# Hybrid Machine Learning Framework for Predicting Global Ocean $CO_2$ Fluxes Using Long Term Observations and Satellite Constraints with Regime Shift Detection

by

Syeda Nazifa Ali
22101025
Fahmida Akter Rimo
21301180
Amit Kumer Halder
21201770

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January, 2026

# Declaration

It is hereby declared that

1. The thesis submitted is our original work while pursuing a degree at Brac University.
2. The thesis does not include previously published or written information by a third party, unless it is properly cited with complete and correct referencing.
3. The thesis must not include any material that has been approved or submitted for another degree or diploma at a university or other institution.
4. We have acknowledged all the major sources of aid.

**Student's Full Name & Signature:**

Syeda Nazifa Ali

22101025

Fahmida Akter Rimo

21301180

Amit Kumer Halder

21201770

# Approval

The thesis titled "Hybrid Machine Learning Framework for Predicting Global Ocean $CO_2$ Fluxes Using Long Term Observations and Satellite Constraints with Regime Shift Detection" submitted by

1. Syeda Nazifa Ali (22101025)
2. Fahmida Akter Rimo (21301180)
3. Amit kumer Halder(21201770)

of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in January 31, 2026.

**Examining Committee:**

Supervisor:
(Member)

Swakkhar Shatabda

Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Designation
Department of Computer Science and Engineering
Brac University

# Abstract

Prediction of global ocean–atmosphere $CO_2$ exchange is often limited by sparse *in situ* observations and the indirect nature of satellite measurements, leading to substantial uncertainty and disagreement among biogeochemical models, observation-based reconstructions, and atmospheric inversion products in both flux magnitude and variability. We develop a hybrid, time-aware machine learning framework that reconstructs monthly global $1^\circ \times 1^\circ$ $CO_2$ flux by learning relationships between air–sea exchange and multi-source environmental drivers. The learning target is the Jena CarboScope ocean flux product derived from SOCAT, which provides spatially and temporally continuous flux fields during 2014–2024 in $2^\circ \times 2^\circ$ and serves as a benchmark consistent with carbon-budget constraints. Predictors integrate physical ocean state from ORAS5 (sea surface temperature, salinity, mixed-layer depth, sea surface height), atmospheric forcing from ERA5 (10 m zonal and meridional winds), and biogeochemical indicators from climate-quality products (chlorophyll-a, phosphate, and net primary production), harmonized to a common monthly grid with seasonality encoded via cyclic month terms and standardized preprocessing. Prediction is formulated as supervised regression at the grid-cell level and evaluated under strict temporal generalization using expanding leave-one-year-out windows with train/validation/test splits to prevent leakage; raw-driver and anomaly-augmented representations are compared to assess the value of emphasizing interannual departures or physics-informed features. Results show that nonlinear tree-based learning generalizes most robustly across unseen years, with Random Forest achieving the lowest mean window errors (actual dataset: MSE $1.067 \times 10^{-7}$, RMSE $3.26 \times 10^{-4}$, MAE $2.16 \times 10^{-4}$; anomaly-augmented dataset: MSE $9.829 \times 10^{-8}$, RMSE $3.13 \times 10^{-4}$, MAE $2.06 \times 10^{-4}$) and best performance in the longest training window (2014–2022 training; 2024 test), indicating improved transfer with increased historical variability. Beyond reconstruction, predicted flux fields are aggregated using RECCAP2 basin masks and conditioned to deseasoned, detrended interannual variability (IAV) signals to examine changes in basin-scale source–sink behavior; regime shifts are detected conservatively using the nonparametric Pettitt change-point test with 12-month persistence confirmation, yielding three candidate shifts (Atlantic: 2016-06; Southern Ocean: 2020-10; Arctic: 2021-07), and interpreted using Niño3.4 ENSO as a physical coherence check. This hybrid framework provides a reproducible pathway for global $CO_2$ flux monitoring that integrates multi-source predictors with observation-based anchoring, supports climate-relevant variability analysis beyond headline accuracy metrics, and establishes foundations for future early-warning prediction using shift labels and precursor indicators.

**Keywords:** Ocean-atmosphere exchange, Supervised regres sion, Leave-one-year-out validation, Interannual variability, Regime shifts.

# Acknowledgement

We convey our deepest appreciation to our research supervisor Swakkhar Shatabda sir for his valuable guidance and insights throughout the research on carbon flux estimation incorporating machine leaning approach. His profound understanding and expertise in machine learning provided full support in the successful completion of this research.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

The ocean serves as a central regulator of Earth's climate by exchanging vast quantities of carbon dioxide ($CO_2$) with the atmosphere and sequestering carbon over diverse timescales. Over recent decades, the global ocean has functioned as a major net sink for anthropogenic $CO_2$, absorbing approximately one quarter of human emissions and thereby substantially mitigating the pace of atmospheric $CO_2$ accumulation and associated global warming [26]. But this shifting climate service inconceivable susceptibility, ocean absorption is intensely influenced by circulation patterns, temperature differentials, wind-based gas exchange and biological productivity all of which seasonally vary and are sensibly active to climate variability and enormous long-term change. [8]. At the same time, the absorption of CO2 alters the seawater carbonate chemistry, causing ocean acidification and initiating a series of effects on the ocean ecosystems [24]. It is against this backdrop that proper perception and subsequent anticipation of air-sea carbon dioxide exchange becomes a vital aspect in believable climate measurements, stringent review of mitigation advancement, and initial identification of impending disturbances in the ocean carbon sink. These disturbances pose a danger to the ability of the ocean to offset additional atmospheric rise of CO2, with considerable consequences on climate pathways and the well-being of marine life on Earth [7], [28].

Proper assessments of air-sea CO 2 fluxes are essential due to a number of interrelated reasons. To begin with, they support the accounting of global carbon budget by dividing anthropogenic emissions between the atmosphere, the land and ocean, and thus allow strict allocation of observed growth rates of atmospheric CO2 and strict quantification of uncertainties in climate forecasting. [28]. Second, spatiotemporal information of sink/source behavior of oceans can enable diagnosis of the control regional processes, such as wind-induced upwelling, stratification, and deepening of the mixed layer, and biological productivity. Third, robust flux reconstructions serve as essential benchmarks to evaluate Earth system models and biogeochemical simulations, thereby exposing structural biases in representations of circulation, gas exchange, and ecosystem dynamics. there is a need to monitor the dynamic dynamics of the ocean carbon sink on a long-term basis, as climate-induced perturbations, such as acidification and deoxygenation, risk reducing the capacity of the carbon sink, changing the average condition as well as variability of CO2 exchange over an interannual to decadal period [7]. Due to the interaction of natural climate variability and anthropogenic forcing, which modulates these rate of uptake in interannual and decadal time scales, a feedback cycle arises wherein, on interannual and decadal time scales, atmospheric CO2 drives climate change which in turn alters carbon-relevant natural systems in the ocean and on land [12].

Nevertheless, even though of paramount importance, the estimation of global air-sea CO2 fluxes is deeply problematic because of inherent constraints of observation. Direct measurements of surface ocean CO2 are sparse in relation to the spatio-temporal scale of ocean variability, easily limited to non-uniform ship track distributions which sample large oceanic areas inefficiently and are prohibitively expensive to scale up to sustained campaigns. One recent instance of community response is the Surface Ocean CO2 Atlas, which assembles and quality controls tens of millions of measurements of surface ocean CO2 globally systematically, providing a solid quantification of basin-scale uptake and a solid model test. However, even the SOCAT with its detailed data set has needed a large number of modeling assumptions and advanced gap-filling techniques to provide complete global fields of flux, and non-negligible uncertainties are present. Therefore, there is a multiplicity of complementary methodologies: process-based ocean biogeochemical models, which model underlying physics and biology; observation -constrained reconstructions of surface CO2 and fluxes; and atmospheric inversion systems which estimate oceanic fluxes by combining in situ atmospheric CO2 measurements with transport modeling. Despite these advances in knowledge, each is faced with long-standing uncertainties in data scarcity, circulation and transport biases, empirical parameterizations, and methodological discrepancies, necessitating the exploration of new data-driven models, such as machine learning, which take advantage of a wide range of environmental covariates and anchor the results of these models to existing products of observation, such as Jena CarboScope inversions, to products based on observation. However, even with these constrained frameworks, the actual magnitude of the magnitude of air-sea fluxes of CO 2 and the time variability of the same have continued to encounter significant uncertainties when considering various methods such as ocean biogeochemical models, observation-based products, and atmospheric inverse models [25].

In this topography of varied approaches full of uncertainties , flux products defined by observation, including the Jena CarboScope ocean inversion, are important to define the primary fields of synthesis, model testing, and comparisons to Earth system simulation. The learning goal in this thesis is the Jena CarboScope product, where quality-controlled surface ocean observations (e.g., of SOCAT) are combined with atmospheric transport modeling to produce observed based estimates of the world air-sea CO2 fluxes. Such a product has a wide range of possible uses in climatology because it provides spatially and temporally continuous reconstructions of flux variability that are used to make powerful inter-regional and intertemporal comparisons that can be used to understand the dynamics and mechanisms of the large-scale carbon-cycle . [28]. However, the presence of several flux products, all of which are based on different assumptions, data streams, and parameterizations, continues to generate significant inter-method differences in magnitude and variability of the fluxes [12], [25]. This methodological divergence compellingly motivates data-driven approaches like machine learning, which can uncover robust, generalizable relationships between fluxes and multifaceted environmental drivers transcending dependence on any single reconstruction pathway. Machine learning represents a particularly promising framework for addressing these challenges precisely because air sea $CO_2$ exchange is governed by nonlinear, interacting thermodynamic, dynamical, and biological controls that are difficult to capture with traditional linear parameterizations [19], [20].

The machine learning in this purpose is thus especially promising as the air-sea CO2 exchange is controlled by nonlinear, interacting thermodynamic, dynamical, and biological controls that cannot be captured using a more traditional linear parameterization [12]. The predictor set used in this thesis is clearly constructed to reflect these coupled controls: physical ocean variables of ORAS5, atmospheric forcing of ERA5, and biogeochemical indicators of climate-quality satellite products. The framework is used to reconstruct the spatiotemporal distribution of ocean CO2 fluxes using widely available drivers to scale-up to global analysis with benchmark capabilities against observation constrained products, and to eliminate methodological divergences [25], [28]. However, reliance on purely statistical mappings raises concerns about physical consistency, necessitating explainable approaches that ensure reconstructed relationships align with established biogeochemical theory and ocean dynamics [21]. To bridge this gap, recent advances integrate explicit physical knowledge into the learning architecture, such as pre-processing $p\mathrm{CO}_2$ data to isolate the thermodynamic effects of temperature and thereby enhancing the biogeochemical signals the model must capture [10].

Beyond predictive reconstruction, an equally important scientific question is whether the ocean carbon sink exhibits persistent shifts in behavior over time often discussed in climate science as regime shifts. In the context of ocean carbon dynamics, regime shifts can be understood as persistent changes in the statistical properties of flux variability (e.g., mean level, variance, persistence) in circulation and climate forcing rather than short lived anomalies. Detecting such changes matters because a shift in basin scale variability or persistence can indicate altered sensitivity of air–sea $CO_2$ exchange implications for predictability, carbon-budget uncertainty, and interpretation of long-term trends. Importantly, many regime-like changes are not most visible in raw flux magnitude, which is dominated by strong seasonality and slower variations; detectability often improves when analyzing anomaly or interannual-variability (IAV) signals at basin scale after removing the seasonal cycle and long-term trend components.[10]

To address these needs, this thesis develops a hybrid machine learning framework for predicting global ocean atmosphere $CO_2$ flux patterns using observation-based flux products as targets and multi-source environmental predictors as inputs. The framework is "hybrid" in the sense that it combines (i) heterogeneous observation-informed datasets (satellite, reanalysis, and in situ–constrained products) and (ii) two complementary scientific objectives: flux prediction and regime-shift characterization. The target variable (co2flux) is derived from the Jena CarboScope inversion product based on SOCAT surface $CO_2$ observation providing a reliable reference for training and evaluation of ML models. The predictor datasets integrate ORAS5 physical ocean fields, ERA5 winds, and climate-quality satellite biogeochemical indicators, enabling data-driven reconstruction of flux variability from indirect drivers of air–sea exchange. A second contribution is the basin-scale interpretation layer. Using RECCAP2 basin masks, the thesis emphasizes basin-level aggregation as a pathway to reduce grid-scale noise and ensure physically interpretable large-scale behavior aligned with carbon-cycle literature. Regime-shift analysis is then framed on anomaly/IAV-type signals to isolate climate-relevant interannual variability coherence checks providing an external climate context for interpretation. This design links the practical goal of reconstructing monthly global flux fields to the scientific goal of identifying persistent carbon dynamics supporting a more climate-relevant assessment than accuracy metrics alone.

## 1.2 Rational of the Study or Motivation

Global ocean atmosphere $CO_2$ flux is a key component of the carbon budget, yet it cannot be directly observed with global coverage at monthly resolution. While flux estimates are available from biogeochemical models, observation based reconstructions, and atmospheric inversions, they remain constrained by sparse and uneven insitu measurements and by substantial inter product uncertainty and sometimes low resolution. In addition, process based Earth system simulations and large ensemble sensitivity experiments are computationally expensive to run repeatedly at global scale. Therefore, a time aware, multi-source machine learning framework that integrates satellite and reanalysis drivers is needed to reconstruct monthly global $CO_2$ flux fields from indirect environmental controls and to enable basin-scale regime-shift detection and attribution within a physically grounded climate context (e.g., ENSO).

## 1.3 Problem Statement

Global monthly air–sea $CO_2$ fluxes cannot be directly observed with sufficient spatial coverage, and existing biogeochemical, observation-based, and inversion products remain uncertain and often disagree in both magnitude and variability due to sparse measurements and methodological divergence. This thesis addresses that gap by developing a strict time-aware, multi-source machine learning framework to reconstruct $1^\circ \times 1^\circ$ monthly global $CO_2$ fluxes (2014–2024) from indirect physical/biogeochemical drivers and to detect persistent basin-scale regime shifts in detrended variability, interpreted within an ENSO (Ni no3.4) physical coherence context.

## 1.4 Objective

Develop and evaluate a time-aware, multi source machine learning framework to reconstruct global ocean $CO_2$ fluxes and identify basin scale regime shifts with ENSO based physical context to support future early-warning prediction.

- Build a global monthly ML reconstruction of ocean atmosphere $CO_2$ flux (2014–2024) using the observation-constrained Jena CarboScope product as the training target, addressing sparse and nonuniform flux observability.
- Integrate multi-source physical and biological drivers to predict flux from indirect controls by harmonizing satellite, reanalysis, and in-situ informed variables common 1x1° monthly grid.
- Evaluate predictive skill under strict time-aware generalization using LOYO year based windows and standard metrics, comparing raw feature vs anomaly transformed representations to assess added value for flux predictions.
- Detect and characterize basin scale regime shifts in true flux variability by aggregating to RECCAP2 basins and applying change-point testing on deseasoned and detrended IAV to isolate persistent baseline changes from short-term noise.
- Contextualize shifts with ENSO and enable a future early-warning classification setup by using Niño3.4 as a physical-coherence check and defining shift months/lagged drivers as labels/features for potential binary prediction of transitions.

## 1.5 Methodology in Brief

This thesis implements a time-aware, integrated framework to reconstruct global ocean–atmosphere $CO_2$ flux patterns and to assess whether the reconstructed variability exhibits persistent regime shifts within a physically grounded climate context. The approach combines multi-source environmental datasets, supervised machine learning, and statistical change-point analysis, using an observation-constrained flux product as a reference. Monthly predictors representing the coupled physical and biogeochemical controls on air–sea $CO_2$ exchange were compiled from ORAS5 (sea surface temperature, salinity, sea surface height, and mixed-layer depth), ERA5 (10 m zonal and meridional winds), and climate-quality satellite and archived biogeochemical products (chlorophyll-a, phosphate, and net primary productivity proxy). The learning target, co2flux_pattern, was taken from the Jena CarboScope ocean inversion, which integrates SOCAT surface $CO_2$ observations with atmospheric transport modeling and provides a spatially and temporally continuous benchmark for model training and evaluation.

All datasets were harmonized to a common monthly timeline spanning 2014–2024 and regridded to a unified $1^\circ \times 1^\circ$ global grid. Preprocessing included duplicate aggregation, physically consistent interpolation of missing values, and masking of land grid points to retain ocean-only data. Predictors were scaled to stabilize learning across heterogeneous units, the target flux was standardized, and seasonality was encoded using cyclic sine–cosine transforms of month, while latitude and longitude were retained to capture persistent spatial structure. Two important feature configurations were evaluated: a raw-feature representation using absolute driver values and an anomaly-transformed representation output selected predictors as departures

from local gridcell-month climatology to important in all interannual variability. Global $CO_2$ flux reconstruction was formulated as a supervised regression problem at the monthly grid-cell level and a hierarchy of models spanning linear to nonlinear capacities was trained under identical temporal exposure, adding Multiple Linear Regression as a baseline and nonlinear models such as Multilayer Perceptron regression, Random Forests, and XGBoost to capture nonlinear responses and interactions among drivers. Model evaluation employed strict time-aware strategies, including fixed chronological train–validation–test splits and an expanding leave-one-year-out protocol to ensure robust out-of-sample assessment.
Beyond prediction skill, reconstructed the flux fields were examined for instability and regime shifts. Seasonal cycles were removed to form anomaly fields, from which early-warning indicators—rolling variance and lag-1 autocorrelation—were calculated over multiple window size to diagnose changes in variability and persistence. Cause formal regime-shift detection, gridded fluxes were aggregated to basin-scale time series with RECCAP2 masks, deseasoned and detrended to isolate interannual variability, and analyzed with the nonparametric Pettitt test with persistence-window confirmation and predicted basin-scale shifts were at last interpreted with the Niño3.4 ENSO index as a physical-coherence check, linking statistically identified changes in ocean carbon dynamics to large-scale climate variability and helping future early-warning prediction frameworks.

## 1.6 Scopes and Challenges

This thesis is very broad in the study of the CO2 flux variations in the ocean atmosphere. Ocean-Relevant Data-driven, climate-relevant research problem at the intersection of ocean. machine learning, biogeochemistry and climate variability. The work is framed around learning patterns of large-scale carbon exchange patterns to emerge through interaction. physical and biological driving forces, and the representation of those patterns, evalu- coded, and decoded in a meaningful manner used in climate. Beyond creating model outputs, the scope involves creating a consistent research work- flow: knowing what the various flux products are and the reason they vary, iden- understanding what it means by predictability of a coupled climate system, and how. to encode domain knowledge into quantifiable modeling decisions and justifiable sci- entific claims. Another scope is interpretation and not optimization only reat- contributing the rebuilt flux field to the basis of basin-scale reasoning, climate-context dissection, and exploration of the steady developments in variability which can be important to. carbon-budget uncertainty and long-term system behaviour. In that sense, the thesis is intended to give not only reconstruction effort, but also a structured contribution. method of examining the relationship between the indirect environmental controls and carbon flux dy- namics, and the responsible use of the statistical learning at climate diagnostics. and problem formulations of early-warnings in the future.

Several challenges shaped both the scientific framing and the practical execution of this research. The target flux field is not directly observed everywhere and therefore inherits uncertainty and non-uniqueness from the underlying reconstruction pathway; as a result, the model learns consistency with a selected reference product rather than an absolute truth, and the interpretation must remain careful about what is being validated. The predictor information is heterogeneous in origin and quality, requiring strong discipline in preprocessing and consistency to avoid introducing artifacts, while still accepting that measurement error, missingness, and regional reliability vary across space and time. The ocean carbon system is also non-stationary: relationships between drivers and flux can change across years due to evolving forcing and episodic climate events, which makes generalization a scientific difficulty rather than only a modeling weakness and demands time-aware evaluation logic. Additionally, in this s summary metrics and global association statistics provided some limited insight for climatological data because dominant seasonal structure can mask errors in interannual variability the component most relevant for regime-like behavior so interpretation must rely on the diagnostic evidence and it is more important than single headline numbers. At last, we see specific faced real computational constraints: global spatiotemporal datasets produce large memory footprints, long runtimes, and occasional crashes that necessitate chunked computation, careful experiment planning, and output persistence, influencing what could be explored within our realistic hardware limits while still maintaining methodological reproducibility and rigor.

