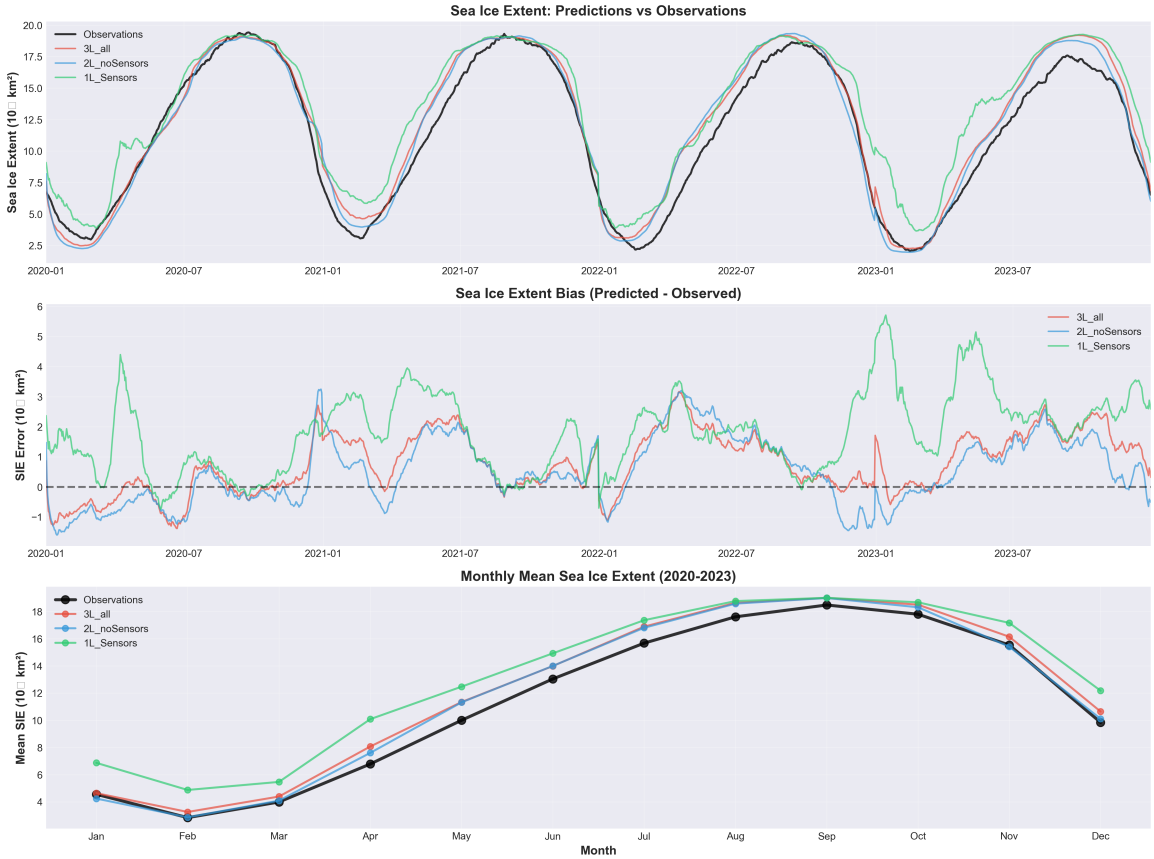


Figure 25: Time series of predicted and observed daily Sea-Ice Extent (SIE) for the test period (2020–2023), comparing distinct scenario predictions against ground truth observations.

7. Conclusions

The forecasting of Antarctic Sea-Ice Concentration remains one of the most formidable challenges in modern climate science, characterised by high-dimensional chaos, non-stationary dynamics, and observational scarcity. In this work, we introduced COMET-QR (Conformal Multi-fidelity Transformer Quantile Regression), a novel hybrid architecture designed to bridge the gap between physically-grounded surrogate modelling and data-driven deep learning. By decomposing the forecasting task into two distinct phases, we successfully isolated the stable, long-term climate dynamics from the highly volatile, short-term anomalies. The offline phase leveraged Bagging Optimised Dynamic Mode Decomposition (BOPDMD) to provide a computationally efficient, zero-cost linear baseline which, however, is limited by its linearity in time and tendency to smooth out fine spatial features, as it relies on single-fidelity data. Therefore, during the online data assimilation process, we employed a Multi-Fidelity Transformer to learn the complex nonlinear corrections and, as we have seen, the empirical results drawn from a rigorous 31-year temporal split demonstrate the superiority of this approach. Key findings of this research include:

- **The Power of Multi-Fidelity Fusion:** Relying on single-source data is insufficient for complex climate modelling. By tokenising Sea Surface Temperature (SST), Sea-Ice Thickness (SIT), and Sparse Sensors into a shared latent space, the Transformer’s attention mechanism successfully learned the cross-correlations between oceanic forcing and ice mechanics. The full multi-fidelity model reduced the Mean Absolute Error by nearly 22% compared to the purely physical BOPDMD surrogate.
- **The Necessity of Localised Error Proxies:** Our ablation studies revealed that global thermodynamic features alone cannot correct localised prediction drift. The inclusion of sparse point-wise residuals proved mandatory to guide the network in correcting the high-frequency errors of the baseline model.
- **Robust Uncertainty Quantification:** By combining Quantile Regression with Conformal Prediction, the framework transitioned from deterministic point estimates to statistically calibrated probabilistic bounds. Applying the conformal correction directly in the clipped physical space ($SIC \in [0, 1]$) guaranteed a valid 90% coverage rate. The resulting uncertainty maps proved physically interpretable, correctly identifying the Marginal Ice Zone and transitional seasons as the primary sources of predictive instability.

The final test evaluation confirms the superiority of the multi-fidelity fusion approach over traditional baselines. The complete 3-level configuration—incorporating SST (91 modes), SIT (95 modes), and 80 Sparse Sensors—achieved the best overall performance with a Mean Absolute Error (MAE) of 0.0746. This translates to a 21.95% improvement over the surrogate BOPDMD model (MAE 0.0955) and a 7.63% improvement over the static Climatology baseline (MAE 0.0807).

The ablation studies clearly isolate the impact of the localised features. Removing the Sparse Sensors degraded the improvement over the single-fidelity BOPDMD baseline to 16.55% (MAE 0.0797). In addition, relying solely on global thermodynamic features without residual correction, having either SST or SIT available, yielded MAEs of 0.0822 and 0.0818, respectively. Given that the Climatology baseline achieves an MAE of 0.0807, these single-modality configurations actually underperformed the static historical average. This proves that while global POD modes provide necessary structural context, the high-frequency Sparse Sensors are strictly required to guide the Transformer in correcting localised, non-linear drift. In contrast, the ablation analysis highlights a fundamental spatial sparsity paradox: while the Sparse Sensors are the critical factor for driving down continuous error (MAE), relying on them in isolation (the $1L_{\text{Sensors}}$ scenario) actively degrades the model’s ability to resolve the binary ice edge (IIEE) and the macroscopic Sea-Ice Extent (SIE). Attempting to reconstruct a continuous physical boundary using only 80 discrete observation points across a 43,736-pixel grid introduces spatial noise and forces the model into a chronic positive bias, heavily overestimating the total ice area. This shows that localised data, despite its high precision, lacks the spatial representation necessary for global classification tasks unless it is fused with the broad thermodynamic context of SST and Thickness.

Finally, the Conformal Quantile Regression bounds proved highly calibrated. The 3L model achieved a strict empirical coverage of 89.5% on the test set, tightly matching the theoretical 90% target ($\alpha = 0.1$), while maintaining a sharp and informative average Confidence Interval width of 0.248.

While COMET-QR represents a significant step forward in operational Sea-Ice forecasting, it also opens several avenues for future research. From an architectural standpoint, scaling the Transformer blocks and increasing the number of attention heads could further enhance the model’s capacity to resolve finer spatial heterogeneities. From a physical perspective, the inclusion of other atmospheric modalities, such as surface wind stress vectors, could lead to better results.

Ultimately, this study demonstrates that the future of climate forecasting does not lie in choosing between computational-physical models and machine learning, but rather in their systematic integration. By embedding physical constraints within the loss landscape and treating diverse observational datasets as interconnected tokens of a unified environment, COMET-QR provides a scalable, risk-aware framework capable of navigating the growing volatility of the Antarctic cryosphere.

References

- [1] Hosking J.S. Pérez-Ortiz M. et al. Andersson, T.R. Seasonal arctic sea ice forecasting with probabilistic deep learning. *Nature Communications*, 12(5124), 2021.
- [2] Antarctic and Southern Ocean Coalition. Changing sea ice, n.d. Scientific consultation by Ted Scambos and Rob Massom.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [4] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual review of fluid mechanics*, 25(1):539–575, 1993.
- [5] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, 5th edition, 2015.
- [6] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, New York, NY, 3rd edition, 2016.
- [7] Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022.
- [8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [9] Paolo Conti, Mengwu Guo, Andrea Manzoni, Attilio Frangi, Steven L. Brunton, and J. Nathan Kutz. Multi-fidelity reduced-order surrogate modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480(2283):20230655, 02 2024.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [11] Delger Erdenesanaa and Leanne Abraham. Where’s all the antarctic sea ice? Annual peak is the lowest ever recorded. *The New York Times*, October 2023. Accessed: 2026-02-03.
- [12] E.U. Copernicus Marine Service Information. Global ocean physics reanalysis (glorys12v1) [data set]. <https://doi.org/10.48670/moi-00021>, 2018. Accessed via Copernicus Marine Service.
- [13] William Gregory, Mitchell Bushuk, Alistair Adcroft, Yongfei Zhang, and Laure Zanna. Deep learning of systematic sea ice model errors from data assimilation increments. *Journal of Advances in Modeling Earth Systems*, 15(10):e2023MS003757, 2023.
- [14] Mengwu Guo, Andrea Manzoni, Maurice Amendt, Paolo Conti, and Jan S. Hesthaven. Multi-fidelity regression using artificial neural networks: Efficient approximation of parameter-dependent output quantities. *Computer Methods in Applied Mechanics and Engineering*, 389:114378, 2022.
- [15] K. H. Hamed and A. R. Rao. A modified mann-kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1–4):182–196, 1998.
- [16] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
- [17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022.
- [18] Zhuoqing Jiang, Bing Guo, Huihui Zhao, Yangming Jiang, and Yi Sun. Sicformer: A 3d-swin transformer for sea ice concentration prediction. *Journal of Marine Science and Engineering*, 12(8), August 2024.
- [19] M. G. Kendall. *Rank correlation methods*. Charles Griffin, 4th edition, 1975.