### 1.6.1 Key Learnings and Insights

Across the full research process, the basic important learnings were less about any single model outcome and more about what it takes to research ocean atmosphere $CO_2$ flux as a climate relevant, data driven problem in a strict way. In this first insight was that air–sea $CO_2$ exchange cannot be treated as a simple "assume one variable from others" task, because flux is an emergent quantity shaped by coupled thermodynamics, circulation, gas-transfer physics, and biology; this demands careful thinking about which indirect drivers directly represent those processes, how seasonality and large-scale structure dominate variance, and why anomaly/IAV assumptions are often important to study climate-relevant behavior. And the second learning was that flux products themselves encode calculations and uncertainty, so "ground truth" must be handled cautiously: the thesis must clearly distinguish learning consistency with an observation-constrained reference from claiming absolute physical truth, and evaluation must be designed to test robustness to time, not just fit to a particular sample. Through dataset construction, it also became clear that scientific credibility depends heavily on preprocessing discipline—harmonizing heterogeneous sources, maintaining unit and scaling consistency, preventing leakage when constructing climatologies/anomalies, and documenting transformations so the pipeline is reproducible rather than ad hoc. Another perspective insight was that time is a first-class dimension in climate MAchine Learning: non-stationarity, event years, and evolving forcing mean that generalization must be framed not permanent, and that interpretation should account for changing connections rather than assuming one fixed mapping holds in any situations and always and the regime-shift component reinforced a related conceptual lesson: shifts must be calculated as persistent changes in statistical behavior rather than visually prominent spikes, and cause require signal conditioning (deseasoning/detrending), scale-aware aggregation for interpretability, and confirmation logic to reduce false alarms. Finally, a practical but very important learning was that methodological ambition must be balanced with computational reality: global gridded datasets impose memory and runtime limits that shape experiment design, force careful batching and output persistence, and create reproducibility, logging, and consistent evaluation protocols important to skip broken conclusions.

In the last, these whole thing establish a research critical mindset in which machine learning is used not only for reconstruction, but as a structured framework for integrating multi-source valuable climate information, testing time-aware predictability, and extracting interpretable variability diagnostics within a physically grounded climate context.

# Chapter 2

# Literature Review

The paper improves the estimation of oceanic carbon flux by machine learning, remote. observation, and air-sea analysis of CO2 flux to enhance climate forecasts.

## 2.1 Preliminaries

The carbon dioxide (CO2) uptake by the ocean and its release into the atmosphere is a process. plays a critical role in controlling the system of world carbon and climate. The most important variable that controls this exchange is the carbon dioxide partial pressure of the sea surface. The direction and magnitude of air-sea CO2 flux is determined by (pCO2). The estimation of pCO2 and FCO2 is thus critical when it comes to determining oceanic. plant absorption, regulation, and climate change variability over time. However, spatially and temporally sparse direct in-situ measurements of surface pCO2 are still present. especially in marginal seas and along coastal areas, and therefore difficult to reconstruct with precision. To mitigate the scarcity of the data, in-situ measurements are usually combined with studies. environmental variables derived by satellite like sea surface temperature (SST) and sea. surface salinity (SSS), chlorophyll-a (Chl-a), wind speed, and depth of the mixed-layer. Satellite sensors have extensive spatial resolution, whereas in-situ data (e.g., High accuracy is provided by SOCAT, mooring buoys, ship-based surveys) etc. Matching and preprocessing- steps- Spatio-temporal collocation, detrending, deseasonalization, conversion between fugacity (fCO2) and partial pressure (pCO2) are necessary steps. to ensure data consistency. PCO2 estimation models can be found between simple empirical regressions on one side and sophisticated machine learning (ML) models on the other. Multiple methods are used traditionally. linear or nonlinear regressions which explicitly link physical and biogeochemical drivers with pCO2. More recent literature uses machine learning models like random forests,The exchange of carbon dioxide (CO2) between the ocean and atmosphere plays a play a very vital part in controlling the carbon cycle and climate system in the world. One of the most important variables that define this exchange is the partial pressure of carbon dioxide in the sea surface. The direction and magnitude of air-sea CO2 flux is determined by (pCO2). Proper estimation of pCO 2 and FCO 2 is thus critical in the interpretation of oceanic. carbon uptake, feedback mechanisms and long term climate variability. However, spatially and temporally-sparse direct in-situ measurements of surface pCO2, especially where sea level is low, and, therefore, difficult to accurately reconstruct. In order to overcome the problem of data scarcity, in-situ measurements are usually combined with studies. environmental data obtained through satellites like sea surface temperature (SST), sea surface salinity (SSS), chlorophyll-a (Chl-a), wind speed and mixed-layer depth. Satellite data offers a wide geographical coverage, whereas in-situ data (e.g., SOCAT, mooring buoys, ship-based surveys) are very accurate. Matching and preprocessing procedures include spatiotemporal collocation, detrending, deseasonalization and the transformation of fugacity (fCO2) into partial pressure (pCO2), which are critical. to ensure data consistency. pCO2 estimation models go between simple empirical regressions, and modern machine learning (ML) models. There are several traditional techniques. linear or nonlinear regressions directly linking pCO2 to physical and biogeochemical drivers. More recent researches include the use of machine learning algorithms like In order to capture nonlinear, they use random forests, kernel ridge regression, and neural networks. contacts and geographical diversity. Evaluation of the model normally depends on measures. as root mean square error (RMSE), mean absolute error (MAE) and coefficient. of determination (R2 ), frequently cross-validation techniques and neural network to model nonlinear. relationships and geographical disparity. Model assessment is normally based on metrics. as root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R2), typically by cross-validation.

Beyond mean-state estimation, understanding interannual variability and regime shifts in air–sea $CO_2$ fluxes has gained increasing attention. These nonlinear transitions, influenced by climate modes such as ENSO and long-term warming, challenge conventional modeling frameworks and motivate the integration of ML, data assimilation, and regime-detection techniques in ocean carbon research.

## 2.2 Review of Existing Research

### 2.2.1 Prediction of $CO_2$ Flux

The paper addresses the challenge of estimating sea surface partial pressure of carbon dioxide ($pCO_2$) and air–sea $CO_2$ flux ($FCO_2$) in the East China Sea (ECS) using limited in-situ observations and well-informed models. Improve understanding of the dynamics of surface $pCO_2$. Develop a regional Multiple Non-Linear Regression (MNR) model to estimate $pCO_2$ based on observed and satellite-derived environmental variables. Assess seasonal and interannual variability of $pCO_2$ and $FCO_2$ over the ECS. We have developed an MNR model to estimate the sea surface $pCO_2$ for the ECS using multiple observed and matched satellite data. Combinations of SST, SSS, MODIS-derived Chl a, JD, longitude, and latitude were tested to obtain the optimal structure of the MNR model. In-Situ Data Hydrographic and carbonate system observations from SOCAT (2006–2016). Data from cruises in the ECS (2013–2018). Sea surface temperature (SST), salinity (SSS), chloro
phyll-a (Chl-a), and carbonate system data. Satellite Data MODIS Aqua Level-3

chlorophyll-a products (2003–2019). Model Outputs Changjiang Biology Finite-Volume Coastal Ocean Model (FVCOM). SOCAT Data: Available online at SOCAT. MODIS Aqua Data: Available from NASA's Goddard Space Flight Center [1]. FVCOM Outputs: Accessible from Changjiang Biology FVCOM. Sea Surface Temperature (SST), Sea Surface Salinity (SSS), Chlorophyll-a (Chl-a) concentrations. Julian Day (JD) represents seasonal cycles. Longitude and Latitude to account for spatial variability. Field Data Processing Conversion of fugacity ($fCO_2$) to partial pressure ($pCO_2$). Correction for in-situ SST and other environmental factors using CO2SYS.XLS. Satellite Data Match MODIS Aqua data matched with in-situ observations within a 3×3 pixel box. Cross-Validation 10-Fold Cross-Validation to optimize model parameters and validate performance. Data Aggregation Seasonal and interannual variability analysis of $pCO_2$ and $FCO_2$ using model outputs. Root Mean Square Error (RMSE), Mean Absolute Error (MAE) Coefficient of Determination ($R^2$). Incorporate Machine Learning Explore advanced machine learning models (e.g., random forests, neural networks) for better handling of nonlinearities and interactions. Utilize Higher-Resolution Data Improve SSS data resolution using newer satellite missions like SMOS or SMAP. Address Domain I Variability Develop submodels or integrate additional variables (e.g., sediment input, river discharge rates) to improve performance in Domain I. Open Source Code Release the code and data preprocessing pipelines to improve reproducibility and enable broader adoption [10].

This Nature Communications Earth & Environment paper covers both regime shifts and $CO_2$ fluxes in ocean contexts. Titled "Machine learning reveals regime shifts in future ocean $CO_2$ fluxes," it uses machine learning to identify abrupt changes (regime shifts) in projected ocean $CO_2$ flux variability. The study analyzes future scenarios, revealing shifts in $CO_2$ flux patterns driven by factors like ENSO, Southern Ocean upwelling, salinity, phosphate, and winds. It highlights potential non-linear changes in interannual variability (IAV) of air–sea $CO_2$ exchanges. .This study addresses the problem of understanding and reconstructing inter-annual variability (IAV) of global ocean–atmosphere $CO_2$ fluxes and identifying regime shifts that are poorly represented in Earth System Models (ESMs). The authors use model-simulated data from CMIP6, covering five ESMs and four scenarios (preindustrial, historical, SSP126, SSP585). Data include monthly surface variables such as $CO_2$ flux (target), dissolved inorganic carbon (DIC), alkalinity (ALK), SST, salinity, phosphate, wind speed, sea-ice concentration, and atmospheric $CO_2$. They apply kernel ridge regression (KRR) with an RBF kernel, combined with permutation feature importance, clustering, and classification. In the preprocessing helping bootstrapping,detrending, normalization,deseasonalization, and detrending. Model performance is evaluated with a modified $R^2$ focused on IAV. Results highlight DIC and ALK as dominant drivers, with shifts under warming scenarios. In this model limitations add reliance on ESM outputs, reduced performance in complex regions (e.g., Southern Ocean), and correlation-based inference [22].

In this paper challenge of accurately predicting the partial pressure of carbon dioxide ($pCO_2$) in the Bay of Bengal (BoB). Presenting models can not restore to capture the region's complex spatiotemporal dynamics due to factors like stratification, monsoonal influences, and limited data availability. The study develops a Multi Parametric Regional Regression (MPRR) algorithm that integrates in-situ and satellite data to improve the accuracy of $pCO_2$ modeling. Insitue and satelight data SST,SSS,Partial pressure $co_2$, MPRR MODEL, Satellite SSS, SST,chla. In-situ Data: National Centers for Environmental Information (NCEI) (NOAA). Bay of Bengal Ocean Acidification (BOBOA) mooring buoy. Field campaign by the National Remote Sensing Centre (NRSC). Satellite Data: Moderate Resolution Imaging Spectroradiometer (MODIS-Aqua): Sea Surface Temperature (SST) and Chlorophyll-a (Chla) (4 km resolution). Multi-Mission Optimally Interpolated Sea Surface Salinity (MMOI-SSS): Sea Surface Salinity (SSS) (25 km resolution). In-situ Data: Available from NOAA's Ocean Carbon Data System and NRSC's field observations. Satellite Data: MODIS-Aqua products accessed from NASA's Ocean Color website. MMOI-SSS data from Jet Propulsion Laboratory's PO.DAAC portal. Physical and Biogeochemical Parameters: Sea Surface Temperature (SST). Sea Surface Salinity (SSS). Chlorophyll-a concentration (Chla). $pCO_2$ (derived from $fCO_2$ using a formula). Data Preprocessing: Spatiotemporal averaging and matchups of in-situ and satellite data. Conversion of $fCO_2$ to $pCO_2$ using temperature-corrected equations. Model Development: A Multi Parametric Regional Regression (MPRR) model was developed, linking $pCO_2$ to SST, SSS, and Chla. Regional coefficients were optimized for better accuracy. Validation: Independent datasets were used to validate the model, including direct in-situ and satellite-derived $pCO_2$ estimates.Validation Metrics for MPRR Model: Mean Relative Error (MRE): 0.012 (in-situ), 0.020 (satellite-derived). Mean Normalized Bias (MNB): 0.022 (in-situ), 0.032 (satellite-derived). Root Mean Square Error (RMSE): 4.75 $\mu$atm (in-situ), 4.94 $\mu$atm (satellite-derived). Correlation Coefficient ($R^2$): 0.92 (in-situ), 0.91 (satellite-derived). Comparison with Other Models: MRE: 0.064–0.146. RMSE: 7.64–14.26. $R^2$: 0.71–0.85 [14]

It proposes a new method for estimating $CO_2$ flux using satellite observations and deep learning models. The method combines two models: Concentration Correction Model: It adjusts $CO_2$ concentration data from a chemical transport model (GEOS-Chem) using satellite data from the Orbiting Carbon Observatory-2 (OCO-2). Concentration-Flux Inversion Model: This model links $CO_2$ concentrations to fluxes and helps in estimating the actual $CO_2$ fluxes. Data type is in situ , model-based data , sattelite data . Weather information (temperature, wind, etc.) from MERRA-2. From ODIAC dataset, providing information on $CO_2$ emissions from human activities like burning coal and oil.From GFED4.1s dataset, estimating $CO_2$ emissions from fires. From JMA Ocean $CO_2$ Map dataset, showing how much $CO_2$ is absorbed by the ocean. From CASA-GFED dataset, representing $CO_2$ exchange with plants and soil. Satellite Data: OCO-2, Ground-Based Measurements:TCCON, Other Carbon Inversion Systems: CAMS, CT2022.Ocean $CO_2$ flux data.The features are 47 levels each for temperature profile, humidity profile, and meridional and zonal wind profile, as well as one layer for simulated $XCO_2$, resulting in a total of 189 neurons. Results showed that both deep learning models showed excellent prediction performance, with a mean bias of 0.461 ppm for the concentration correction model and an annual mean correlation coefficient of 0.920 for the concentration–flux inversion model.[23]

This study addresses the problem of objectively detecting and tracking marine carbon biomes to understand changes in air–sea $CO_2$ fluxes under climate variability.

The authors used global ocean biogeochemistry model outputs (ORCA025-MOPS) from 1958–2018, including monthly sea-surface $fCO_2$, SST, dissolved inorganic carbon (DIC), and alkalinity (ALK). First, they constructed local multivariate linear regressions (2°×2° grids) to capture target–driver relationships between $fCO_2$ and its drivers. These regression coefficients were then clustered using adaptive agglomerative hierarchical clustering, enabling data-driven biome detection without predefined regions. For temporal tracking, a feed-forward neural network (FNN) classified biomes across time. Model performance was evaluated using accuracy, precision, and recall (∼95%). Preprocessing included outlier removal, normalization of regression coefficients, and feature standardization. Limitations include reliance on model-simulated data, linear assumptions in local regressions, and exclusion of some drivers (e.g., salinity, biology). The dataset comes from the ORCA025-MOPS global ocean biogeochemical model, which covers the entire world ocean at approximately 0.25° spatial resolution and spans 1958–2018. All major ocean basins are included. Although the data are global, the authors perform regional analyses (e.g., the North Atlantic and the Southern Ocean) to better interpret and explain the global results. These regions are used for analysis and discussion, not because the dataset itself is regional. Future improvements could integrate observational datasets, nonlinear regressions, uncertainty quantification, and explainable AI methods. [15]

The paper demonstrates that machine learning (kernel ridge regression) can reconstruct current and future $CO_2$ flux variability using ESM data, identifying key drivers and their changing roles under climate change scenarios. The data type is In situ and this data source is ACCESS-ESM1-5. Data were collected from sediment traps with polyacrylamide gel layers to capture particles, which were then imaged using stereomicroscopy. Regions of interest (ROIs) were extracted for particle classification. Data main feature is Alkalinity (ALK). Mean Absolute Error (MAE): Compared model predictions of carbon flux to: Human annotations Bulk carbon measurements. Precision and Recall: Assessed classification accuracy for particle types. Variance and Flux Profiles: Compared estimated carbon flux across depths. The methodology includes human-in-the-loop scripts and ResNet-18 implementations. Code specifics were not explicitly provided in the text but can be inferred as tailored image processing pipelines and flux calculation scripts. [16]

The researchers addressed the challenge of estimating global surface-ocean CO partial pressure fields and sea-air CO flux variability using the sparse SOCAT v1.5 database ( 6 million surface pCO data points until 2007) combined with atmospheric CO mixing ratios from monitoring stations. Their approach employed a diagnostic mixed-layer scheme integrating Bayesian inversion with ocean biogeochemistry parameterizations, including sea-air gas exchange, carbonate chemistry, and mixed-layer carbon budgets that utilize spatial/temporal correlations for data interpolation. The model incorporated mixed-layer DIC concentration, temperature/alkalinity effects, wind speed, solubility changes, and ocean-internal carbon sources/sinks as adjustable parameters. Key limitations include sparse data coverage, arbitrary dampening factors for interannual variations, and reliance on smoothness assumptions for regularization. Potential improvements could incorporate oxygen measurements, satellite ocean color data, multi-linear regression approaches, and neural networks for enhanced process parameterization, though no specific code repository was mentioned in their work.[27]

The paper addresses how ENSO (El Niño–Southern Oscillation) influences sea–air CO flux variability in the equatorial Pacific and how this relationship may change under a high-warming future scenario. The authors use observational datasets (SST, CO flux, DIC, alkalinity) and CMIP6 Earth System Model outputs, all at monthly resolution on gridded spatio-temporal data. The study relies on process-based Earth system modeling and statistical analysis, not ML; ENSO is quantified using the Niño3.4 index, and anomalies are derived via spline detrending. Preprocessing includes regridding, vertical interpolation, deseasonalization, and separation of thermal vs non-thermal pCO components. Model performance is evaluated using correlations, variance, and spatial pattern comparisons against observations. Results show that while many models reproduce historical ENSO–CO relationships, about half project a future reversal, linked to buffering capacity and DIC dynamics. Limitations include reliance on a single high-emission scenario, model mean-state biases, and sparse subsurface observations. Improvements could involve multi-scenario analysis, better carbonate chemistry constraints, and open, standardized analysis code (currently not fully provided).[9]

Zeng et al. addressed the critical problem of accurately estimating oceanic CO trends when data scarcity forces models to pool multi-year observations. The researchers used SOCAT version 2021 database (273,434 data points from 1980-2020) alongside satellite-derived predictors including sea surface temperature, salinity, chlorophyll-a, and mixed layer depth. They deployed three machine learning algorithms—Random Forest (RF), Gradient Boost Machine (GBM), and Feedforward Neural Network (FNN)—employing a novel variable-rate normalization method instead of constant trends. Key improvements include: decadal-scale trend extraction reducing bias in early/late model periods, and leave-one-year-out validation achieving $R^2$=0.718 (RF). Their model estimated oceanic sink increased from 1.79±0.47 PgC/yr (1980s) to 2.58±0.20 PgC/yr (2010s). Limitations: Missing coastal areas ( 0.11 GtC/yr), limited pre-1990 data reliability, and assumptions about post-2015 rates. Potential improvements: incorporating physics-informed neural networks, attention mechanisms for temporal dependencies, or ensemble deep learning architectures with uncertainty quantification. Code available upon request from authors.[13]

The paper addresses equifinality and regularization bias in hybrid modeling—where multiple parameter combinations describe data equally well, undermining interpretability. They propose Double Machine Learning (DML)-based hybrid modeling for causal effect estimation in Earth sciences. Data: FLUXNET2015 eddy covariance measurements from 36 sites, measuring CO fluxes, temperature, radiation, and meteorological variables. Synthetic data generated from Q10 models for controlled validation. Algorithm: DML framework with two-stage estimation: (1) fit non-parametric models (neural networks/random forests) for confounders, (2) estimate causal treatment effects on residuals. Applied to Q10 temperature sensitivity and carbon flux partitioning problems. Improvements: The work could extend beyond partially linear models, incorporate unobserved confounders using sensitivity analysis, and scale to real-time applications. The computational inefficiency of multi-stage fitting versus end-to-end learning needs addressing. Limited validation on only one site (Neustift) for Q10 restricts generalizability claims.[17]