- [20] Bernard O Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.
- [21] J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.
- [22] Jean-Michel Lellouche, Eric Greiner, Olivier Le Galloudec, Gilles Garric, Charly Regnier, Marie Drevillon, Mounir Benkiran, Charles-Emmanuel Testut, Yann Le Tron, Maeva Faucher, et al. Recent updates to the copernicus marine service global ocean monitoring and forecasting real-time 1/12° high-resolution system. *Ocean Science*, 14(5):1093–1126, 2018.
- [23] H. B. Mann. Nonparametric tests against trend. *Econometrica*, 13(3):245–259, 1945.
- [24] Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, 2005.
- [25] Marco Da Mommio. MELTER: a transformer architecture for multi-fidelity problems. Master’s thesis, Politecnico di Milano, Milan, Italy, 2024. Advisor: Andrea Manzoni, Co-advisor: Paolo Conti.
- [26] Christopher Olah. Understanding LSTM networks. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>, August 2015. Accessed: 2024-05-23.
- [27] Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, et al. AION-1: Omnimodal foundation model for astronomical sciences. *arXiv preprint*, 2024.
- [28] Luke A. Roach, Sally M. Dean, and Cecilia M. Bitz. Antarctic sea ice area in cmip6. *Geophysical Research Letters*, 47(9):e2020GL087965, 2020.
- [29] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3538–3548. Curran Associates, Inc., 2019.
- [30] Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.
- [31] Lawrence Sirovich. Turbulence and the dynamics of coherent structures. Part I: Coherent structures. *Quarterly of Applied Mathematics*, 45(3):561–571, 1987.
- [32] Gregory C. Smith et al. Polar operational oceanography and sea ice forecasting. *Frontiers in Marine Science*, 6:429, 2019.
- [33] J. Stark, D.S. Broomhead, M.E. Davies, and J. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods Applications*, 30(8):5303–5314, 1997. Proceedings of the Second World Congress of Nonlinear Analysts.
- [34] John Turner, J. S. Hosking, G. J. Marshall, T. Phillips, and T. J. Bracegirdle. Antarctic sea ice increase consistent with intrinsic variability of the amundsen sea low. *Climate Dynamics*, 46:2391–2402, 2016.
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [36] D. S. Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, 3rd edition, 2011.
- [37] Qinghua Yang et al. Deep learning shows promise for seasonal prediction of antarctic sea ice in a rapid decline scenario. *Advances in Atmospheric Sciences*, 41(8):1–5, 2024.

A. Standard Exploration for Climate Spatiotemporal Data

This appendix details the foundational statistical analysis performed to characterise the long-term evolution of the Antarctic cryosphere within the study domain. The objective of this preliminary phase is twofold: first, to quantify the average rate of change (trends) for the primary physical variables; and second, to identify regions where these changes are statistically significant. By establishing these secular trends, we provide a baseline against which the high-frequency residuals modelled by our multi-fidelity architecture can be evaluated.

Consistent with established diagnostic protocols in climate research, our analysis begins by quantifying long-term secular trends through the application of standard parametric linear regression. This approach assumes that the underlying trends are linear and monotonic over the observation period (1993–2023).

A.1. Yearly Aggregation and Ordinary Least Squares (OLS) Regression

To isolate inter-annual variability from seasonal cycles, we first aggregate the daily data into yearly mean time series. For a generic spatiotemporal variable y (representing Ice Thickness, SST, or Concentration), the yearly mean y_k for year k at a specific grid point is computed as:

$$y_k(i, j) = \frac{1}{N_d} \sum_{t=1}^{N_d} y_{k,t,i,j} \quad (38)$$

where $N_d = 365$ is the number of days in the standardised year. This operation yields a sequence of yearly spatial fields $\{y_k\}_{k=1}^{N_y}$, where $N_y = 31$.

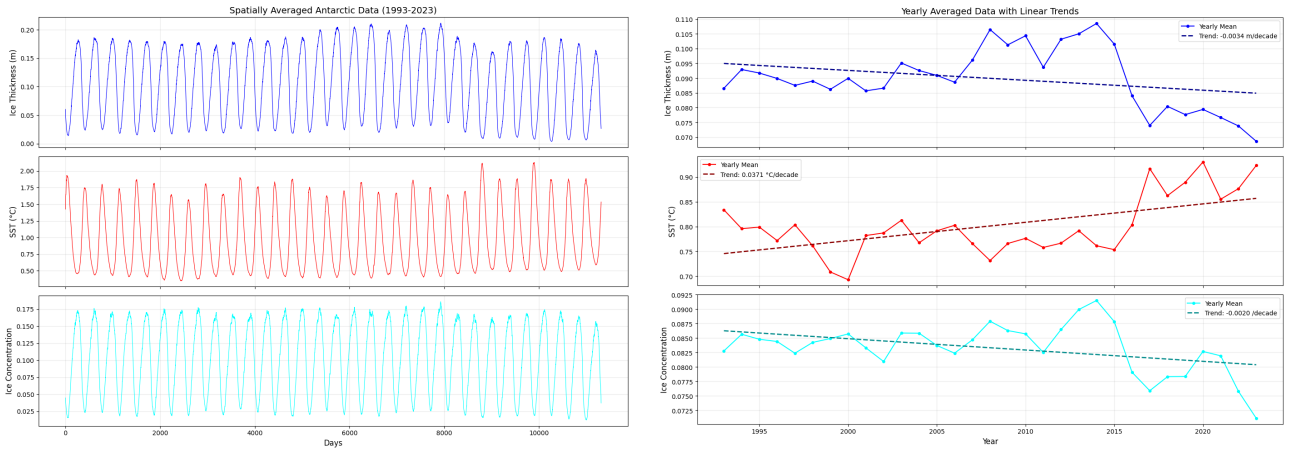


Figure 26: Time Series vs Yearly averaged data with linear trends

We model the temporal evolution of the yearly means [36] using a simple linear regression model:

$$y_k = \beta_0 + \beta_1 k + \epsilon_k, \quad k = 1, \dots, N_y \quad (39)$$

where β_0 is the intercept, β_1 represents the trend slope (rate of change per year), and ϵ_k is the residual error term. The parameters are estimated using the Ordinary Least Squares (OLS) method, which minimises the sum of squared residuals:

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{N_y} (k - \bar{k})(y_k - \bar{y})}{\sum_{k=1}^{N_y} (k - \bar{k})^2} \quad (40)$$

Here, $\bar{k}$ and $\bar{y}$ denote the temporal means of the time index and the climate variable, respectively. To ensure the results are physically interpretable and comparable across literature, we scale the estimated slope $\hat{\beta}_1$ to represent the trend per decade [19]:

Table 4: Summary of linear trends.

Linear Trends	
SIC	-0.0020 /decade
SIT	-0.0034 m/decade
SST	+0.0371 °C/decade

A.2. Mann-Kendall Spatial Yearly Trend Analysis

To robustly assess the temporal evolution of Antarctic climate variables, we employ the Mann-Kendall (MK) test [23]. Unlike the Ordinary Least Squares (OLS) regression described in previous sections, the MK test is a non-parametric statistical method. It assesses the presence of a monotonic trend without assuming a specific distribution (e.g., normality) for the data, making it particularly suitable for hydro-meteorological time series which often exhibit non-normality and outliers. The Mann-Kendall test evaluates the null hypothesis H_0 that there is no monotonic trend in the series, against the alternative hypothesis H_1 that a monotonic trend exists. For a time series $X = \{x_1, x_2, \dots, x_n\}$, the Mann-Kendall statistic is calculated by examining all pairs of data:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(x_j - x_k) \quad (41)$$

To determine the statistical significance of the trend, the variance of S must be computed [15]. This computation requires a correction for "tied" groups (identical values), which are common in ice concentration data.

$$\text{Var}(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^q t_p(t_p-1)(2t_p+5)}{18} \quad (42)$$

where n is the number of data points, q is the number of tied groups and t_p is the number of data points in the p -th tied group. For sample sizes $n > 10$, the statistic S approximately follows a normal distribution. We calculate the standardised test statistic Z_{MK} using a continuity correction:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & \text{if } S < 0 \end{cases} \quad (43)$$

The null hypothesis H_0 is rejected at the significance level α if $|Z_{MK}| > Z_{1-\alpha/2}$. A positive Z_{MK} indicates an increasing trend, a negative one indicates a decreasing trend. To distill this dense information into a format capable of revealing long-term climatic shifts, a pixel-wise temporal aggregation procedure is applied as the preliminary step. During this phase, the algorithm isolates the time series for every individual geographic coordinate and mathematically compresses the full sequence of daily observations into a single representative annual mean, effectively filtering out short-term seasonal noise to expose the underlying inter-annual signal. Once this temporal reduction establishes a clean yearly timeline for every point on the map, the workflow shifts to a systematic spatial examination where each pixel is treated as an independent monitoring station. A rigorous data screening protocol is immediately applied to these generated time series, ensuring that only coordinates possessing a sufficient density of valid historical observations are permitted to advance to the statistical testing phase.

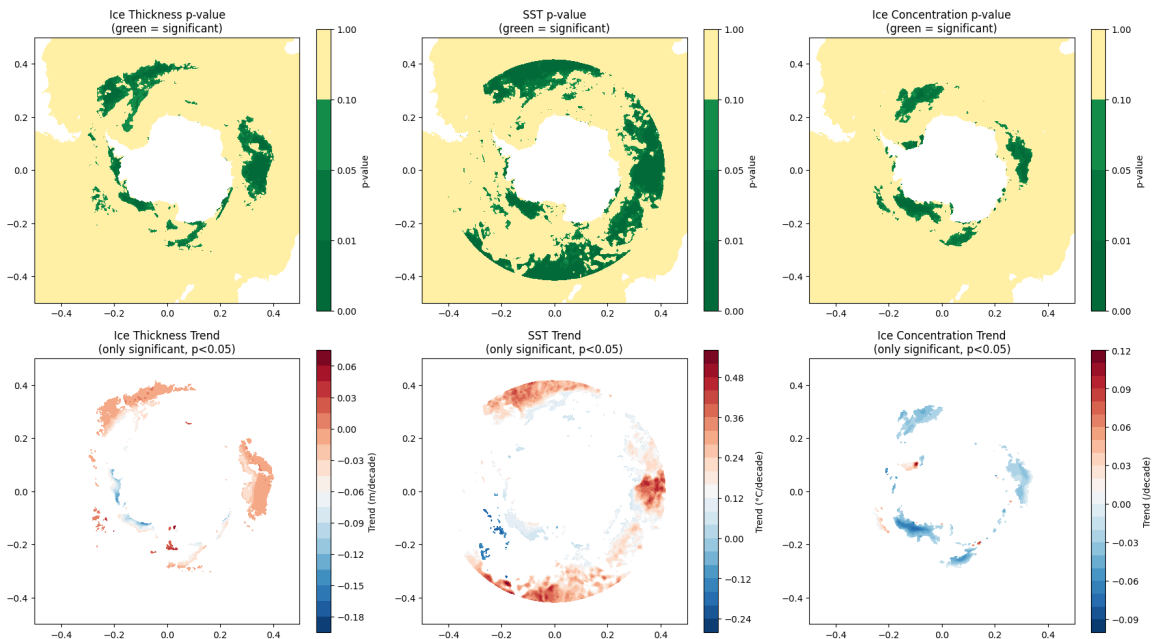


Figure 27: Mann-Kendall test on pixel-wise yearly averaged variables

B. SARIMA (Seasonal AutoRegressive Integrated Moving Average)

While the COMET-QR framework relies on Bagging Optimised Dynamic Mode Decomposition (BOPDMD) and Multi-Fidelity Transformers to capture the complex spatiotemporal dynamics of Antarctic Sea-Ice, it is instructive to establish a classical statistical baseline. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model provides a rigorous mathematical framework for univariate time series exhibiting periodic fluctuations, such as the seasonal freezing and melting cycles of the cryosphere.