### 2.2.2 Model Derive Data Based prediction

The study addresses how to correctly estimate long-term air quality trends attributable to anthropogenic emission changes while accounting for meteorological variability. It uses simulated and observational air quality data for $PM_{2.5}$ and $O_3$ over the US (2011–2017) and China (2013–2017). The main data sources are GEOS-Chem chemical transport model simulations, surface monitoring networks, and MERRA-2 meteorological reanalysis dataThe authors evaluate statistical and machine learning algorithms, including multiple linear regression (MLR), generalized additive models (GAM), LASSO, and random forest (RF). Model performance is measured using trend estimation error compared to counterfactual simulations with constant meteorology. Results show that traditional MLR and GAM perform poorly, while random forest with local and regional meteorological features reduces errors by 30–42%.Limitations include incomplete separation of emission–meteorology interactions and dependence on model-based simulations. Improvements could include physics-informed ML, causal modeling, and better representation of emission–weather feedbacks. For understanding It uses both global and regional data, but the analysis focus is regional. Global data: The GEOS-Chem chemical transport model is global in structure and uses global meteorological reanalysis (MERRA-2) and global-scale emissions as boundary conditions. Regional data: The core experiments are regional (nested) simulations over the United States and China, using higher-resolution grids, regional emission inventories (NEI for the US, MEIC for China), and regional air-quality monitoring data. So, the study is globally driven but regionally analyzed acp-22-10551-2022 (1) Sources method and main purpose Main purpose: To accurately separate emission-driven air quality trends from meteorological effects, and to evaluate how well commonly used statistical and machine learning methods can correct for weather variability when assessing long-term trends in $PM_{2.5}$ and $O_3$. Method: The study uses GEOS-Chem chemical transport model simulations to create two scenarios: Observational scenarios (changing emissions + changing meteorology) Counterfactual scenarios (changing emissions + constant meteorology) Statistical and machine learning models—MLR, GAM, LASSO, and Random Forest (local and regional)—are applied to meteorology-correct observed trends. Model performance is evaluated by comparing corrected trends against counterfactual (emission-only) trends, identifying which methods best recover true emission impacts. **meteorological effects**, what is this?? Meteorological effects mean the influence of weather and atmospheric conditions on air pollution levels, independent of human emissions. In simple terms: Even if emissions stay the same, the weather can make pollution look higher or lower. Examples of meteorological effects: Temperature → affects ozone ($O_3$) formation Wind speed & direction → disperses or accumulates pollutants Rainfall → washes out $PM_{2.5}$ Boundary layer height → controls how concentrated pollutants become Humidity & pressure → influence chemical reactions in the air So in this paper, meteorological effects refer to changes in air pollutant concentrations caused by weather variability, not by emission reductions. The study tries to remove (correct for) these weather effects to reveal the true impact of emission changes. which model best and poor .and how can they solve?? Best-performing model: Random Forest with local + regional meteorological features (RF-regional) It best captures non-linear relationships, feature interactions, and regional transport effects. It reduced trend-estimation errors by ∼30–42% compared to traditional methods. Poor-performing models: Multiple Linear Regression (MLR) → worst overall Generalized Additive Models (GAM) → slightly better than MLR but still weak These models assume simple or smooth relationships and fail to capture complex meteorology–pollution interactions. How they solved the problem: Created counterfactual simulations using GEOS-Chem with constant meteorology to represent "true" emission-driven trends. Applied different statistical/ML models to meteorology-correct trends. Compared corrected trends to counterfactual trends to objectively measure error. Used RF-regional models to incorporate nonlinearity and regional weather influence, achieving the most accurate correction. [11]

The researchers developed hybrid machine learning data assimilation methods for marine biogeochemistry to update unobserved variables (like nitrate) using only total chlorophyll observations, addressing computational limitations of traditional ensemble-based approaches. Marine biogeochemical data from L4 location (2000–2014), including chlorophyll concentrations, nitrate, temperature, and other pelagic variables from 100-member ensemble runs with weekly sampling intervals. ML-OI (machine learning optimal interpolation) estimates correlations, and ML-EtE (end-to-end) directly predicts analysis increments. Both use fully connected artificial neural networks optimized via AutoKeras. Synthetic observations with 10% uncertainty, ensemble free-runs covering training period. Standardization to unit variance/zero mean, normalization by climatological maximum values. RMSE for error assessment, SHAP analysis for feature importance. Heavy reliance on existing DA systems for training data, dependence on climatological statistics, and need for robust reanalysis datasets.[27]

The study addresses how ENSO modulates interannual variability of tropical Pacific air–sea CO flux and why some CMIP5 Earth System Models fail to reproduce observed responses. It uses model-simulated spatiotemporal climate–biogeochemical data (1960–2000) from 14 CMIP5 models, evaluated against observational and reanalysis datasets (HadISST SST, SOCAT-based pCO/fCO products, wind reanalysis).The methodology is physics-based statistical analysis, relying on standardized EOFs, regression, and correlation analysis, rather than machine learning. Key features include SST, DIC, TALK, wind speed, and biological productivity, with preprocessing via detrending, deseasonalization, interpolation, and anomaly computation. Performance is assessed using regression coefficients, correlations, variance explained, and ENSO consistency, not predictive metrics. Limitations include model biases in upwelling–DIC coupling, coarse resolution, and weak biogeochemical realism. Improvements could involve higher-resolution models, better ocean biogeochemistry, data assimilation, or ML-based bias correction.[6]

### 2.2.3 Regime Shift Detection

The paper addresses the challenge of accurately estimating global carbon fluxes (terrestrial and oceanic carbon sinks) by combining in situ (ground-based) and satellite $CO_2$ observation data. It tackles the issue of heterogeneity between these data sources, which has historically led to inconsistent or less accurate results when using only single-source observations. The authors developed the Multi-observation Carbon Assimilation System (MCAS), which modifies the Ensemble Kalman Filter (EnKF) to balance the observational errors from different data types and achieve more precise global carbon flux estimates. Solves the heterogeneity problem using a mathematical method called the modified Ensemble Kalman Filter (EnKF). This approach applies different adjustments (called inflation factors) to each data source to balance their differences. In situ $CO_2$ Data: ObsPack: obspack_co2_1_GLOBALVIEWplus_v9.0_2023-09-09.Available on NOAA's Global Monitoring Laboratory website. OCO-2 data: Accessible via NASA's Earth Science Data Systems. This dataset contains measurements from 74 laboratories across 28 countries. Satellite $CO_2$ Data: Orbiting Carbon Observatory-2 (OCO-2): Level 2 bias-corrected $XCO_2$ retrievals. Other Reference Data (Validation): Total Carbon Column Observing Network (TCCON): A ground-based network measuring $CO_2$ column averages. Prior Carbon Flux Data: Fossil fuel emissions from CDIAC and ODIAC. Wildfire emissions from GFED 4.1s and GFED_CMS. Ocean flux estimates from $pCO_2$-Clim and Ocean Inversion Fluxes (OIF). Terrestrial biosphere fluxes from CASA-based models. ObsPack data: Available at tccondata.org. Prior Flux Data: Available through repositories like CarbonTracker.Multi-source $CO_2$ Observations: Combines ground-based (high accuracy) and satellite (broad coverage) data. Spatiotemporal Resolution: Carbon flux estimates with $1^\circ \times 1^\circ$ resolution. Temporal granularity of 3-hourly flux updates. Dynamic Inflation Factors: Used to handle heterogeneity in observational errors. Validation Using Independent Datasets: ObsPack and TCCON for evaluation. Preprocessing: ObsPack and OCO-2 data were downsampled to match the resolution of the GEOS-Chem transport model. Observation thinning was applied to reduce computational costs. Assimilation Framework: A modified Ensemble Kalman Filter (EnKF). Introduced dynamic inflation factors for observational and forecast errors. Implemented global localization by dividing Earth into seven regions to account for regional differences in observations. Carbon Flux Inversion: Joint assimilation of in situ and satellite data. Utilized prior fluxes from fossil fuel, wildfire, terrestrial biosphere, and ocean models. Root Mean Square Error (RMSE): Quantifies the difference between modeled and observed $CO_2$ concentrations. Correlation Coefficient (CORR): Measures the agreement between predictions and observations. Bias: Quantifies systematic differences between predictions and observations and regional Sparsity of Observations: Some regions, like Africa and South America, have limited in situ observation coverage. Additionally,Satellite Data Limitations: Satellites like OCO-2 are less sensitive to near-surface $CO_2$ emissions, which affects accuracy in the lower atmosphere. High-latitude regions (e.g., boreal zones) face constraints due to limited satellite coverage.In the dependence on Prior Models: Results are influenced by the selection of prior flux calculation, to variability in the outcomes. In the computational complexity: High-dimensional data assimilation with the dynamic inflation increases in computational costs. Validation Constraints: Validation datasets, such as TCCON, are geographically sparse and concentrated in the developed regions (e.g., North America and Europe) [18].

## 2.3 Summary of Key Findings

Machine learning (ML) introduces have a significant effect in the research of oceanic carbon flux estimation and ocean acidification observing.Recent studies such as *Detection and Tracking of Carbon Biomes via Integrated Machine Learning* (Mohanty et al., 2024) mentions how ML models increase $CO_2$ flux calculation and biogeochemical trend calculation. From the results we could see that XGBoost and artificial neural networks (ANN) models can have comparatively better performance than traditional regression methods in terms of predicting $pCO_2$. However, handling sparse data still remains a challenge. Integrating observationally constrained statistical models with ML Techniques may improve further carbon flux trend predictions. Few more challenges need to be considered like the high spatial and temporal variability in ocean carbon fluxes for models to perform accurately. A research on ocean acidification, *A Mapped Dataset of Surface Ocean Acidification Indicators* (Sharp et al., 2024) focuses on substantial regional variations, specifically in the coastal and Arctic regions integrating RFR-LME model. In other approaches, whereas ensemble neural networks and XGBoost models capture high-resolution variability in sea-air $CO_2$ fluxes, the struggle remains with wind-driven events. Besides the challenges in the mentioned approaches, a notable progress has been made in $pCO_2$ and air-sea $CO_2$ flux estimation in the mentioned hybrid modeling suggested in the papers such as studies such as *Explicit Physical Knowledge in Machine Learning for Ocean Carbon Flux Reconstruction* (Bennington et al., 2022) and *Advancements in Carbon Dioxide Modeling in the Bay of Bengal* (Shaik, 2023). The hybrid ML-regression models significantly improved in $pCO_2$ estimations. A study about the Mediterranean Sea found that XGBoost-based $pCO_2$ estimation showed a good result though limited by small datasets. For tackling the limitation, expansion in observational data sources still remains crucial. The Ocean Carbon and Acidification Data System (OCADS) plays an important role in ensuring long-term data preservation. However, inconsistency in submitting data creates obstacles for the machine learning approaches. Also, Coupled Model Intercomparison Project Phase 6 (CMIP6) which provides ocean carbon sink simulations, reveals significant internal variability and model biases in ocean carbon sink estimates. The need for larger datasets and advanced multi-model comparison 14 processes emerges from the informations to improve estimation and minimize uncertainties in ocean carbon cycles. Regionally and seasonally different data shows discrepancies between observation-based and model-based estimates. A study on the East China Sea and Northeast Pacific highlighted this issue of dependence of $CO_2$ fluxes on regional oceanographic conditions. For further improvement of ocean carbon cycle predictions, it is necessary to modify the biogeochemical models. Future research could emphasize integrating satellite, in situ, and model-based data for better assessment along with the expansion of observational networks which could include BCG-Argo floats and real-time data assimilation. For further improvement in real-time flux assessments, AI-driven carbon monitoring, especially deep learning can improve real-time flux assessments. Additionally, the integration of high-resolution satellite observations, advanced machine learning models and incorporating ocean interior carbon data alongside Total Matrix Intercomparison (TMI) methods may increase the accuracy of anthropogenic carbon estimates resulting in more precise climate trend predictions.

# Chapter 3

# Requirements, Impacts and Constraints

## 3.1 Final Specifications and Requirements

This thesis requires a research environment capable of handling large spatiotemporal gridded datasets and running iterative machine learning experiments without data loss or memory failure. Primary data sources are provided in NetCDF and CSV formats; therefore, a NetCDF inspection tool (e.g., Panoply) is required to verify variable names, dimensions, coordinate conventions (lat–lon order, time encoding), missing-value flags, and basic spatial/temporal coverage before ingestion. Implementation requires a Python development environment (Jupyter Notebook or local IDE such as VS Code or PyCharm) with core scientific libraries for array and tabular processing (NumPy, Pandas, xarray, netCDF4), machine learning (scikit-learn; optional deep learning via PyTorch), and visualization (Matplotlib). Because the workflow involves global 1° monthly grids across multiple predictors, high RAM capacity is essential. Practical execution benefits from GPU-enabled environments for any neural model (even though not implemented in this research) and from memory-safe processing strategies (chunked reading, on-disk caching, and saving intermediate outputs). Stable storage is required to persist merged datasets, trained model artifacts, prediction files, and evaluation summaries to avoid recomputation after kernel restarts. Lastly, in order for the experiments to be reproducible, the random seeds, package versions and the directory structure for the logs, inputs and outputs need to be organized.

## 3.2 Societal Impact

The study relates to a social demand as it will enhance our knowledge about the processes of ocean carbon that directly can impact climate policy and mitigation measures. The machine learning framework that can be created in this case allows better reconstruction of global ocean atmosphere CO fluxes, which can help in making evidence-based decisions related to climate action. This work makes a contribution to international carbon budget accounting by offering reliable estimates of the flux at monthly resolution, which is crucially needed to track the progress of the emission reduction targets of international climate change frameworks like the Paris Agreement. The early warning of changes in the ocean carbon sink provided by the regime-shift detection capability has far-reaching consequences to the coastal societies, fisheries, and maritime economies that rely on the ocean conditions. Moreover, ENSO perspective allows foreseeing the risks in the climate in the vulnerable area, especially in developing countries where the impact of extreme weather associated with an El Niño cycle is disproportionate in terms of harm to food security, water supplies, and human health. The fact that the framework relies on free access to satellite data and reanalysis data makes the framework accessible and democratizes the ability of climate science across institutions with small computational resources, encouraging global scientific cooperation and capacity-building in climate studies.

## 3.3 Environmental Impact

The environmental impact of the research is automatically favorable as it will serve to develop a scientific potential in terms of monitoring and in order to appreciate the significance of the ocean in the control of the climate system of the earth. This study enhances the ability to identify long-term dynamics in ocean carbon uptake as well as the process of flux reconstruction itself, complementing the process of identifying the early signs of stress in ecosystems, such as ocean acidification, deoxygenation, which endangers marine life and marine services. Even the computational method itself has a relatively small environmental impact compared to conventional techniques like machine learning models, after training, have much lower computational cost than running repeated global-scale simulations of an Earth system or carrying out large-scale observational campaigns in the form of ship expeditions. Nonetheless, the training step does use resources of energy, and responsible introduction ought to focus on effective optimization of codes and, when possible, use computing resources that run on renewable energy. The results of the research have a direct impact on the environmental sustainability, which is the improvement of our ability to determine whether the ocean carbon sink is becoming fragile under the influence of climate change that aids in the conservation of marine and protected areas and the sustainable management of ocean resources. In the end, the improved knowledge of air-sea CO exchange processes will help to project climatic conditions more accurately, which is the key to effective environmental policies and adaptation strategy to avoid the destruction of ecosystems and the reduction of planetary limits.

## 3.4 Ethical Issues

A few ethical implications arise out of this study, which are primarily data stewardship, transparency in science, and responsible expression of uncertainty. To start with, the utilization of observation-constrained products such as Jena CarboScope or publicly funded data (ORAS5, ERA5, satellite products) comes with a social responsibility to the adequate credit data providers and upholding open principles of science in general by distributing methodologies and, where feasible, model results to the wider research community. Second, machine learning models by nature represent a risk of overfitting or generating a physically impossible prediction; the ethical requirement in this case is to strictly test and report outputs to existing oceanographic theory and to communicate the limitations of the model to avoid misuse in policy. Third, regime shifts should be found and interpreted rigorously and without resort to alarmism, but with truthfulness about the implication of the perceived changes to carbon-budget uncertainty and climate paths. Ethically, there is a need to report results in a balanced manner, being sensitive to inter-method differences and the tentative quality of machine-learning reconstructions alongside process-based knowledge. Moreover, since this framework may have direct implications on climate policy that affect vulnerable people, researchers will have to make sure that uncertainty is clearly measured and that early-warning capacities should not be exaggerated farther than the data and methods can be reliably justified. Professional responsibility is also concerned with reflecting on dual-use issues: although this study is aimed at climate science, one should be careful not to misuse the methodologies in order to achieve the goals that are considered to negatively affect the climate monitoring work or climate action. Lastly, fair access to research products is an ethical consideration, which means that the results are available to the scientific and policy communities of both developed and developing countries without generating knowledge gaps that would benefit climatically exposed areas to a disadvantage.

## 3.5 Research Management Plan

### 3.5.1 Phase 1: Topic Selection and Pre-Thesis-1 (November 2024 - January 2025)

Focused on building machine learning as the main research direction and finalizing a thesis scope. During this stage, we studied ML fundamentals, reviewed literature, and prepared the Pre Thesis 1 paper. The main outcome were the Pre Thesis 1 report, a structured literature review summary, and an initial shortlist of candidate models with an evaluation plan.

### 3.5.2 Phase 2: $CO_2$ Flux prediction using machine learning (February 2025 - October 2025)

focusing on building the global $CO_2$ flux prediction framework using direct and indirect environmental drivers of $pCO_2$. This phase included dataset selection and integration (drivers and target flux), preprocessing and feature engineering, and model selection with testing of multiple methodologies and baseline/nonlinear models. We implemented strict Leave-One-Year-Out (LOYO) evaluation to ensure time-aware validation. The deliverables were the finalized dataset, a reproducible modeling workflow, and prediction results supported by metrics and plots.

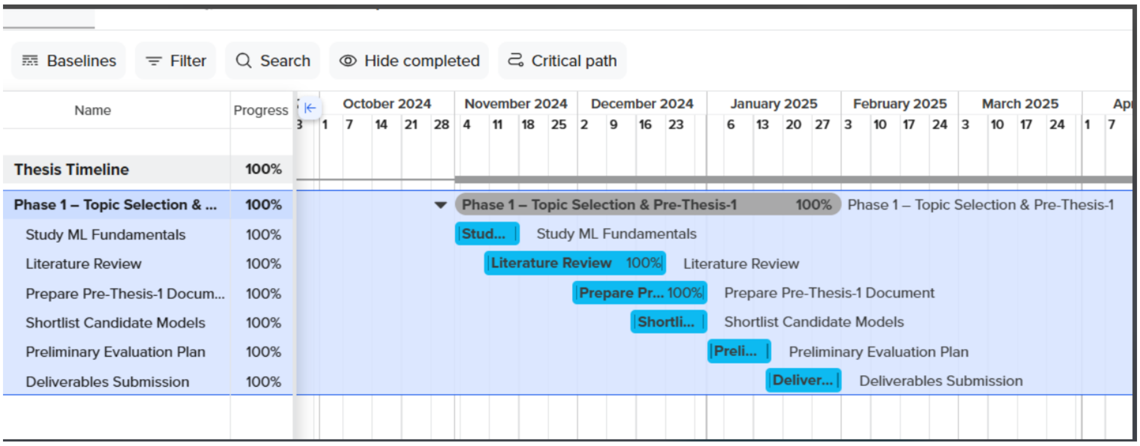



Figure 3.1: Gantt chart for Phase 1 showing the timeline and completion of research tasks from November 2024 to January 2025.

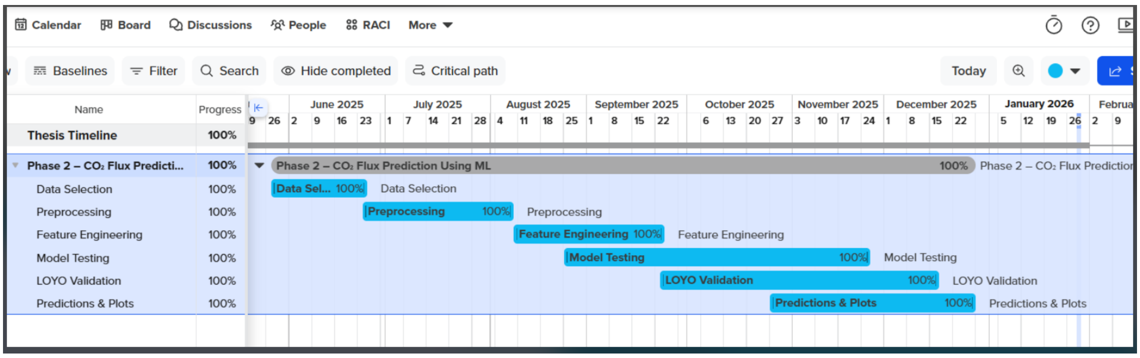



Figure 3.2: Gantt chart for Phase 2 ($CO_2$ Flux Prediction) outlining research timeline and task completion from February to October 2025.

### 3.5.3 Phase 3: Regime shift detection and early warning analysis (November 2025 -January-2026)

Extended the work from prediction to a hybrid framework by adding regime shift detection and early-warning interpretation, while completing the thesis for submission. Key work included updating the scope/title, refining and fixing models, basin-level aggregation, IAV preparation (deseasonating and detrending), Pettitt change-point detection, ENSO (Niño3.4) coherence checks, and early-warning indicators. The deliverables were confirmed regime-shift results with supporting indicators/plots, a revised and polished final manuscript, and the completed thesis formatted and submitted according to requirements. Deliverables: detected regime shifts + supporting indicators/plots, final thesis report aligned with formatting requirements.

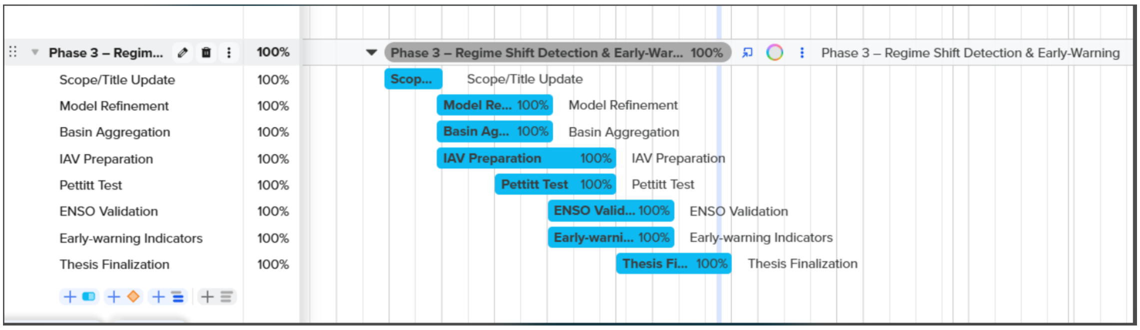



Figure 3.3: Gantt chart for Phase 3 (Regime Shift Detection) outlining final analysis and thesis submission timeline from November 2025 to January 2026.

## 3.6 Risk Management

This study has a number of main threats which need mitigation measures. Satellite and reanalysis interruptions that pose a risk to the availability of data are overcome by redundancy and the use of local backup. The small 10-year time range is addressed by model overfitting through a rigorous technique of leave-one-year-out cross-validation and regularization. The risks of statistical mappings which include physical inconsistency are addressed using physics-informed constraints and testing against known ocean dynamics. False positives to regime-shift changes are reduced by applying conservative significance levels and coherence tests with other independent climatic indices such as ENSO. The issue of the lack of computational resources is taken care of by optimizing the code and choosing the efficient algorithm. Risks of interpretation when attributing the patterns of flux to physical drivers are overcome through explainable AI tools and analysis of feature importance. The version-controlled code repositories and detailed documentation ensure the reproducibility. Lastly, translation hazards between policy applications and scientific outputs are reduced through clear unsafe communication and basin-scale aggregations within decision-specific requirements. It is the framework robustness and credibility that is provided by continuous monitoring and adaptive strategies throughout the lifecycle of the research.

# Chapter 4

# Proposed Methodology

## 4.1 Design Process or Methodology Overview

This study follows a structured, multi-stage methodology designed to (i) construct a physically consistent, machine-learning-ready dataset of global ocean–atmosphere CO flux drivers, (ii) evaluate the skill of linear and nonlinear models in predicting gridded ocean CO flux patterns under strict temporal generalization, and (iii) assess whether the predicted flux variability exhibits early-warning signals and statistically robust regime shifts, with physical contextualization using ENSO.

### 4.1.1 Data processing and integration

Monthly datasets spanning 2014–2024 were assembled from multiple authoritative sources, including ORAS5 (physical ocean state), ERA5 (surface winds), ESA OC-CCI and related products (biogeochemistry), and the Jena CarboScope ocean inversion (target CO flux). All variables were regridded to a common $1^\circ \times 1^\circ$ global grid, temporally aligned to monthly resolution, and integrated into a unified spatiotemporal cube. Data cleaning addressed missing values through a combination of temporal interpolation, spatial nearest-neighbor filling, and climatological replacement where necessary, while duplicate observations were aggregated to monthly means. Land and invalid polar grid points were masked using chlorophyll-based ocean masks, yielding a spatially coherent, ocean-only dataset. Continuous predictors were standard scaled and seasonal effects were encoded using cyclic sine–cosine transforms. This preprocessing produced a consistent, high-dimensional dataset suitable for both machine learning and IAV analysis.

### 4.1.2 Carbon flux prediction using machine learning

Global ocean CO flux prediction was framed as a supervised regression problem at the monthly grid-cell level, with each sample indexed by (year, month, latitude, longitude). Two feature representations were evaluated: a raw-feature configuration using absolute environmental drivers and spatiotemporal encodings, and an anomaly transformed configuration emphasizing departures from local seasonal climatology. Multiple model classes were trained under identical data exposure: Multiple Linear Regression as a linear baseline, and nonlinear models including Multilayer Perceptron regression, Random Forests, and XGBoost.

Model evaluation employed strict time-aware splitting to prevent temporal leakage. Expanding leave-one-year-out (LOYO) framework provided year by year out-of-sample evaluation. This design ensures comparison of model structuresperformance and feature representations under realistic climate situations.

### 4.1.3 Early warning indicators and regime-shift analysis

After prediction using machine learning, this thesis examined whether predicted CO flux shows pattern of increasing instability. Seasonal cycles were removed to form anomaly fields, and early-warning indicators rolling variance and lag-1 autocorrelation were computed using multiple window lengths (6, 12, and 24 months). These indicators quantify changes in variability and temporal persistence that may precede structural transitions. Analyses were conducted globally, spatially, and by ocean basin to assess detection ability across scales.

### 4.1.4 Statistical change-point detection at basin scale

For formal regime-shift detection, gridded flux predictions were aggregated to basin-mean monthly time series using RECCAP2 basin masks. After deseasoning and detrending, interannual variability (IAV) residuals were analyzed using the non-parametric Pettitt test to identify candidate abrupt shifts in mean state. To reduce false positives, candidate change points were confirmed using a persistence-window criterion requiring sustained differences before and after the detected transition. This conservative approach prioritizes robust, basin-scale regime shifts over transient extremes.

### 4.1.5 ENSO-based physical contextualization

Finally, detected regime shifts were interpreted in the context of large-scale climate variability using the Niño 3.4 ENSO index. ENSO was not treated as ground truth, but as an external physical-consistency check. Lead-lag correlations, block-permutation significance testing, and event-alignment analyses were used to evaluate whether detected shifts cluster near strong ENSO phases. This step links statistically detected changes in ocean carbon flux dynamics to known climate drivers while acknowledging that other modes may dominate in non-ENSO sensitive regions.

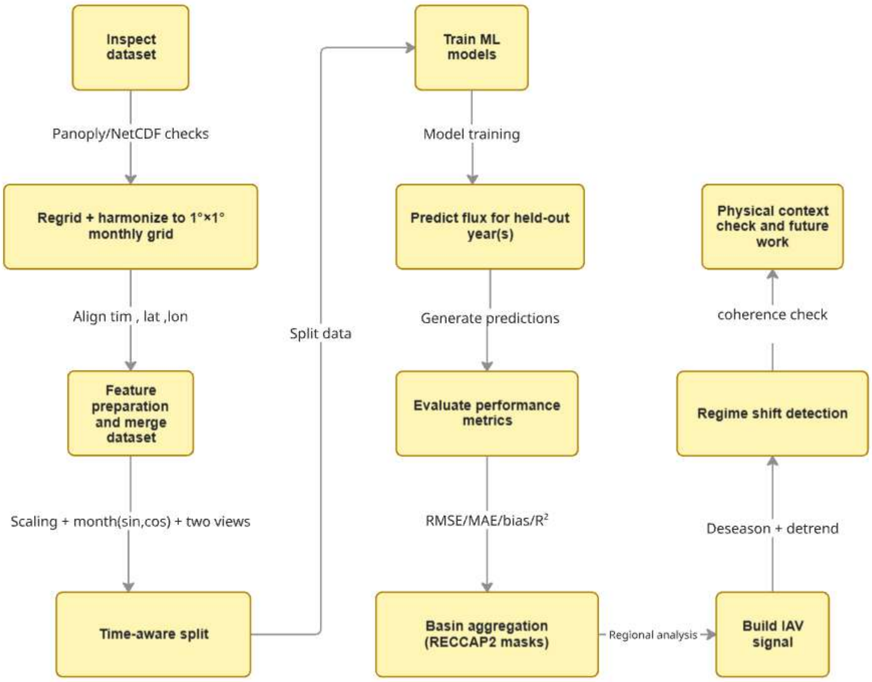



Figure 4.1: Methodology Workflow Diagram

Together, this end-to-end framework integrates data-driven prediction, variability diagnostics, and physically motivated validation to assess both the predictability and structural evolution of global ocean CO fluxes.

## 4.2 Preliminary Design or Design (Model) Specification

### 4.2.1 Multiple Linear Regression (MLR)

Multiple Linear Regression represents the target as a linear combination of predictors defined on the spatio-temporal grid. For each sample $\mathbf{x} \in R^p$, the model is written as

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \;\; = \;\; \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}.$$

Parameter estimation is obtained by minimizing the sum of squared residuals,

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \, ,$$

which can be expressed in closed form (when $X^\top X$ is invertible) as

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

This formulation yields a parametric baseline where the coefficients describe the direction and relative magnitude of the learned linear associations between each driver and $y$. Because the mapping is linear, the model expresses interactions only through explicitly provided transformed features, while correlated predictors can destabilize coefficient estimates and amplify sensitivity to outliers. In this design, the training procedure is defined over strict time-aware splits so the fitted coefficients reflect genuine out-of-sample relationships rather than temporal leakage.

### 4.2.2 Multilayer Perceptron Regressor (MLP)

The Multilayer Perceptron is a feedforward neural network that learns a nonlinear mapping from predictors to the flux target through stacked affine transformations and nonlinear activations. With $L$ hidden layers, the forward computation is

$$\mathbf{h}^{(1)} = \phi(W^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \quad \mathbf{h}^{(l)} = \phi(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \quad \hat{y} = W^{(L+1)}\mathbf{h}^{(L)} + b^{(L+1)},$$

where $\phi(\cdot)$ is a nonlinear activation such as ReLU. Training is defined by minimizing a regression loss, commonly mean squared error with an $L_2$ penalty on parameters,

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \; + \; \lambda \|\theta\|_2^2,$$

and parameter updates are obtained by backpropagation with gradient-based optimization. This model class expresses smooth and highly nonlinear response surfaces, allowing interactions among drivers to be captured implicitly through hidden units. The specification in this study treats the network as a non-sequential regressor operating on monthly gridpoint samples, with the nonlinearity providing capacity beyond linear baselines while maintaining a consistent supervised objective aligned with the target flux field.

### 4.2.3 Random Forest Regressor (RF)

Random Forest regression is an ensemble learning method that aggregates the predictions of many decision trees trained under bootstrap sampling and randomized feature selection at splits. Each tree performs recursive partitioning of the predictor space so that terminal nodes represent locally homogeneous response regions, and the forest prediction is the average of all tree predictions:

$$\hat{y}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{x}),$$

where $f_t$ denotes the $t$-th tree and $T$ is the number of trees. This architecture captures nonlinearities and higher-order interactions through axis-aligned splits without requiring manual feature interaction terms, and the aggregation over trees reduces variance relative to a single tree. The effective capacity of the forest is controlled through constraints on tree growth and node sizes, while the randomization mechanisms improve generalization by decorrelating individual trees. In this modeling design, the forest provides a robust nonlinear baseline whose training is conducted using temporally blocked splits so that performance reflects generalization to unseen years.

### 4.2.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient boosting framework that builds an additive ensemble of regression trees sequentially, where each new tree fits residual structure not yet captured by the current ensemble. The prediction is expressed as

$$\hat{y}(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}), \quad f_m \in \mathcal{F},$$

with $\mathcal{F}$ denoting the space of regression trees and $M$ the number of boosting rounds. The model is trained by minimizing a regularized objective,

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m),$$

where $\ell$ is commonly squared error for regression. Regularization penalizes complexity and stabilizes learning, typically written as

$$\Omega(f) = \gamma\,(\#\text{leaves}) + \frac{\lambda}{2}\sum_{j} w_j^2,$$

where $w_j$ are leaf weights, $\gamma$ discourages excessively complex trees, and $\lambda$ provides $L_2$ shrinkage on leaf scores. The boosting procedure uses gradient and curvature information to select splits efficiently, which supports strong predictive accuracy for heterogeneous nonlinear relationships across ocean regions. In this specification, boosting is applied under strict temporal validation, with subsampling and regularization forming part of the defined model structure for controlling overfitting.

## 4.3 Data Collection

The dataset used in this study was compiled to support global-scale prediction of ocean–atmosphere $CO_2$ fluxes and the detection of regime shifts over the period 2014–2024. A harmonized monthly dataset was constructed on a global latitude–longitude grid, where each record corresponds to a unique combination of year, month, latitude, and longitude. The dataset integrates physical, biogeochemical, and atmospheric drivers of air–sea $CO_2$ exchange, together with an observation-based $CO_2$ flux product serving as the target variable. Key physical ocean variables sea surface temperature (SST), salinity, sea surface height (sossheig), and mixed layer depth (MLD) were obtained from the ORAS5 (Ocean Reanalysis System 5) global ocean reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service. ORAS5 combines numerical ocean modeling with in situ and satellite observations through data assimilation, providing spatially and temporally continuous fields without observational gaps. Atmospheric forcing variables influencing air–sea gas exchange, specifically the 10 m zonal (u10) and meridional (v10) wind components, were collected from the ERA5 monthly mean reanalysis on single levels, also produced by ECMWF.

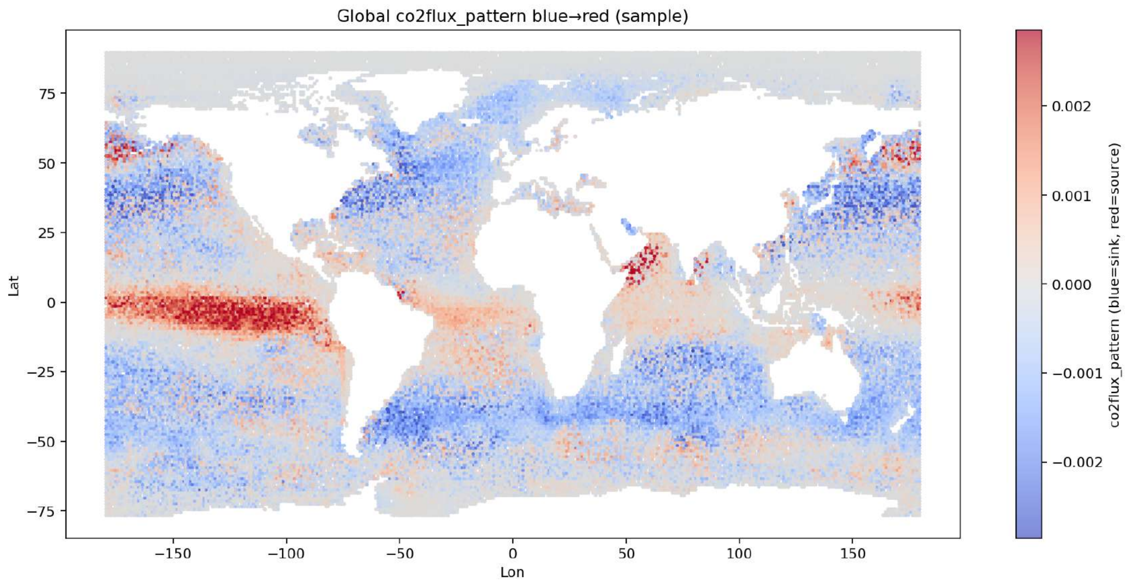



Figure 4.2: Visual representation of carbon flux dataset in global view indicates ocean sink and source.

ERA5 integrates a wide range of global observations using a state-of-the-art data assimilation system to deliver consistent and physically constrained atmospheric fields. Monthly mean wind data were used to match the temporal resolution of the oceanic variables. Biogeochemical variables representing biological activity and nutrient availability chlorophyll-a concentration (chl), phosphate (po4), and net primary productivity (nppv) were obtained from the ESA Ocean Colour Climate Change Initiative (CCI) products and related global biogeochemical datasets archived at the CEDA Data Catalogue. These satellite-based, climate-quality datasets provide long-term, spatially resolved estimates of marine ecosystem dynamics that are directly linked to variations in ocean carbon uptake. The target variable, ocean–atmosphere $CO_2$ flux (co2flux_pattern), was derived from the JENA CarboScope ocean inversion product, which is based on quality-controlled surface ocean $CO_2$ observations from the Surface Ocean $CO_2$ Atlas (SOCAT). This dataset integrates observed $pCO_2$ measurements with atmospheric transport modeling to resolve seasonal to interannual variability in global ocean $CO_2$ fluxes, making it a reliable reference for machine learning model training and evaluation. To characterize large-scale climate variability, the El Niño–Southern Oscillation (ENSO) signal was incorporated using the Niño 3.4 index obtained from the NOAA Physical Sciences Laboratory. ENSO information was used to contextualize climate-driven variability and interpret regime shifts in ocean carbon dynamics. Additionally, ocean basin masks from the RECCAP2 framework were applied to classify grid cells into major ocean basins, enabling basin-level aggregation and comparative analysis of predicted $CO_2$ fluxes and regime shift indicators. All datasets were acquired in NetCDF format, temporally aligned to monthly resolution, and spatially regridded to a common global grid prior to preprocessing and model implementation.

| index | year | month | lat | lon | chl | .... | sst | ... | co2flux_pattern |
|---|---|---|---|---|---|---|---|---|---|
| 626487 | 2015 | 3 | 82.500000 | -155.500000 | 0.033404 | .... | -1.668038 | ... | -0.000018 |
| 4111711 | 2022 | 3 | -73.500000 | -120.500000 | 0.506519 | .... | -1.056691 | ... | -0.001694 |
| 4897334 | 2023 | 9 | 34.500000 | -163.500000 | 0.083854 | .... | 26.131485 | ... | 0.000538 |
| 3009994 | 2019 | 12 | 36.500000 | 133.500000 | 0.582657 | .... | 16.136407 | ... | -0.000818 |
| 3454575 | 2020 | 11 | -29.500000 | 32.500000 | 0.278401 | .... | 23.861613 | ... | -0.000143 |

Figure 4.3: Sample dataset for prediction of ocean carbon flux (the dots indicate columns not shown here but include other drivers for visual representation).

The dataset consists of 5,537,796 rows and 14 columns, representing a large spatiotemporal sample of oceanographic and atmospheric variables. Each row corresponds to a unique year–month–latitude–longitude grid cell, with 12 continuous predictors and one target variable (`co2flux_pattern`) which is the ocean carbon flux stored as floating-point values.

### 4.3.1 Data Cleaning

All datasets spanning 2014–2024 were carefully inspected for missing values, duplicates, and unrealistic measurements. Temporal gaps in the monthly data were first addressed using linear interpolation, filling missing months for each grid cell. Remaining missing values, particularly where entire grid points were absent, were estimated using spatial nearest-neighbor interpolation or replaced with monthly mean values to preserve seasonal patterns. Duplicate observations at identical (time, latitude, longitude) locations were aggregated using monthly averages to en- sure a unique value per grid cell. Land area and polar regions were masked using false ocean mask to keep only valid ocean data. After these preprocessing steps, all datasets were complete and consistent, with missing values either filled or masked where data were unavailable and removed.

### 4.3.2 Data Transformation

For maintaining consistency across all heterogeneous oceanographic variable's datasets, all were converted to a common $1^\circ \times 1^\circ$ spatial grid and temporally aligned by same year and month spanning 2014–2024. Predictor variables like (*chl*, *po4*, *nppv*, *MLD*, *salinity*, *sst*, *sossheig*, *u10*, *v10*) were normalized using min max scale and the target variable (*co2flux_pattern*) was standardized using $z$-score normalization to address unit inconsistencies. Temporal variables such as year and month were encoded using cyclic sine and cosine transformations to preserve seasonal periodicity. to maintain same longitude format , it was converted from 0–359° to −180°–180°, multiple observations within each grid cell were aggregated into monthly means. Collectively, these transformations helped for physically consistent feature representations which is suitable for machine learning modeles to learn and predict.

### 4.3.3 Data Integration

Heterogeneous datasets from ORAS5, ERA5, OC-CCI, and Jena CarboScope were harmonized into a unified 3D cube (384 time-steps × 167 lats × 360 lons) using `xarray`, regridding all variables to the ocean-only reference grid (−76.5° to 89.5° lat, −179.5° to 179.5° lon) via nearest-neighbor interpolation with conservative mass preservation. Grid mismatches, cell-centered conventions, and wind longitude formats were resolved; duplicates per (time,lat,lon) aggregated to monthly means; minor temporal gaps filled with monthly climatologies.