Let $Y_t \in [0, 1]$ be the realization of the Sea-Ice Concentration (SIC) at a specific spatial location at time t . We define the backshift operator B such that $B^k Y_t = Y_{t-k}$ [16]. To achieve stationarity in the mean, a necessary condition for autoregressive moving average processes, linear difference operators are introduced [6]. These include the non-seasonal difference operator, $\nabla^d = (1 - B)^d$, which removes local polynomial trends, and the seasonal difference operator, $\nabla_s^D = (1 - B^s)^D$, which addresses non-stationarity linked to the seasonal cycle of period s (e.g., $s = 365$ for daily data).

A stochastic process denoted as $SARIMA(p, d, q) \times (P, D, Q)_s$ is governed by the multiplicative interaction of four complex polynomials in the operator B [5]:

- Non-seasonal Autoregressive $AR(p)$: Captures the linear dependence on the p immediate past observations.

$$\phi(B) = 1 - \sum_{i=1}^p \phi_i B^i$$

- Non-seasonal Moving Average $MA(q)$: Models the memory of past stochastic shocks up to lag q .

$$\theta(B) = 1 + \sum_{j=1}^q \theta_j B^j$$

- Seasonal Autoregressive $SAR(P)$: Models the dependence on observations from the same calendar days in previous years.

$$\Phi(B^s) = 1 - \sum_{i=1}^P \Phi_i B^{is}$$

- Seasonal Moving Average $SMA(Q)$: Models the persistence of seasonal stochastic shocks.

$$\Theta(B^s) = 1 + \sum_{j=1}^Q \Theta_j B^{js}$$

The complete theoretical dynamic of the SIC within this mathematical framework is expressed by the multiplicative application of these polynomial filters to the differenced series:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d Y_t = \Theta(B^s)\theta(B)\epsilon_t \quad (44)$$

where ϵ_t represents a pure Gaussian white noise process, such that $\mathbb{E}[\epsilon_t] = 0$ and $Var(\epsilon_t) = \sigma^2$.

Expanding the operators for a first-order model, such as a $SARIMA(1, 0, 0) \times (1, 0, 0)_{365}$, explicitly demonstrates the multiplicative nature that generates interaction terms at mixed lags (e.g., at day 366):

$$(1 - \phi_1 B)(1 - \Phi_1 B^{365})Y_t = \epsilon_t \implies Y_t = \phi_1 Y_{t-1} + \Phi_1 Y_{t-365} - \phi_1 \Phi_1 Y_{t-366} + \epsilon_t \quad (45)$$

B.1. Limitations for Sea-Ice Forecasting

Despite its analytical elegance, the SARIMA framework exhibits structural limitations when applied to Sea-Ice forecasting, which inherently justify the transition to the Multi-Fidelity Transformer architecture proposed in this study. The Antarctic system is governed by high-dimensional spatiotemporal chaos and unbounded interactions between atmospheric forcing and oceanic currents. Within this context, the SARIMA model is fundamentally limited because:

1. It is strictly linear: It cannot map the complex thermodynamic bifurcations and extreme volatility observed in recent regime shifts of the Antarctic pack.
2. It is intrinsically univariate: It treats each spatial grid point in isolation, thereby ignoring cross-correlations with other thermodynamic features such as Sea-Ice Thickness (SIT) and Sea Surface Temperature (SST).

C. COMET-QR Insights

C.1. Proper Orthogonal Decomposition (POD)

Following the framework established by [4], Proper Orthogonal Decomposition (POD) is defined as a method to identify the most energetic coherent structures within a dynamical system. It is a dimensionality reduction technique which provides an optimal basis: for any given number of modes k , the POD basis represents the k -th directions which explain the most of the variability of the system.

Let $X \in \mathbb{R}^{n \times m}$ be the data matrix containing m temporal snapshots, where each column $x^{(k)}$ represents the fluctuating field at n spatial points. The goal is to find a set of orthogonal basis functions (modes) ϕ that maximise the projection of the data onto the basis. This is equivalent to solving the eigenvalue problem for the spatial correlation tensor (or covariance matrix) $C \in \mathbb{R}^{n \times n}$:

$$C\phi = \lambda\phi \quad \text{where} \quad C = \frac{1}{m}XX^T \quad (46)$$

Here, the eigenvalue λ represents the energy captured by the corresponding mode ϕ .

C.1.1 The Method of Snapshots & SVD

Since the spatial dimension n is typically much larger than the number of snapshots m (i.e., $n \gg m$), directly computing the eigenvectors of the $n \times n$ matrix C is computationally prohibitive.

Instead, we employ the method of Snapshots [31], which relates the POD modes to the Singular Value Decomposition (SVD) of the data matrix:

By employing the Method of Snapshots, we factorise the data matrix as:

$$X = U\Sigma V^T \quad (47)$$

Where the components possess the following mathematical and physical properties:

- $U \in \mathbb{R}^{n \times m}$ is an orthogonal matrix ($U^T U = I$) whose columns u_i represent the left singular vectors. Physically, these are the spatial POD modes, representing fixed, spatially uncorrelated patterns of the system's variance (e.g., stationary geographic anomalies in the ice pack).
- $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_m)$ is a diagonal matrix containing the singular values sorted in strictly descending order ($\sigma_1 \geq \sigma_2 \geq \dots \geq 0$). These values mathematically quantify the absolute variance captured by each respective mode, where the corresponding eigenvalue of the covariance matrix is given by $\lambda_i = \sigma_i^2/m$.
- $V \in \mathbb{R}^{m \times m}$ is an orthogonal matrix ($V^T V = I$) whose columns v_i represent the right singular vectors. Physically, these dictate the uncorrelated temporal evolution (the time-series coefficients) that modulate the amplitude of each corresponding spatial mode u_i over the observation period.