### 4.3.4 Data Reduction

When checked for ocean and land data, approximately 30% of grid points were identified as land and 70% as ocean. An ocean mask from chlorophyll (chl) data was then applied, deleting the land-masked data for data reduction purposes and retaining $\sim$ 70% valid ocean points ($\sim$ 500k grid-months), yielding a complete, spatially coherent ocean-only dataset. For regime shift detection, data points were aggregated to monthly mean per basin using basin mask from RECCAP2 ocean mask regions.

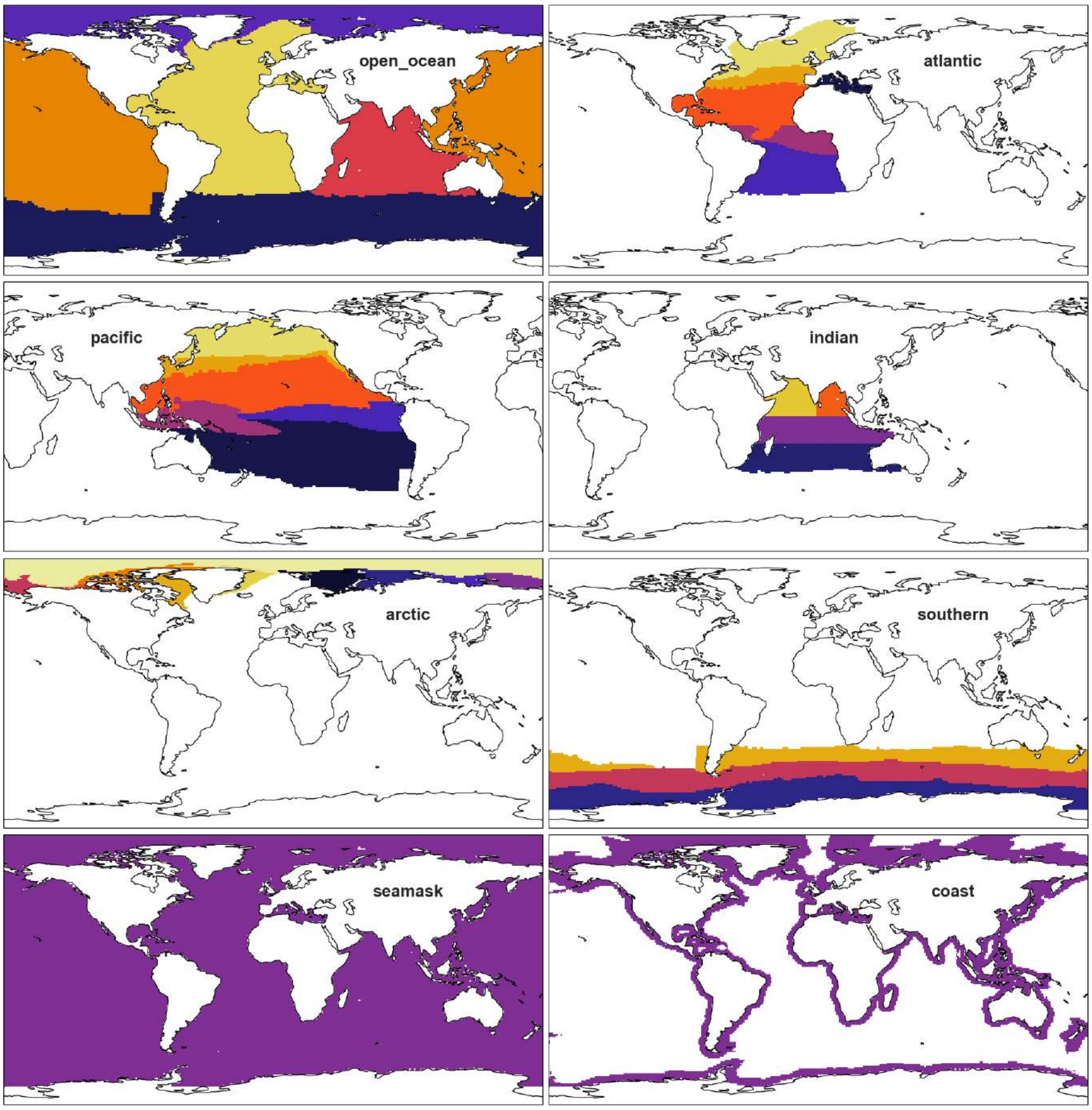



Figure 4.4: RECCAP2 ocean mask regions

### 4.3.5 Summary of Preprocessed Data

The preprocessing pipeline produced a complete, ocean-only dataset spanning 2014–2024 (384 monthly time-steps) on a unified $1^\circ \times 1^\circ$ global grid ($-76.5^\circ$ to $89.5^\circ$ lat, $-179.5^\circ$ to $179.5^\circ$ lon), comprising 12 variables: 9 predictors (SST, salinity, sossheig, MLD, u10, v10, chl, po4, nppv), 1 target (co2flux_pattern), After harmonization into a 3D cube ($384 \times 167 \times 360$), data cleaning eliminated missing values via interpolation and climatologies, while duplicates were aggregated to monthly means. Land points ($\sim 30\%$ of original grid) were masked using chlorophyll data, reducing storage to $\sim 70\%$ valid ocean grid-months ($\sim 500$k points) without signal loss. Continuous features underwent Min–Max scaling (predictors) and z-score standardization (target); time variables received cyclic encoding; and longitude was unified to $-180^\circ$–$180^\circ$. The final dataset is spatially coherent, computationally efficient, and machine-learning ready, enabling robust global $CO_2$ flux prediction and regime shift detection.

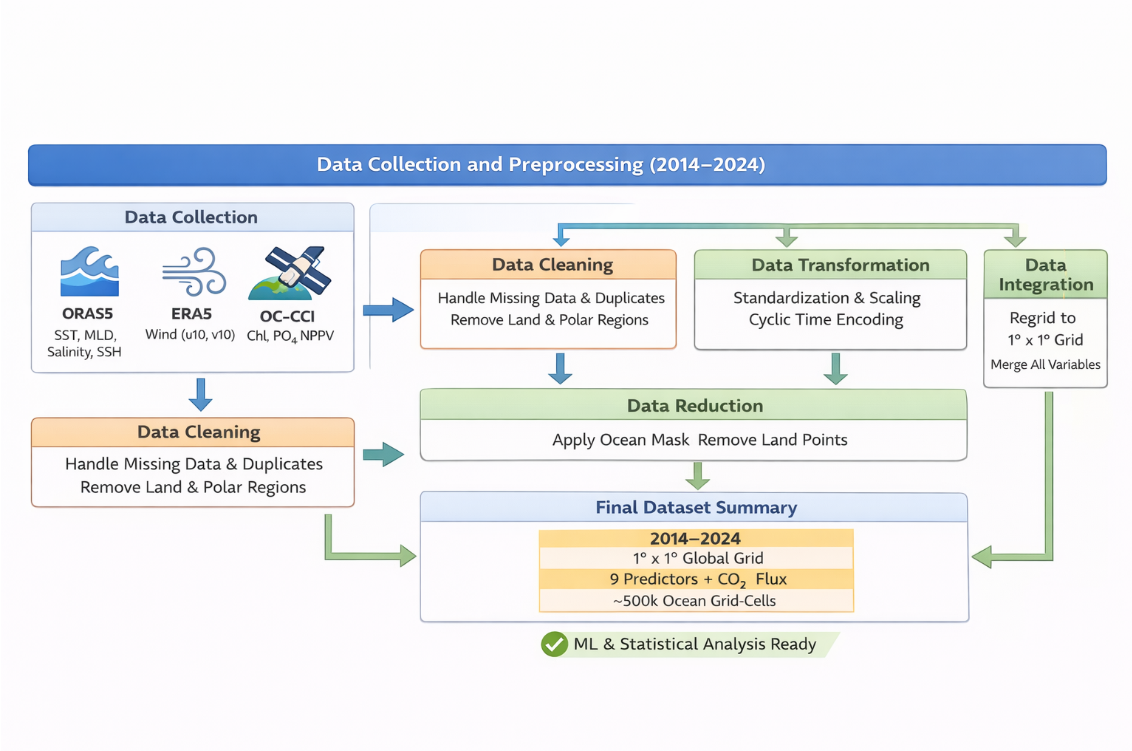



Figure 4.5: Data collection and preprocessing

### 4.3.6 Pearson and Spearman Correlation

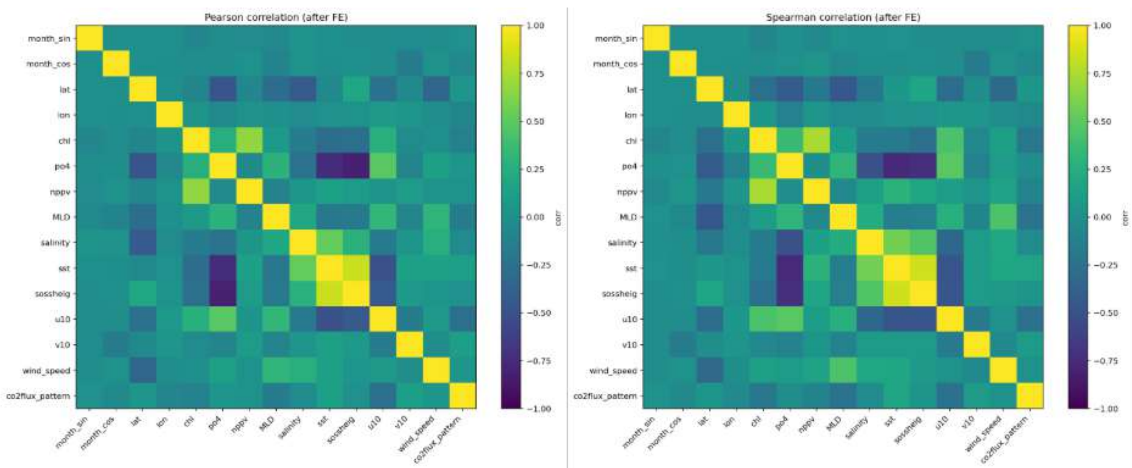



Figure 4.6: Pearson and Spearman correlation graph

Using both Pearson (linear) and Spearman (rank-based) correlations, the pre–feature-engineering analysis shows that `co2flux_pattern` has only weak univariate associations, with all absolute correlations below $\approx 0.27$. The strongest and most consistent relationship is with zonal wind `u10` (Pearson $r = -0.263$; Spearman $\rho = -0.254$), indicating that stronger east–west winds are generally accompanied by lower flux pattern values. MLD is also negatively related, with a markedly stronger monotonic component ($r = -0.153$ vs. $\rho = -0.239$), suggesting a nonlinear response in which deeper mixing shifts the pattern consistently but not proportionally.

Thermodynamic predictors exhibit weak positive associations: SST ($r = 0.131$; $\rho = 0.167$) and `v10` ($r = 0.126$; $\rho = 0.103$) imply modest increases in the pattern under warmer conditions and enhanced meridional wind variability. Biological proxies are primarily negative in rank space: `chl` ($r = -0.118$; $\rho = -0.198$) and NPPV (near-zero Pearson $r = 0.024$ but $\rho = -0.100$), consistent with nonlinear or regime-dependent behavior.

Spatial/seasonal encodings contribute little in isolation (e.g., latitude $\rho = 0.060$; longitude $\rho = -0.048$; harmonics $\approx 0$), and salinity, SSH, and phosphate are negligible ($|r|, |\rho| 0.04$). Overall, the weak single-variable correlations imply multivariate coupling and nonlinear interactions, motivating nonlinear models that learn cross-driver structure rather than relying on individual linear effects.

Regime-dependent behavior refers to the conditional response of air–sea $CO_2$ flux to environmental drivers, whereby the sign and magnitude of the relationship vary across oceanic states, regions, or climatic conditions, leading to weak global correlations despite strong localized physical coupling. A low Spearman coefficient does not imply the absence of a relationship; it only indicates the absence of a monotonic relationship. Many ocean–carbon processes are non-monotonic, thresholded, or regime-dependent.

Therefore we implemented mutual information to clarify the non-linear relationship between target and features. Mutual Information quantifies the amount of shared information between a driver and $CO_2$ flux, capturing general (non-linear and non-monotonic) dependence that may not be detected by traditional correlation measures.

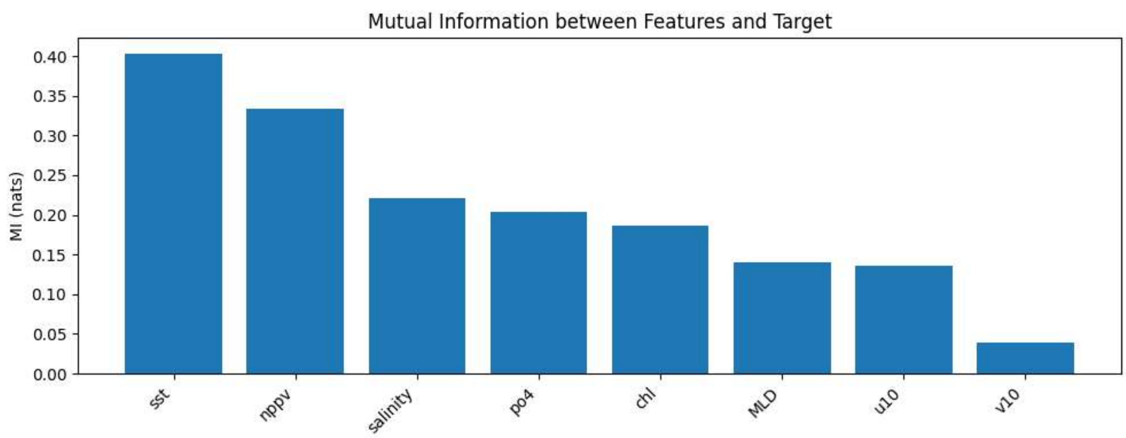



Figure 4.7: Mutual information between features and target

Fig Mutual information between features and target

| Feature | Interpretation |
|---|---|
| **SST (0.40)** | Strongest overall dependence with $CO_2$ flux |
| **NPPV (0.33)** | Strong non-linear / interaction-driven influence |
| **Salinity (0.22)** | Moderate but meaningful coupling |
| **$PO_4$ (0.20)** | Biogeochemical control present |
| **Chlorophyll (0.19)** | Biological mediation of flux |
| **MLD (0.14)** | Physical modulation via mixing |
| **u10 (0.14)** | Wind influence present but indirect |
| **v10 (0.04)** | Very weak or regime-limited influence |

Table : Mutual information values

Mutual information (MI) indicates that SST and NPPV carry the strongest dependence with $CO_2$ flux, which is physically expected because their effects are non-linear and often regime-dependent. SST influences flux via solubility and carbonate chemistry, and its relationship can vary with season, stratification, and circulation. NPPV reflects biological uptake/export, but its impact emerges through interactions with nutrients and mixing, so dependence can be high even when correlations are low. Salinity, $PO_4$, chlorophyll, and MLD show moderate MI, consistent with indirect/conditional controls through water masses, buffering, biology, and mixing. The low MI for v10 suggests directional wind alone is weak; gas exchange depends more on wind magnitude/turbulence, explaining u10 > v10.

Although rank correlations were generally weak, mutual information analysis revealed substantial dependence between $CO_2$ flux and several environmental drivers, particularly SST and net primary production. This indicates that the flux–driver relationships are predominantly non-linear and regime-dependent, justifying the use of non-linear machine learning models.

### 4.3.7 Implementation of Selected Design

**Prediction of Ocean Carbon Flux Using Machine Learning**

The predictive modeling stage was implemented as a controlled set of supervised learning experiments designed to evaluate how well indirect oceanographic drivers can reproduce the target field, (co2flux_pattern), under strict time aware generalization. Each sample corresponds to a monthly gridpoint observation indexed by $((\text{year}, \text{month}, \text{lat}, \text{lon}))$ and associated environmental predictors. Model training and evaluation were structured explicitly to prevent temporal mixing between training and evaluation periods, ensuring that reported performance reflects out-of-sample skill on unseen years rather than interpolation within the same time period.

Two feature configurations were evaluated in parallel to isolate the impact of anomaly-based predictors. The first configuration used a "raw feature" representation, in which the input vector contained the spatial coordinates $((\text{lat}, \text{lon}))$, seasonal phase terms $((\text{month\_sin}, \text{month\_cos}))$, and the full set of environmental drivers available in the cleaned dataset, including chlorophyll $((\text{chl}))$, phosphate $((\text{po4}))$, net primary productivity proxy $((\text{nppv}))$, mixed-layer depth $((\text{MLD}))$, salinity, sea surface temperature ((sst)), sea surface height ((sossheig)), and wind components ((u10), (v10)) together with their derived magnitude ((wind_speed)). The second configuration used an "anomaly-transformed" representation, where a subset of key predictors was replaced by their anomaly forms relative to the local gridcell-month climatology. This configuration was constructed to emphasize interannual departures and event-scale variability while retaining the remaining predictors in their original form. By training and evaluating the same model classes under these two parallel feature sets, the experiments quantify whether predictive skill improves when the input representation focuses on departures from the seasonal mean state rather than absolute values.

Model evaluation was conducted under time-aware splitting strategies.The strategy implemented an expanding "leave-one-year-out" (LOYO) evaluation, designed to approximate a repeated operational setting in which predictions are made for one target year at a time. For each test year ($t$), validation was assigned to the immediately preceding year (($t-1$)), and training was defined as all available years strictly earlier than (($t-1$)). This yields a sequence of independent test evaluations across many years, allowing performance to be assessed as a function of time and training record length. Unlike a single held-out test period, LOYO provides a year-by-year characterization of generalization that is particularly relevant for climate applications where interannual variability can strongly influence apparent model skill. The LOYO design also allows comparison of how each model responds to different test-year conditions, including anomalous years, without changing the overall training protocol.

Across both splitting strategies, the same family of regression models was trained and evaluated to provide a spectrum of linear and nonlinear mapping capacities. Multiple Linear Regression served as a linear baseline, while three nonlinear machine learning models Multilayer Perceptron regression, Random Forest regression, and XGBoost regression provided increasing capacity to represent nonlinear responses and interactions among drivers. For each split and each feature configuration (raw versus anomaly-transformed), all models were trained on the same training years, tuned against the same validation years, and evaluated on the same held-out test years, ensuring that differences in performance arise from model structure and feature representation rather than differences in data exposure.

The experimental design therefore supports two central comparative questions. First, it enables direct ranking of model classes under identical temporal generalization constraints, clarifying whether nonlinear methods provide measurable skill gains over linear baselines when predicting (co2flux_pattern) from indirect drivers. Second, it enables a controlled assessment of anomaly-based representation by repeating the entire modeling and evaluation procedure on the anomaly-transformed dataset and comparing the resulting performance distributions against those obtained from raw predictors. In this way, the implementation isolates the value of incorporating anomaly information in the predictors while preserving consistent data partitions, consistent evaluation years, and consistent model families.

## DIAGNOSING INSTABILITY AND REGIME SHIFTS IN OCEAN CARBON FLUX

| Window | Train year(s) | Validation year | Test year |
|---|---|---|---|
| 1 | 2014–2015 | 2016 | 2017 |
| 2 | 2014–2016 | 2017 | 2018 |
| 3 | 2014–2017 | 2018 | 2019 |
| 4 | 2014–2018 | 2019 | 2020 |
| 5 | 2014–2019 | 2020 | 2021 |
| 6 | 2014–2020 | 2021 | 2022 |
| 7 | 2014–2021 | 2022 | 2023 |
| 8 | 2014–2022 | 2023 | 2024 |

Table 4.1: Leave-One-Year-Out window for train, validation and test dataset

To go beyond simply predicting ocean $CO_2$ flux and to assess whether the system exhibits signs of increasing instability, this study incorporated an early-warning indicator (EWS) analysis. The purpose of early warning signs is to capture a change in the system with respect to the stability attributes of systems which may include rapid structural changes.The ability to capture such signals will allow a window to highlight the chances of the critical threshold crossing on a temporal scale.In that respect, the signs indicate potential trends and risks.However, capturing these signs is not an exact science and remains on the ambiguous side with respect to the regime change and the timing on it.What the signs show is an increased uncertainty and risk of a regime change..

### Anomaly Construction

For isolating unusual random variability from the seasonal cycle, the predicted $CO_2$ flux was deseasoned at each grid cell to compute monthly anomalies of it. For each grid cell $(i, j)$ and time $t$, the anomaly calculated as:

$$x'_{i,j,t} = x_{i,j,t} - \overline{x}_{i,j,m}$$

where $x_{i,j,t}$ is the predicted $CO_2$ flux at lat $i$, lon $j$, and time $t$, $\overline{x}_{i,j,m}$ represents the long term mean flux for that grid cell and month $m$, and $x'_{i,j,t}$ is the resulting anomaly time series. This transformation removes the climatological seasonal cycle which helps the analysis to focus on abrupt changes rather than expected seasonal behavior.

### Early-Warning Indicators

Early warning indicators were computed from the anomalies using rolling variance and lag-1 autocorrelation for capturing the complementary parts of instability.

### Rolling Variance

Rolling variance measures the changes in short term variability for the predicted flux within a moving window of length $W$ months. For each time step $t$, rolling variance calculated as:

$$\mathrm{Var}_t = \frac{1}{W-1} \sum_{k=0}^{W-1} \left(x'_{t-k} - \overline{x'}_t\right)^2$$

where $x'_{t-k}$ is the anomaly at lag $k$ within the window and $\overline{x'}_t$ is the mean anomaly over the same window:

$$\overline{x'}_t = \frac{1}{W} \sum_{k=0}^{W-1} x'_{t-k}$$

Increase in rolling variance means increased fluctuations in flux which may reflect in growing instability in the underlying system.

**Lag-1 Autocorrelation**

Lag-1 autocorrelation tells us the persistence of anomalies across consecutive time steps within the same rolling window. Defined as:

$$\rho_{1,t} = \frac{\sum_{k=1}^{W-1} \left(x'_{t-k} - \overline{x'}_t\right) \left(x'_{t-k-1} - \overline{x'}_t\right)}{\sum_{k=0}^{W-1} \left(x'_{t-k} - \overline{x'}_t\right)^2}$$

Higher value of lag-1 autocorrelation indicates that anomalies decay more slowly which shows increased persistence and reduced system flexibility.

Indicators are analyzed using various rolling window lengths (6, 12, and 24 months) to assess short-term variations and long-term trends.Due to the need for early-warning indicators to have uninterrupted multi-year time series, all Random Forest prediction windows had to be merged into one time series for time depth to permit meaningful signal analysis.This involved the absence of the prediction window.To determine the emergence and scale-wise robustness of early warning signals, this analysis focused on the spatial grids, ocean basins, and global mean of the anomaly, rolling variance, and lag-1 auto-correlation.This derived the regions and timeframes of greater variability and persistence concerning ocean CO2 fluxes.

These indicators are more like warning lights that often tell you that the system could be less stable and more shift prone, but still do not tell you when and what structural changes take place.This is a particularly important oversight because predictions based on models are only for five years, a time span that is too short for the reliable detection of regime shifts.Because of this, early warning analyses can only be used diagnostically.To define regime shifts more precisely, the Pettitt test, combined with the Henderson persistence criterion, was used on the time series of basin-scale CO2 flux observations, which were of longer duration.This meant that only strong and persistent shifts were considered regime changes, in order to bypass the short-term variability that could be present.

This study shocases two approach: first, a 12-month persistence check confirms the change is sustained and large compared to normal variability and second the Pettitt test to find significant change points in deseasoned and detrended CO flux.Finally checked with ENSO whih acted only as an external physical coherence checking factor.

## Data Preparation and Basin Aggregation

Monthly basin-scale time series were constructed from gridded ocean $CO_2$ flux fields by aggregating all valid ocean grid cells within each basin. For each basin and month, the workflow retained:

- **true_mean**: the basin-mean flux value (target signal),
- **n_cells**: number of contributing grid cells (coverage/robustness indicator),
- **date**: monthly timestamp.

This produced a balanced panel dataset of 5 basins $\times$ 132 months (2014–2024). Signal Decomposition: Seasonal Adjustment and IAV Construction Direct regime-shift detection on raw flux can be confounded by strong seasonal cycles and long-term trends. Therefore, the flux series was decomposed into progressively cleaner components: Seasonal anomaly (deseasoning) For each basin, a monthly climatology was computed across all years:

$$\mu_{b,m} = \frac{1}{N_{b,m}} \sum_{t:\, month(t)=m} x_b(t)$$

where $x_b(t)$ is the basin mean flux and $m$ is month-of-year. The deseasoned anomaly was then:

$$\text{anom}_b(t) = x_b(t) - \mu_{b,\, month(t)}$$

## Trend removal (detrending)

A linear trend was fitted to $\text{anom}_b(t)$ using time index $t$, and removed:

$$\text{anom}_b(t) = \alpha_b + \beta_b t + \epsilon_b(t)$$

$$\text{iav}_b(t) = \text{anom}_b(t) - (\alpha_b + \beta_b t)$$

The resulting iav is the interannual variability residual, intended to emphasize persistent changes and suppress regular seasonal structure.

## Regime Shift Detection Using the Pettitt Test

The Pettitt method (Pettitt 1979) is a rank-based nonparametric statistical test method. The null hypothesis of the test is that, when arbitrarily splitting the sample in two segments, there is no change in the mean value of each segment.

Regime shifts were detected independently for each basin using the Pettitt test, a non-parametric single-change-point method that identifies an abrupt shift in location (mean/edian) without assuming normality. For a time series $\{x_1, \ldots, x_n\}$, Pettitt evaluates all candidate split points $t$ and computes:

$$U_t = 2\sum_{i=1}^{t} r_i - t(n+1)$$

where $r_i$ is the rank of $x_i$ among all observations. The estimated change point is:

$$t^* = \arg\max_t |U_t|$$

and the test statistic is:

$$K = \max_t |U_t|$$

An approximate p-value is:

$$p \approx 2\exp\left(\frac{-6K^2}{n^3 + n^2}\right)$$

Pettitt was applied to $\mathrm{iav}_b(t)$ to detect candidate change points while reducing season/trend artifacts.

## Confirmation via Persistence Window (Robust Shift Flagging)

To reduce false detections from short lived noise, each candidate change point was evaluated with a persistence window confirmation. For a candidate month $t^*$, two windows were defined:

pre-window: $[t^* - 12, \ldots, t^* - 1]$

post-window: $[t^*, \ldots, t^* + 11]$

A change point was confirmed if the step change exceeded an effect-size threshold:

$$|\mu_{post} - \mu_{pre}| \geq \gamma \cdot \sigma_{pre}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation of the pre-window, and $\gamma$ is a tunable sensitivity parameter (e.g., 0.5). This step produced the final binary indicator:

`pettitt_cp_confirmed` (1 = confirmed regime shift month)

## ENSO-Based Contextual Validation

To evaluate whether detected regime shifts in basin-scale ocean $CO_2$ flux variability are physically plausible, this study uses the El Niño–Southern Oscillation (ENSO) as an external climate benchmark. ENSO is one of the dominant modes of interannual climate variability and is widely recognized as a key driver of interannual variability (IAV) in air–sea $CO_2$ exchange, especially through its strong influence on tropical Pacific carbon dynamics [5]. Prior literature shows that the equatorial Pacific contributes substantially to global $CO_2$-flux variability and exhibits a strong relationship with ENSO, often expressed as a pronounced correlation between $CO_2$-flux anomalies and Niño3.4 variability.

This ENSO–flux linkage is physically motivated by the standard bulk formulation of air–sea $CO_2$ flux:

$$F = kK_0(T, S)\,(pCO_{2,w} - pCO_{2,a})$$

where $F$ is the sea–air $CO_2$ flux, $k$ is the gas-transfer velocity (primarily wind-driven), $K_0(T, S)$ is $CO_2$ solubility (dependent on sea-surface temperature $T$ and salinity $S$), and $(pCO_{2,w} - pCO_{2,a})$ is the air–sea $CO_2$ partial-pressure gradient. ENSO affects flux through two key pathways represented in this equation. First, ENSO-related SST anomalies modify $K_0$, altering $CO_2$ solubility and therefore the thermodynamic component of exchange. Second, ENSO-driven changes in upper-ocean circulation—particularly upwelling strength and thermocline depth—modify surface carbon supply and $pCO_{2,w}$, producing large, physically interpretable flux anomalies (e.g., reduced upwelling during El Niño and enhanced upwelling during La Niña). Because regime shifts in this thesis are defined as sustained changes in the dynamics of flux variability (not isolated month-to-month noise), coherence with ENSO timing and phases provides a meaningful physical-consistency check. Methodologically, ENSO validation is performed on anomaly/IAV-type signals rather than raw flux magnitude, consistent with established practice: long-term and seasonal components are removed to isolate interannual variability attributable to climate modes. In this work, basin-level flux IAV is compared with Niño3.4 using lag-aware association tests (allowing ENSO to lead the ocean carbon response) and event-alignment analyses (testing whether detected change points cluster near strong ENSO extremes). This is treated as a physical-coherence validation rather than "ground-truth confirmation," since ENSO influence varies by basin and other climate modes may dominate outside ENSO-sensitive regions.

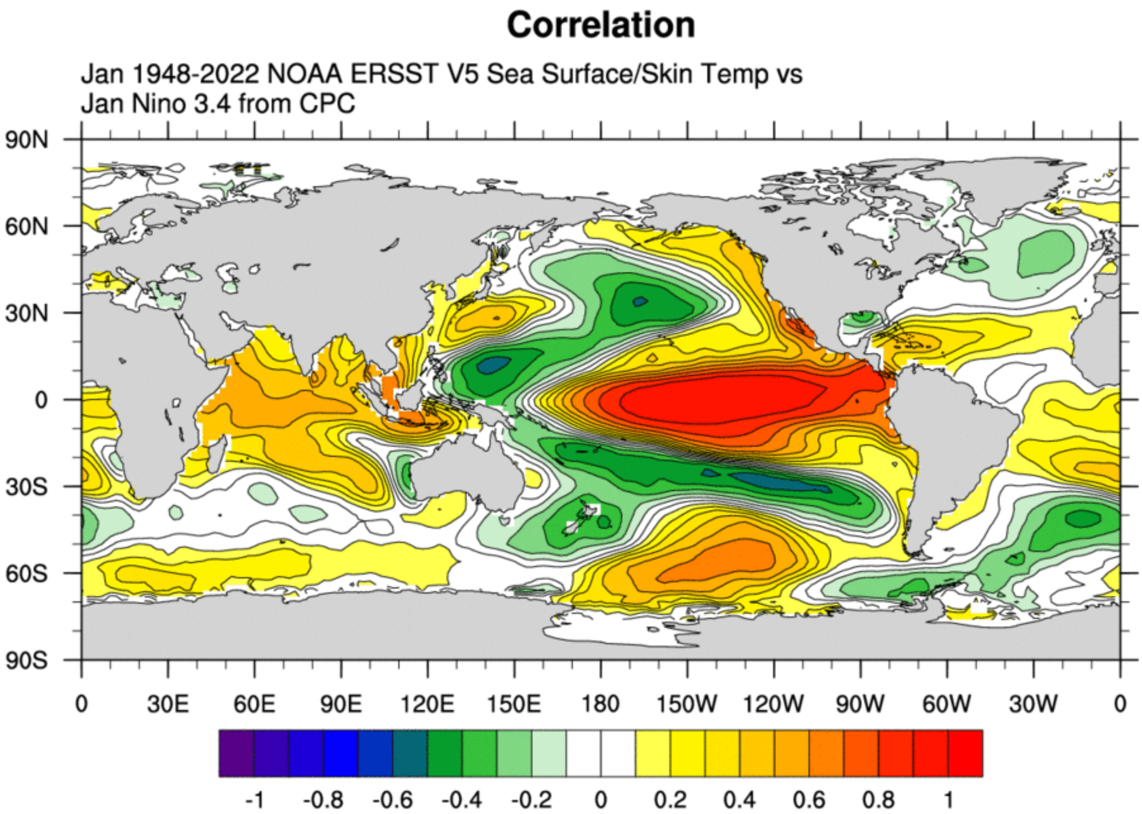



Figure 4.8: ENSO is a periodic fluctuation (i.e., every 2–7 years in wind and sea surface temperature affecting the climate of many parts of the world.

ENSO was used after detection to assess whether detected shifts occur near large ENSO anomalies or whether ENSO leads basin variability.

- ENSO index used: monthly Niño3.4 (nino34), where positive indicates El Niño and negative indicates La Niña.
- Lead–lag correlation: correlations were computed for lags 0–12 months (ENSO leading IAV).
- Autocorrelation-aware significance: a block-permutation test was used to form a null distribution that preserves ENSO autocorrelation while breaking coupling with IAV.

Additionally, a shift-date alignment check measured the distance (in months) from each confirmed shift month to the nearest ENSO "extreme" month defined by $|\text{Niño3.4}| \geq 1.0$, and compared it against random-month baselines.

Importantly, ENSO did not determine where shifts were detected; it was only used to interpret and contextualize detected shifts.

# Chapter 5

# Result Analysis

## 5.1 Performance Evaluation

This chapter summarizes how predictions were evaluated using performance metrics and analyzing coherence of climate events with statistically detected regime shift.

Below are the performance metrics (with equations) for the test set each window for global carbon flux prediction.

### Notation

In this $N$ test samples (grid-cells $\times$ months):

True flux: $y_i$

Predicted flux: $\hat{y}_i$

Error: $e_i = y_i - \hat{y}_i$

Mean of truth:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

### Error magnitude metrics

#### Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

#### Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

**Root Mean Squared Error (RMSE)**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

## Goodness-of-fit metric

**Coefficient of Determination ($R^2$)**

Let

$$\text{SSE} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \qquad \text{SST} = \sum_{i=1}^{N}(y_i - \bar{y})^2$$

Then

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

## LOYO window averaging (reporting across years)

If each LOYO window $w$ has $N_w$ samples and metric $M_w$, a weighted mean is:

$$M_{\text{LOYO}} = \frac{\sum_w N_w M_w}{\sum_w N_w}$$

## Regime shift analysis

For regime shift analysis, after spatial aggregation, the analysis produces a basin–monthly time series for the five RECCAP2 basins (Atlantic, Pacific, Indian, Arctic, Southern) over 2014-01 to 2024-12 (132 months per basin). Each row of the basin table represents one basin-month value:

- `true_mean`: spatial mean flux across all grid cells belonging to the basin in that month
- `n_cells`: number of basin grid cells contributing to the mean

Latitude and longitude do not appear in the basin table because the spatial field has already been summarized into a single basin mean value per month. No specific performance evaluation has been performed. Only checked with ENSO extreme to identify if ENSO can have an effect on carbon flux regime shift.

## 5.2 Analysis of Design Solutions

### Carbon Flux prediction

Model performance is summarized under the Leave-One-Year-Out evaluation scheme, where lower values of MSE, RMSE, and MAE indicate better predictive accuracy, the following tables report window wise results, mean performance across all windows, and best-window performance for both actual dataset and anomaly transformed dataset.

| model | MSE | RMSE | MAE |
|---|---|---|---|
| MLR | 6.138e-07 | 0.000782 | 0.000552 |
| MLP | 3.095e-07 | 0.000556 | 0.000392 |
| RF | 1.067e-07 | 0.000326 | 0.000216 |
| XGB | 1.172e-07 | 0.000342 | 0.000231 |

Table 5.1: Mean across all windows for actual dataset

Across all windows using RAW features, RF is the best overall model on every error metric: MSE = 1.067e07, RMSE = 0.000326, MAE = 0.000216. XGB is consistently second (MSE = 1.172e07, RMSE = 0.000342, MAE = 0.000231), showing slightly higher error than RF but still far better than the linear baselines. MLP performs moderately (MSE = 3.095e07, RMSE = 0.000556, MAE = 0.000392), while MLR performs worst (MSE = 6.138e07, RMSE = 0.000782, MAE = 0.000552). In short, RAW mean performance ranks: RF > XGB > MLP > MLR, indicating tree-based models generalize best under time-aware evaluation when learning absolute flux values.

| Model | MSE Win | MSE | RMSE Win | RMSE | MAE Win | MAE |
|---|---|---|---|---|---|---|
| MLR | 1 | 5.024e-07 | 1 | 0.000709 | 1 | 0.000510 |
| MLP | 3 | 2.864e-07 | 3 | 0.000535 | 3 | 0.000365 |
| RF | 8 | 9.482e-08 | 8 | 0.000308 | 8 | 0.000206 |
| XGB | 8 | 1.077e-07 | 8 | 0.000328 | 5 | 0.000224 |

Table 5.2: Best window per model for actual dataset. Win = Window

The "best window" results show each model achieves its minimum errors at different points in the window sequence. RF peaks in Window 8 with the best overall RAW performance (MSE = 9.482e08, RMSE = 0.000308, MAE = 0.000206), confirming that RF improves as the training period grows. XGB also performs best in Window 8 for MSE/RMSE (MSE = 1.077e07, RMSE = 0.000328), but its best MAE occurs earlier (Window 5, MAE = 0.000224), suggesting slightly different sensitivity in absolute error. MLP's best window is Window 3 across all metrics (MSE = 2.864e07, RMSE = 0.000535, MAE = 0.000365), meaning it performs best at moderate training length but does not keep improving later. MLR's best window is Window 1 across metrics, implying that expanding training years does not reduce error for the linear model.

Under anomaly transformed features, RF remains the best model and improves further: MSE = 9.829e08, RMSE = 0.000313, MAE = 0.000206, outperforming all

| model | MSE | RMSE | MAE |
|---|---|---|---|
| MLR | 6.952e-07 | 0.000833 | 0.000579 |
| MLP | 3.506e-07 | 0.000592 | 0.000415 |
| RF | 9.829e-08 | 0.000313 | 0.000206 |
| XGB | 1.283e-07 | 0.000358 | 0.000242 |

Table 5.3: Mean across all windows for feature's anomaly transformed dataset

other methods. XGB remains second but is weaker than its RAW counterpart (MSE = 1.283e07, RMSE = 0.000358, MAE = 0.000242). MLP and MLR both degrade under ANOM, with MLP at (RMSE = 0.000592, MAE = 0.000415) and MLR worst (RMSE = 0.000833, MAE = 0.000579). This indicates anomaly learning benefits RF most, but reduces performance for the regression and neural baseline in this configuration. Mean ANOM ranking stays: RF > XGB > MLP > MLR.

| Model | MSE Win | MSE | RMSE Win | RMSE | MAE Win | MAE |
|---|---|---|---|---|---|---|
| MLR | 1 | 6.470e-07 | 1 | 0.000804 | 1 | 0.000561 |
| MLP | 3 | 3.154e-07 | 3 | 0.000562 | 3 | 0.000388 |
| RF | 8 | 8.428e-08 | 8 | 0.000290 | 8 | 0.000192 |
| XGB | 8 | 1.182e-07 | 8 | 0.000344 | 8 | 0.000233 |

Table 5.4: Best window per model for feature's anomaly transformed dataset. Win = Window

Best-window behavior under anomaly transformed dataset is clearer: RF achieves its strongest anomaly performance in Window 8 (best overall): MSE = 8.428e08, RMSE = 0.000290, MAE = 0.000192, which is also the best result among all ANOM models. XGB's best ANOM window is also Window 8 for all metrics (MSE = 1.182e07, RMSE = 0.000344, MAE = 0.000233), showing steady benefit from expanding training years. MLP again peaks at Window 3 (MSE = 3.154e07, RMSE = 0.000562, MAE = 0.000388), consistent with its RAW pattern. MLR peaks at Window 1 (MSE = 6.470e07, RMSE = 0.000804, MAE = 0.000561). Overall, window-wise best results show tree models gain from longer historical training, while MLP/MLR do not show monotonic improvement with additional years.