By virtue of the Eckart-Young-Mirsky theorem, the SVD provides the optimal low-rank approximation of the original high-dimensional dataset in the least-squares sense. The optimal reduced-order approximation of rank k is obtained by truncating the series to retain only the most energetic coherent structures:

$$X \approx \sum_{i=1}^k \sigma_i u_i v_i^T \quad (48)$$

C.1.2 Energy Criterion (Variance Threshold)

The eigenvalues λ_i (or squared singular values σ_i^2) represent the energy associated with each mode. To select the cutoff k , we define the Relative Information Content (RIC):

$$\text{RIC}(k) = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^m \lambda_i} \geq \tau \quad (49)$$

We retain the first k modes such that the cumulative energy exceeds a user-defined threshold τ (e.g., $\tau = 0.90$).

C.2. Other Plots

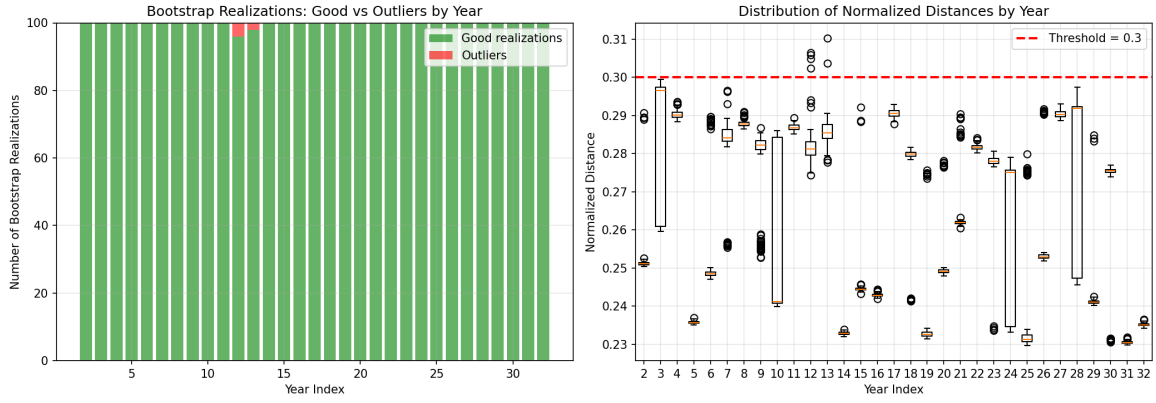


Figure 28: BOPDMD Outliers Identification and Removal for years 1993-2023

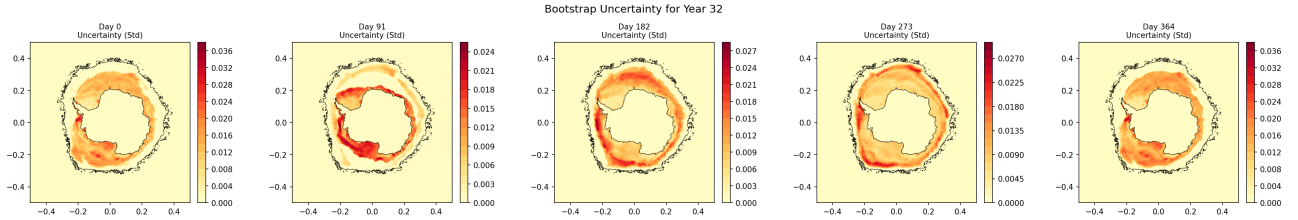


Figure 29: Uncertainty of BOPDMD forecasted SIC over 5 days of year 2023

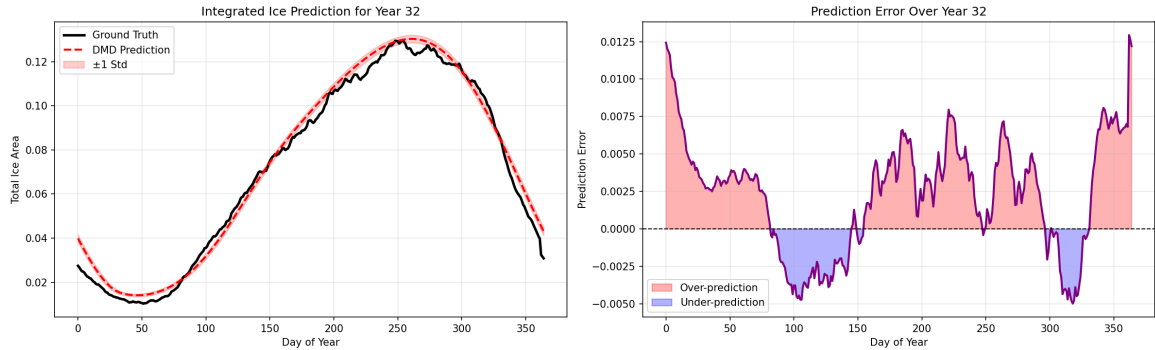


Figure 30: (Left) Integrated Ice BOPDMD prediction for 2023. (Right) Prediction error for 2023

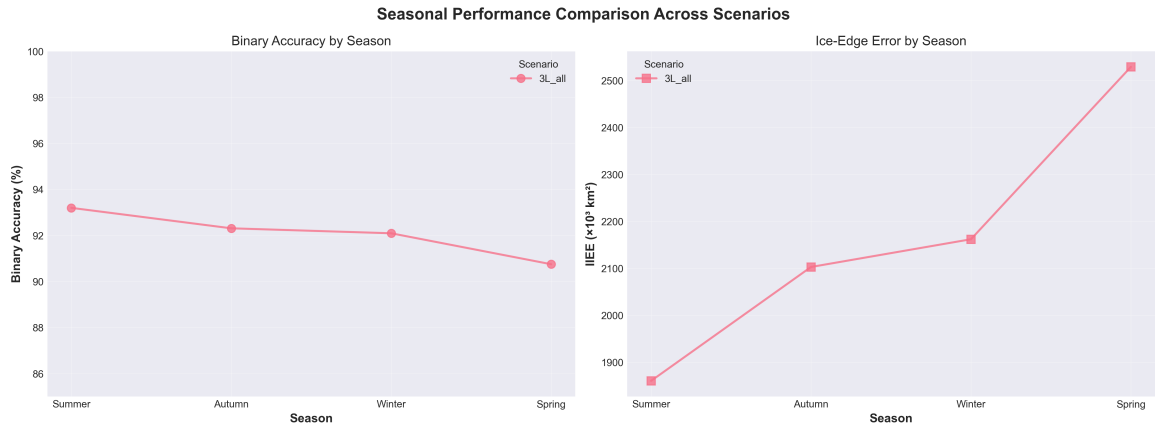


Figure 31: Seasonal performance of COMET-QR for the 3L scenario

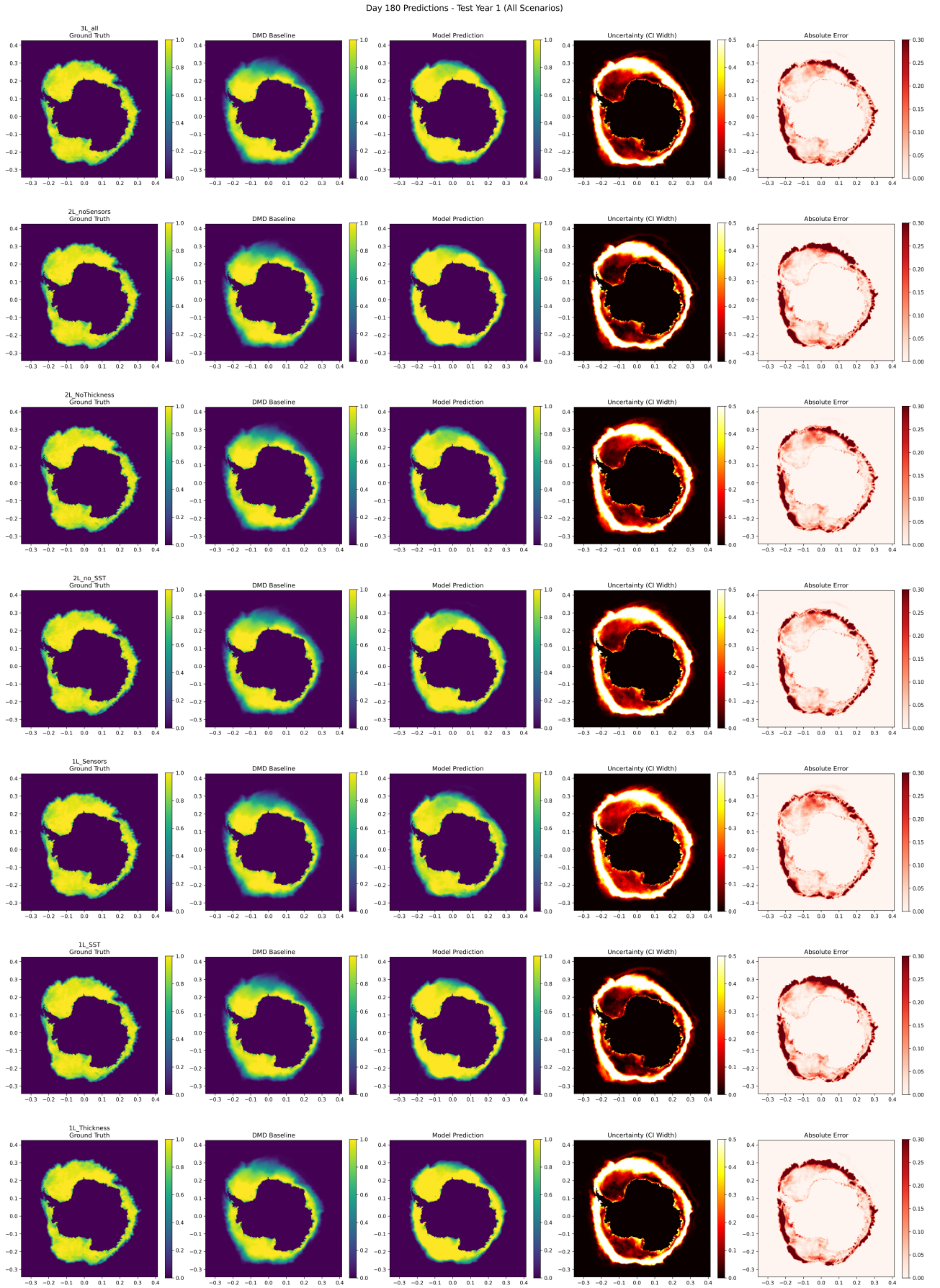


Figure 32: Qualitative assessment of data availability impacts: median SIC predictions for all seven input scenarios on day 180 of the initial test year (2020)