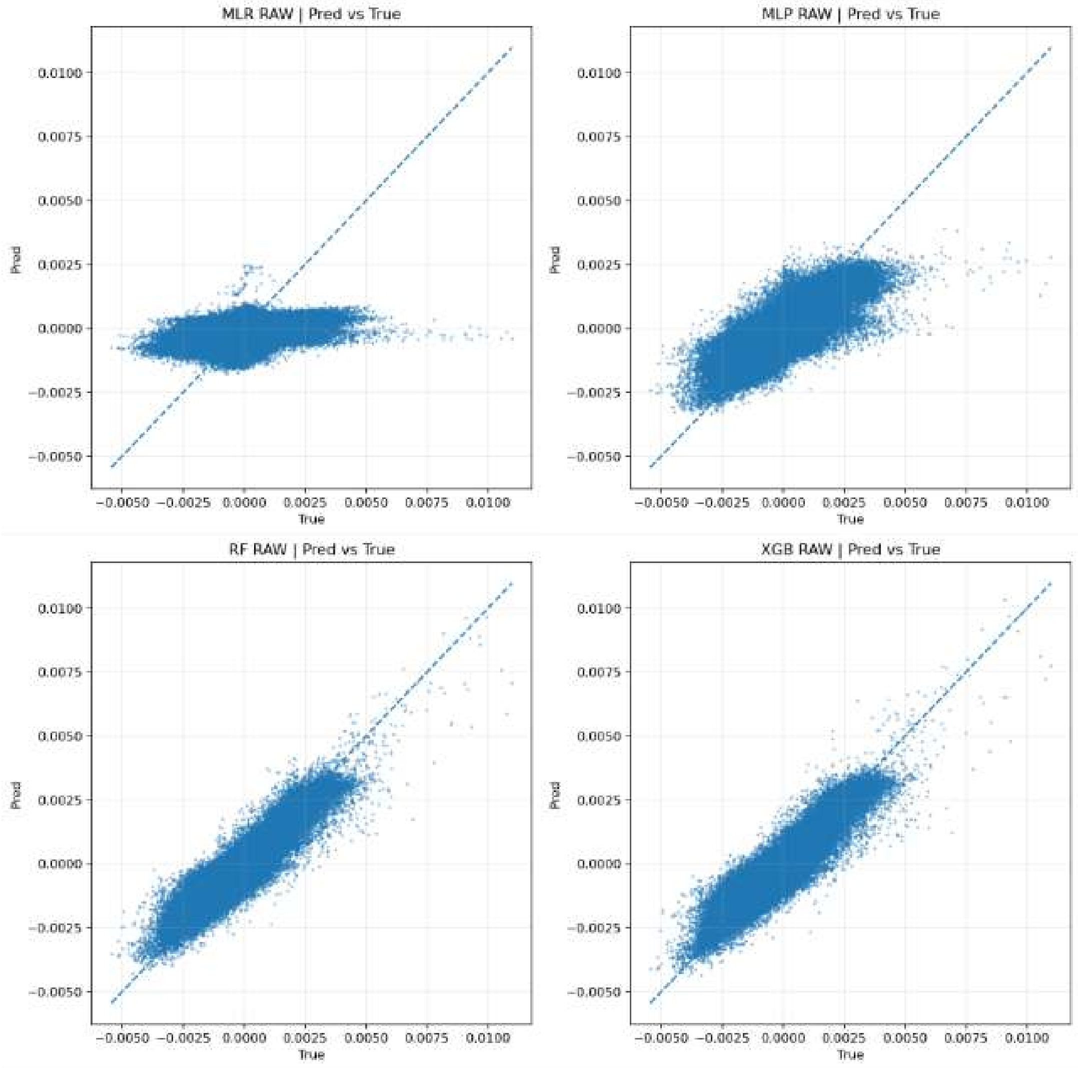



Figure 5.1: Scatter Plots of Machine Learning regression (MLR ) , Multi Layer Perceptron ( MLP) , Random Forest (RF) , XGBoost (XGB)

**Regime shift detection**

For regime shift detection, basin aggregation was chosen to reduce grid-scale noise and to ensure physically interpretable units aligned with carbon-cycle literature. Because raw basin-mean flux is dominated by seasonality and slow variations, regime analysis is performed on anomaly/IAV-type signals (seasonal structure removed, then detrended) to isolate interannual variability relevant to climate forcing. Regime shifts are detected on the basin IAV series using a non-parametric change-point test. Significant shifts are identified in:

- Atlantic: ∼2016-06
- Southern Ocean: ∼2020-10
- Arctic: ∼2021-07

No statistically significant single dominant step-like transition is detected for the Indian and Pacific basins over 2014–2024 under the same threshold. In the ENSO IAV plots, the IAV curve shows basin variability, Niño3.4 provides external climate context, and the vertical dashed line marks the detected regime-shift month.

## 5.3 Final Design Adjustments

For carbon flux prediction although $R^2$ is reported as a conventional summary metric, it provides limited scientific insight for gridded air–sea $CO_2$ flux in our setting. The target field is strongly climatological: large-scale spatial gradients and a dominant seasonal cycle account for a substantial fraction of total variance. As a result, a model can obtain a seemingly "acceptable" $R^2$ by reproducing the mean seasonal structure, even if it fails to capture the climate-relevant variability that matters for carbon-cycle interpretation. In our results, the overall $R^2$ scores were not consistently strong across LOYO test years, reinforcing that variance-explained alone does not reflect robust generalization.

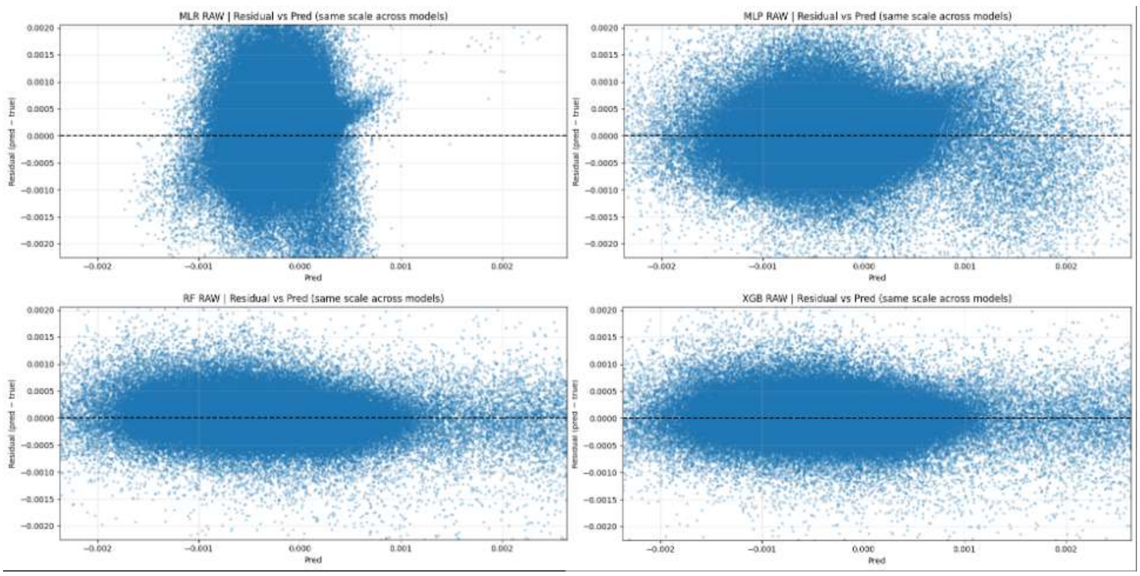



Figure 5.2: Residual plots of machine learning regression models: Multiple Linear Regression (MLR), Multi-Layer Perceptron (MLP), Random Forest (RF), and XGBoost (XGB).

These residual–vs–predicted plots compare how four models distribute their errors across the range of fitted values. In an ideal diagnostic, residuals form a patternless horizontal band centered on zero with roughly constant spread; any visible structure (curvature, slope, or "fan" shape) suggests systematic model misspecification or non-constant error variance.

MLR (linear regression) shows the strongest departure from this ideal. The predictions shows a relatively narrow range while residuals remain widely spread, producing a dense vertical blob. This is consistent with underfitting because a linear model cannot capture the nonlinear relationships so it collapses toward a mean like prediction and leaves structured error. Any slight asymmetry or thickening indicates that the linear form is missing signal rather than being only noisy.

MLP (nonlinear neural network) spreads predictions more than MLR and residuals are closer to a horizontal cloud whic indicates improvement in fit to nonlinear structure. However, the band is still fairly thick and slightly uneven across the prediction range and suggests that remaining bias in certain regimes or mild heteroscedasticity.

RF and XGB (tree-based nonlinear models) show the most "textbook" residual behavior: residuals are tightly centered around zero with minimal curvature. This indicates low systematic bias and that nonlinear effects are largely captured. There is some widening at the extremes, which often reflects fewer samples and harder-to-predict regimes rather than a global variance problem.

$R^2$ remains a valid goodness of fit summary across regression models, but for nonlinear methods it should be interpreted as predictive skill rather than variance explained under linear assumptions. That's why residual diagnostics are essential: they show whether the model captures the signal without leaving systematic error patterns. In our case, the residual and scatter plots confirm that tree based models (RF/XGB) capture nonlinear structure far better than MLR and somewhat better than MLP.

For the second part of our thesis for regime shift detection , early warning indicators were applied to predicted $CO_2$ flux to assess increasing instability. This analysis tells the spatiotemporal patterns of these indicators and basin scale variability in global ocean $CO_2$ fluxes which indicates how signal strength and persistence change across different time and ocean basins. %

Firstly, the analysis based on rolling windows revealed that the early-warning detection.Signals highly relies on the selected window length. Using a 6-month window, the portion of grid points with high early-warning patterns rose steeply since he percentage changed by 0.89% (1,328 grids) in 2020 to 19.1% (28,661 grids) in 2023, then slightly dropped to 18.5% (27,739 grids) in 2024. This is compared to more that was captured by the 24-month window. Immediate developments, increasing by 0.35 (529 grids) in 2020 to 4.55 (6,829 grids) in 2024 and the 12-month window had intermediate values. These patterns indicate that shorter windows are more responsive to short-term changes, and longer. window displays show trends of system variability that are longer-term.

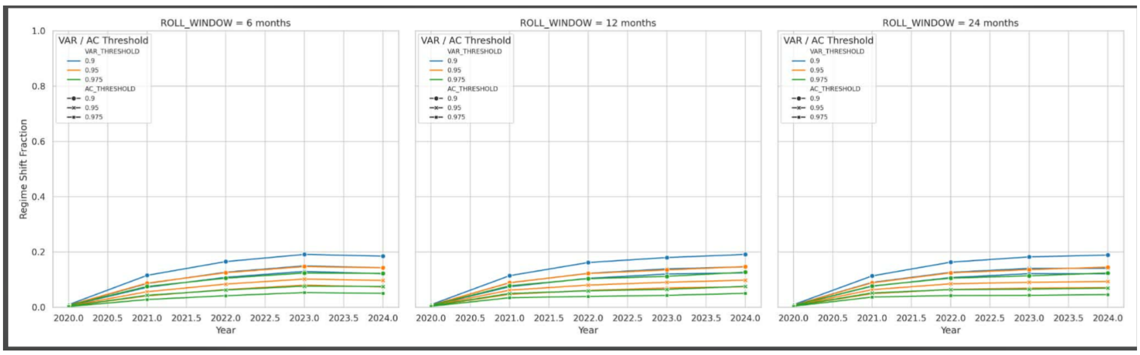


Figure 5.3: Comparison of early warning signal detection across 6 month,12 month and 24 month rolling windows that reflect the sensitivity of signal fragment towards window length.

Secondly, global mean metrics explains the overall system behavior which highlights trends in variability and temporal persistence across all the ocean. The global mean rolling variance increased normally from approximately $3.1 \times 10^{-8}$ in 2020 to $4.0 \times 10^{-8}$ in 2024 for the 6 month window. Similarly lag-1 autocorrelation increased from $-0.040$ in 2020 to $+0.083$ in 2024 for the 12 month window and to $+0.138$ for the 24 month window. The shift from negative to positive autocorrelation means increasing temporal persistence also a potential reduction, consistent with the behavior expected from EWS.

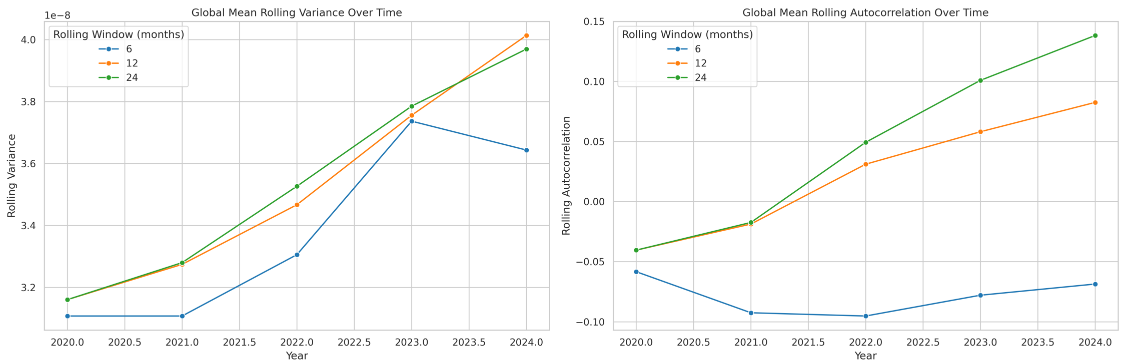
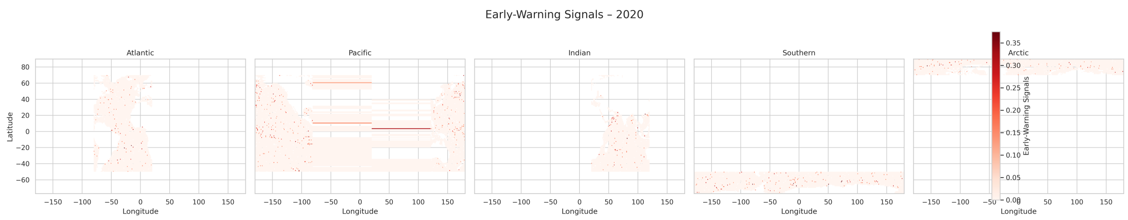



Figure 5.4: Trends in global mean rolling variance and lag-1 autocorrelation across different window lengths, reflecting increasing temporal persistence and system variability.

Thirdly the basin wise analysis on five major ocean basins ( the Pacific, Atlantic, Indian, Southern, and Arctic Oceans) . to evaluate regional differences in EWS while covering all major oceanic systems. Results shows a strong spatial variability specifically, the Pacific Ocean showed the highest part of flagged grid pointswhich increased from 0.7% (371 grids) in 2020 to 17.3% (9,127 grids) in 2024. The Indian and Atlantic Oceans moderately increased to 9.8% (1,966 grids) and 6.2% (1,757 grids). While the Southern Ocean remained relatively stable below 3.5%, the Arctic Ocean displayed persistent positive autocorrelation which rechead 7.5% in year 2024. Spatial maps confirmed these patterns highlighting the hotspots in the equatorial Pacific and Arctic regions which had maximum rolling variance rising from $1.1 \times 10^{-7}$ in 2020 to $1.75 \times 10^{-7}$ in 2024 more in the equatorial Pacific.



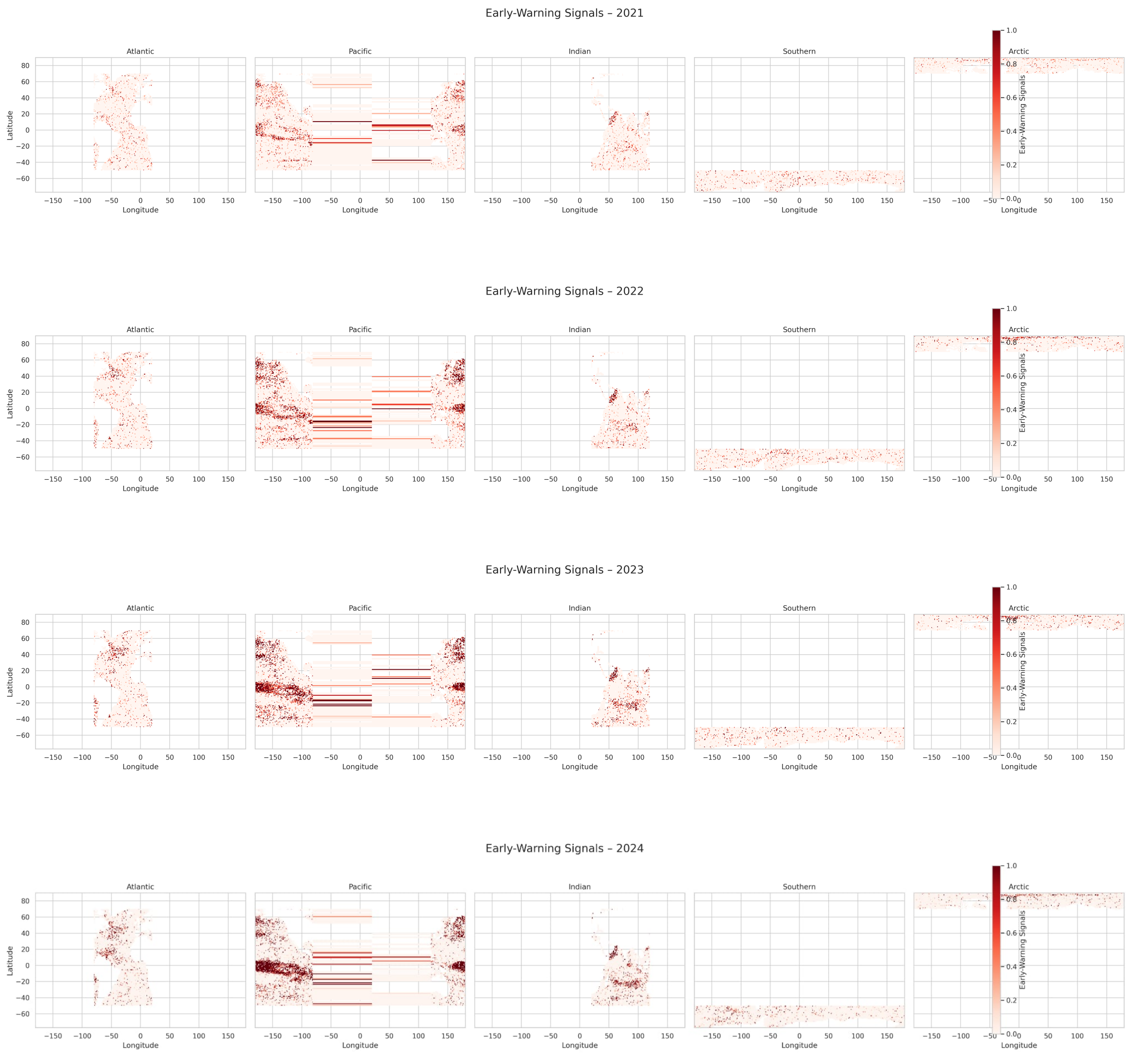



Figure 5.5: Spatial distribution of early-warning signals across major ocean basins from 2020 to 2024, highlighting regional hotspots and variability in signal strength.

It must be noted that that EWS show increasing variability and persistence but actual regime shift or abrupt changes may not be confirmed . Some patterns can result from data processing rather than real physical changes so these results should be viewd as indicative trends rather than exact transitions. Moreover our focus shifted toward the second methodology for statistical regime shift detection.