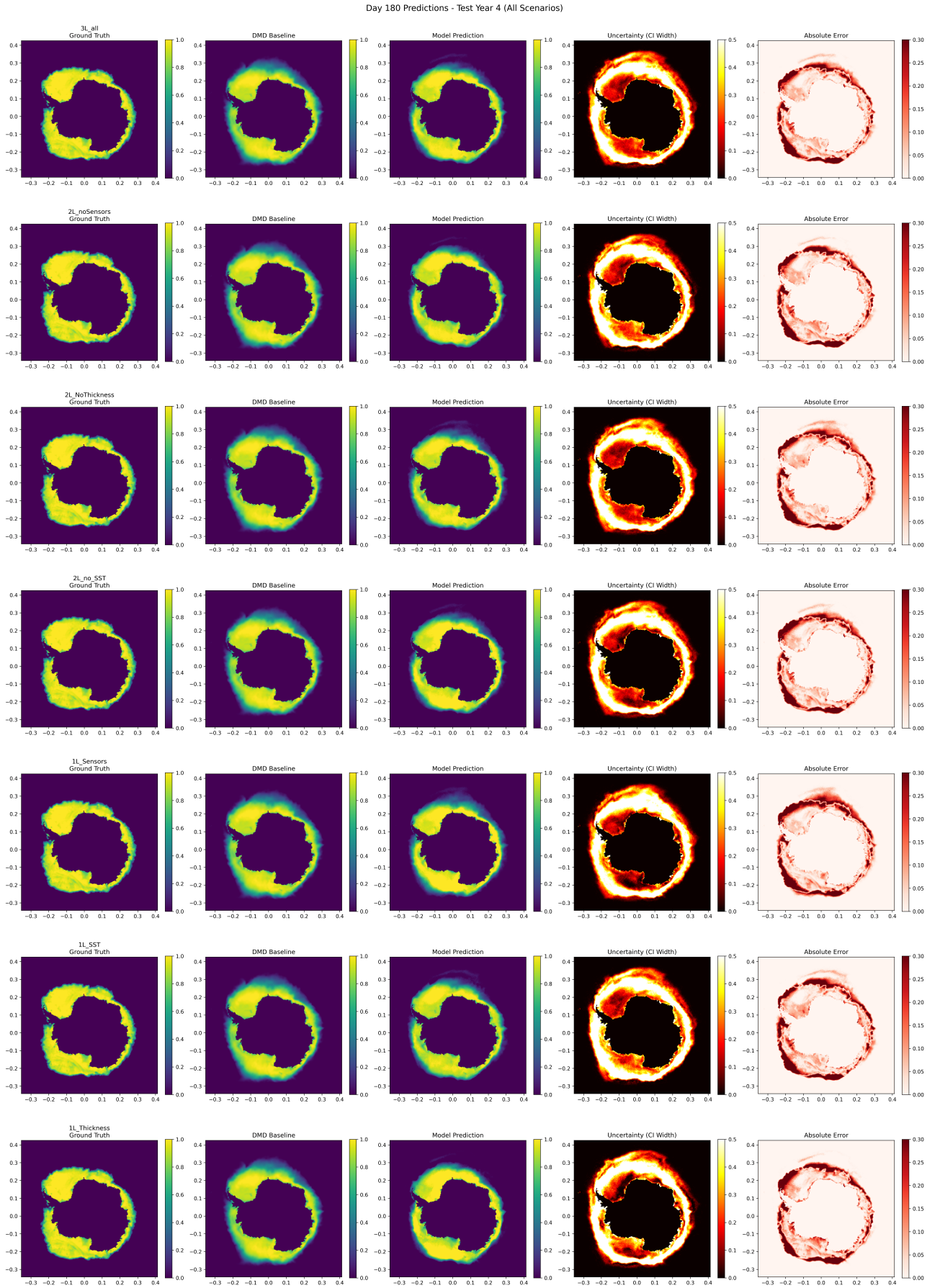


Figure 33: Model predictions for day 180 of 2023. Although the observed Sea-Ice Concentration (SIC) ground truth deviates significantly from previous years (e.g., 2020), the COMET-QR architecture demonstrates robust predictive performance. This highlights the framework’s ability to generalise to unprecedented climate anomalies, despite being trained exclusively on historical data from 1993 to 2015.

Abstract in lingua italiana

Il ghiaccio marino antartico svolge un ruolo fondamentale nella regolazione del sistema climatico globale. Nonostante decenni di analisi satellitari e modellistiche, le dinamiche che governano la variabilità del ghiaccio marino antartico rimangono poco comprese a causa dei forti contrasti regionali, dell'elevata variabilità naturale e dei complessi meccanismi di accoppiamento tra i processi atmosferici, oceanici e criosferici. Per questa ragione, anche i modelli più promettenti, che siano essi puramente fisici o data driven, faticano a descrivere le interazioni non lineari, in particolar modo quando si combinano dati eterogenei caratterizzati da differenti risoluzioni spaziali e temporali. Al fine di migliorare le previsioni del ghiaccio marino, questo studio sviluppa un innovativo approccio a due fasi: (i) un modello surrogato lineare, costituito da una Bagging optimised Dynamic Mode Decomposition, e (ii) un framework di deep learning basato sull'architettura Transformer che, attraverso la fusione di molteplici fonti di dati, ha l'obiettivo di correggere le previsioni del surrogato grazie all'apprendimento delle variazioni non lineari del ghiaccio marino nel corso del tempo. In particolare, sfruttando il meccanismo di attention, il modello cattura le dipendenze spazio-temporali e le relazioni cross-modali tra i diversi livelli di fedeltà. Inoltre, consapevoli dell'importanza di quantificare l'incertezza oltre a fornire una stima puntuale, abbiamo utilizzato la regressione quantilica (quantile regression) per produrre intervalli di confidenza calibrati per le previsioni della concentrazione del ghiaccio marino. Le analisi preliminari indicano che gli approcci basati sui Transformer riescono a rappresentare la variabilità su larga scala e le anomalie regionali in modo molto più flessibile rispetto ai metodi convenzionali. Nel complesso, la strategia proposta consente di migliorare la comprensione del comportamento del ghiaccio marino antartico, tenendo conto anche degli eventi estremi, e costruendo un ponte tra l'inferenza puramente guidata dai dati e la modellazione di natura fisica.

Parole chiave: Sea-Ice Concentration Forecasting, Progressive Multi-Fidelity Learning, Dynamic Mode Decomposition, Transformers, Conformal Quantile Regression, Spatiotemporal Predictions