## 5.4 Statistical Analysis

### Carbon Flux prediction

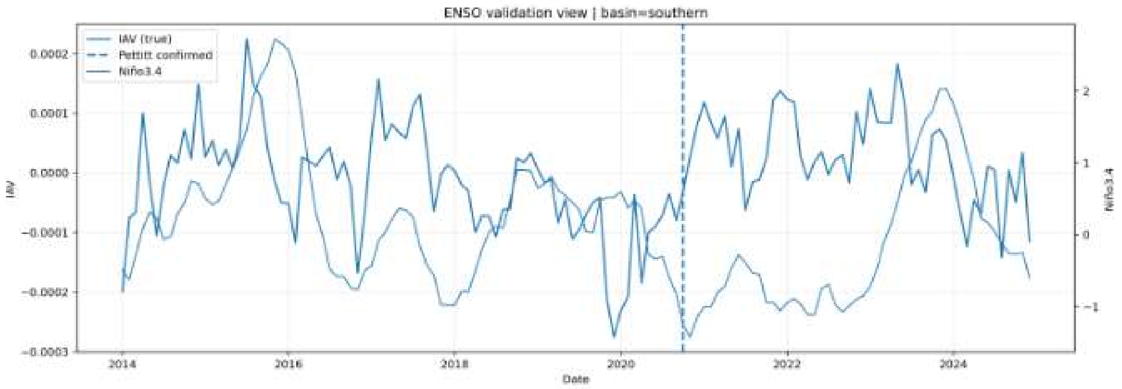
Looking across the windows (where each higher window means more training years), the clearest pattern is that longer training improves generalization for the tree-based models, especially RF (and also XGB). In both RAW and ANOM, RF achieves its best performance in Window 8 (the longest training span: 2014–2022), and its errors generally trend lower in later windows, meaning it benefits from having more historical variability to learn stable relationships that transfer to unseen years. XGB shows a similar behavior: its best MSE/RMSE occurs in Window 8, indicating better generalization when trained on longer time spans. In contrast, MLP does not show monotonic improvement with increasing training years; its best results occur around Window 3 (moderate training length), after which adding more years does not consistently reduce error suggesting either saturation or increased difficulty capturing additional nonstationarity. MLR performs best in the earliest window and degrades as the training period expands, implying that a simple linear structure does not gain robustness from longer training and may be more sensitive to distribution shifts. Overall, long training generalizes best for RF/XGB, while short-to-moderate training is sufficient for MLP, and MLR does not benefit from longer training under this LOYO setup.[4]

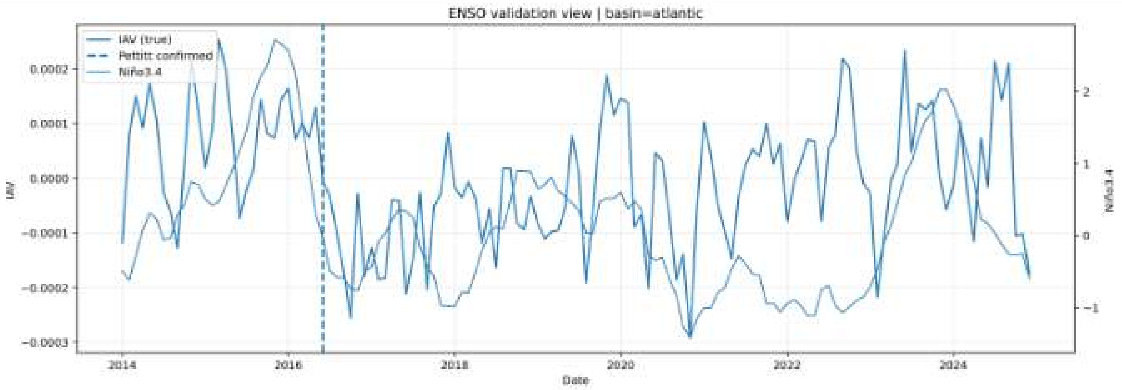
### Regime shift detection using method 2

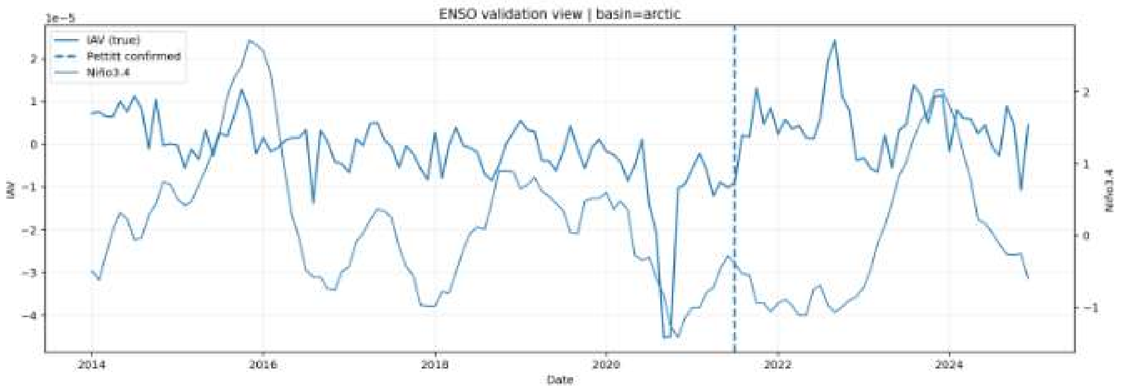
ENSO validation is not used as a ground-truth confirmation ("ENSO proves the shift"). Instead, it is a physical-coherence check.

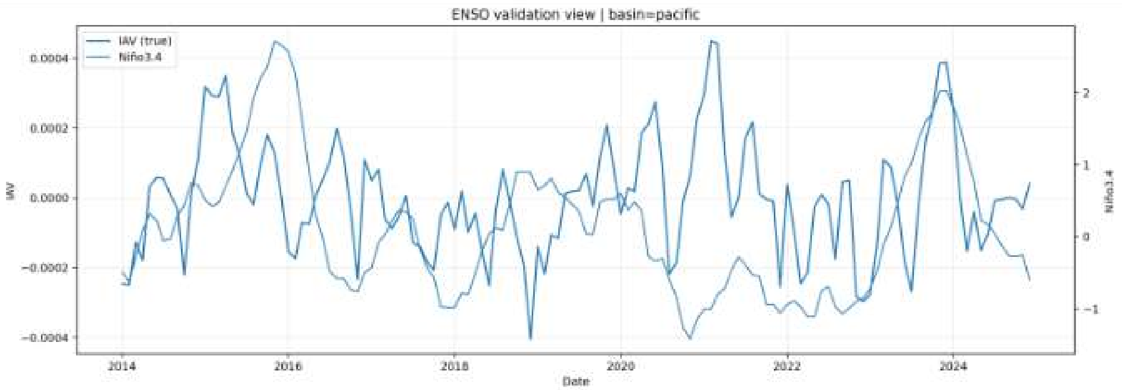
- ENSO (Niño3.4) is a dominant source of interannual climate variability.
- It is known to influence air–sea $CO_2$ exchange through changes in winds, SST, circulation, and upwelling (especially strong in the Pacific, with teleconnections affecting other basins).
- Therefore, if a detected flux-IAV regime shift is physically meaningful, it should often show some coherence with ENSO timing/phases, especially in basins where ENSO influence is expected.

Across basins, lagged ENSO–IAV correlations indicate that Niño3.4 explains part of the IAV structure (with basin-dependent lags), but the strength is not uniform. Pacific shows the strongest instantaneous relationship (largest $|r|$ at lag 0). Southern shows a notable lagged relationship (ENSO leading by several months), consistent with delayed basin-scale adjustment mechanisms. Atlantic/Indian show moderate correlations, which is plausible due to atmospheric teleconnections rather than direct ENSO forcing. Arctic shows weak/non-robust correlation, which is expected because Arctic variability is often dominated by other modes and cryosphere–atmosphere coupling [2].


ENSO validation view | basin=southern
IAV (true)
Pettitt confirmed
Niño3.4
IAV
Date
Niño3.4


ENSO validation view | basin=atlantic
IAV (true)
Pettitt confirmed
Niño3.4
IAV
Date
Niño3.4


ENSO validation view | basin=arctic
1e−5
IAV (true)
Pettitt confirmed
Niño3.4
IAV
Date
Niño3.4


ENSO validation view | basin=pacific
IAV (true)
Niño3.4
IAV
Date
Niño3.4

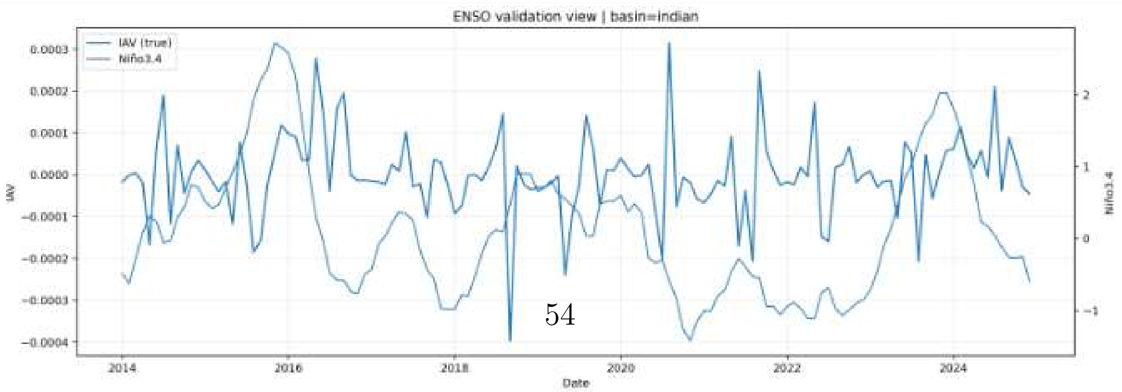

ENSO validation view | basin=indian
IAV (true)
Niño3.4
IAV
Date
Niño3.4

Figure 5.6: It demonstrates temporal context: detected regime-shift dates occur during periods when Ni˜no3.4 is transitioning or sustained in one phase. It also shows that ENSO does not explain everything (some basins show weaker visual coupling).

**Interpreting the figure:** ENSO is a dominant mode of interannual climate variability that originates in the tropical Pacific, but it also influences other ocean basins through atmospheric teleconnections. These teleconnections modulate surface winds, upper-ocean mixing, upwelling intensity, and sea surface temperature patterns/processes that directly affect air sea $CO_2$ exchange through changes in gas-transfer velocity, $CO_2$ solubility, and the surface ocean carbon supply. Accordingly, ENSO is not used as a "ground-truth confirmation" of regime shifts. Instead, it is treated as a physical coherence check, if detected shift timing clusters around strong El Niño/La Niña episodes or exhibits consistent lagged association between Niño3.4 and basin scale flux variability, the interpretation of the shift as climate-relevant becomes more credible. Conversely, weak ENSO alignment does not invalidate a detected shift, because basin scale $CO_2$ flux variability can be driven by other modes of climate variability and regional mechanisms, such as the Southern Annular Mode (SAM), Indian Ocean Dipole (IOD), North Atlantic Oscillation (NAO), or basin-specific circulation and ecosystem dynamics [3].

In addition to interpretation, this validation framework provides a pathway to formulate regime shift detection as a supervised learning problem. Once regime shifts are identified from the basin scale flux variability series, each basin month can be treated as an input sample and assigned a label indicating "shift" versus "no shift" (or, alternatively, multi-state regime classes). Predictors can be constructed from lagged basin anomaly/IAV features, early-warning indicators (e.g., rolling variance and lag-1 autocorrelation) and external climate indices such as Niño3.4. Under this formulation, a classifier can be trained to learn precursor patterns that increase the probability of an upcoming shift, enabling supervised early-warning detection rather than purely retrospective statistical change point identification.

## 5.5 Comparisons and Relationships

Here presented the best overall model across both dataset mentioning the best window as well.

**Best overall in actual dataset (lowest error)**

Random Forest (RF)

MSE: $9.482 \times 10^{-8}$ (Window 8)

RMSE: 0.000308 (Window 8)

MAE: 0.000206 (Window 8)

**Best overall in anomaly transformed dataset (lowest error)**
Random Forest (RF)
MSE: $8.428 \times 10^{-8}$ (Window 8)

RMSE: 0.000290 (Window 8)

MAE: 0.000192 (Window 8)

RF is the best model in both actual dataset and anomaly transformed dataset, anomaly transformed dataset slightly improves RF further.

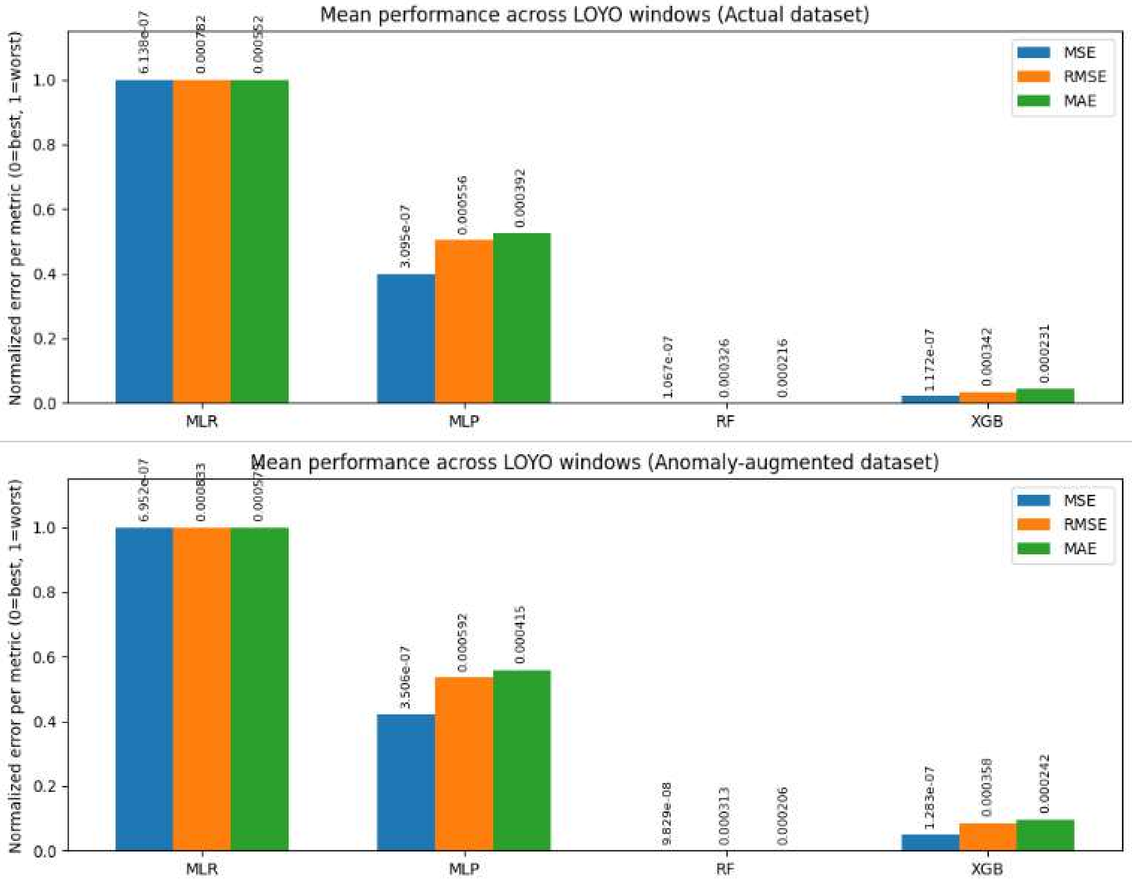



Figure 5.7: Mean performance across all windows of actual and anomaly transformed dataset

| Model | Better Mode Overall | Notes (Based on Best Scores) |
|---|---|---|
| MLR | RAW | RAW beats ANOM for MSE/RMSE/MAE |
| MLP | RAW | RAW beats ANOM for all 3 metrics |
| RF | ANOM | ANOM beats RAW for all 3 metrics |
| XGB | RAW | RAW beats ANOM for all 3 metrics |

Table 5.5: Actual vs anomaly transformed winner per model

## 5.6 Comprehensive Comparison of the Model and the Windows and the Dates for Predicting Ocean Carbon Flux

The actual dataset and anomaly transformed dataset comparison reveals a clear relationship between what each model is able to learn and why anomaly information can matter under LOYO. RAW features preserve the full mean state (seasonal cycle + background gradients), so models can gain skill by learning stable climatological structure. This is exactly what we see for MLR, MLP, and XGB, where the best results are consistently actual data > anomaly transformed across MSE/RMSE/MAE. In these models, removing the mean state component (pure anomaly transformed) reduces an easy to learn, high signal baseline, so the remaining variability is harder to predict and error increases.

In contrast, RF behaves differently, it is the best model in both modes, and it is the only one that becomes better under anomaly transformed dataset (all three metrics improve). That pattern supports the interpretation that RF benefits from learning climate relevant variability rather than relying on the mean seasonal structure when trained on anomalies of the features. RF can focus on relationships between flux and driver departures (e.g., unusual winds, SST anomalies, mixed-layer changes), which are more directly tied to interannual dynamics and regime-like behavior.

This observation provides strong justification for using an anomaly transformed dataset (features and their anomaly together) even though anomaly transformed dataset only does not universally outperform actual dataset only. The key idea is not "anomalies always win," but that anomalies often encode physically meaningful signals more cleanly than raw values. By including both raw and anomalies features, the model can learn two complementary components at once: (i) the baseline mapping between flux and the mean state which helps general predictive accuracy and (ii) the sensitivity to deviations from that state which supports generalization to unseen years and aligns with regime shift objective.

Under LOYO, this dual representation is valuable because it reduces over reliance on memorizing seasonal patterns and encourages the model to use variability that is more transferable across years. Results are consistent with this rationale: actual raw variables are important for several models (mean state skill), while anomaly learning can enhance robustness in at least one strong learner (RF), indicating that an anomaly transformed design is a defensible methodological choice for both prediction quality and climate-process relevance. Below here is the feature importance of Random Forest model (high performing among other models ) of window 8

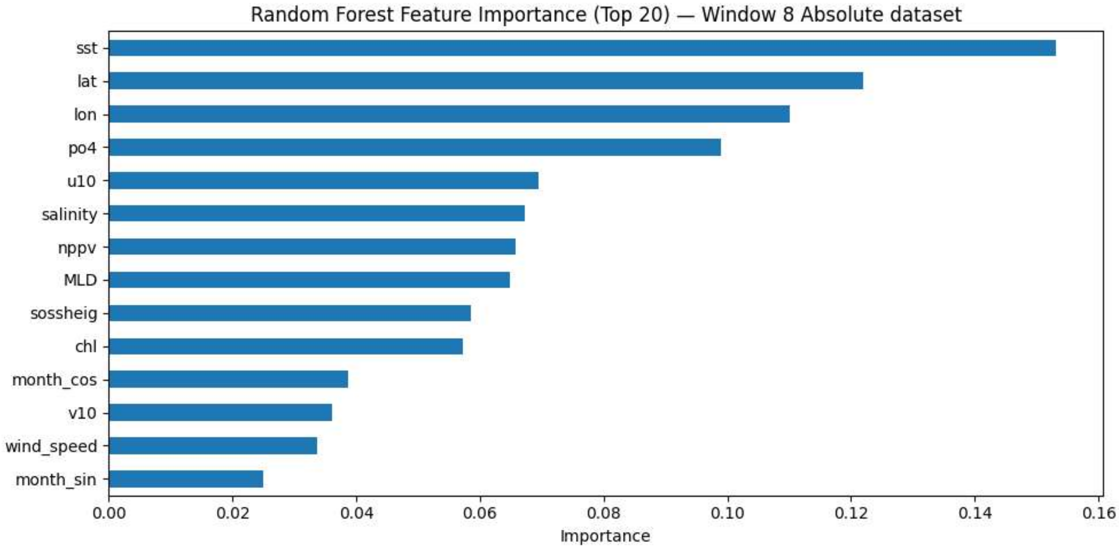



Figure 5.8: Feature Importance of Random Forest (window 8)

Random Forest feature importance summarizes which predictors the model relied on most to reduce prediction error, computed as the average decrease in impurity from splits using each variable (and then normalized so importances sum to 1). It does not indicate causal influence; it reflects predictive usefulness given the other features and the training data.

In the Window-8 absolute dataset, SST is the dominant predictor, consistent with the strong thermodynamic control of temperature on $CO_2$ solubility and air–sea exchange conditions. Latitude and longitude rank next, indicating that the model uses spatial position to encode large-scale biogeochemical and circulation regimes (e.g., basin structure, boundary currents, upwelling zones) that are not fully captured by the available drivers. Among process variables, $PO_4$, wind (u10), salinity, NPPV, and MLD contribute substantially, aligning with nutrient supply, wind-driven gas transfer and circulation, freshwater/buffering effects, biological production, and vertical mixing controls on surface carbon availability. Sea surface height (sossheig) and chlorophyll provide additional information on circulation and ecosystem state. Finally, month_sin/cos and wind components show smaller but non-zero importance, meaning seasonality is partly explained by physical biological drivers rather than calendar alone. Because correlated predictors can share importance, these rankings should be interpreted as relative contributions, not standalone drivers.

## 5.7 Discussions

The Leave out one year results provide practical guidance on selecting the training window, model, and dataset formulation for predicting ocean $CO_2$ flux from indirect environmental drivers (e.g., wind, SST, mixed-layer depth, chlorophyll, nutrients, and related reanalysis/satellite variables). Overall, nonlinear tree-based learners generalize most reliably across unseen years, indicating that the driver–flux mapping is dominated by nonlinear interactions and regime-dependent responses. Random Forest delivers the lowest errors and the most stable performance across windows, making it the strongest choice when the objective is robust year-to-year generalization and physically plausible responses to indirect forcing. XGBoost is consistently competitive (typically second-best) and can be preferred when sharper thresholds or higher-order interactions are expected, though its improvements under variability-focused learning are less pronounced than Random Forest. In contrast, multiple linear regression performs worst and does not benefit from longer training, implying that a linear mapping cannot represent the multivariate, nonlinear controls of air–sea $CO_2$ exchange. The neural baseline (MLP) improves on linear regression but does not show monotonic gains as the training span increases, suggesting sensitivity to temporal distribution shifts and limited benefit from additional years under the current setup.

Window-wise, the tree models achieve their best performance in later windows with the largest historical training span, supporting the use of the longest feasible training window when constructing a general-purpose predictor. A longer window likely captures a broader range of climate variability and reduces dependence on year-specific patterns, improving temporal transfer.

Regarding dataset choice, an anomaly-transformed dataset is defensible because it allows the model to learn both (i) the baseline relationship between flux and mean state and (ii) the sensitivity of flux to departures from that state. This supports LOYO generalization by reducing reliance on seasonal memorization and emphasizing physically meaningful variability. Empirically, Random Forest benefits most from this design, while other models retain more skill when mean-state structure remains dominant. Hence, for ocean carbon flux prediction driven by indirect variables and intended for robust downstream analysis, a long-window tree-based model with anomaly augmentation is the most generalizable configuration.

Addition to that, regime shifts were detected retrospectively using a change-point framework based on the Pettitt test applied to basin-scale $CO_2$ flux variability, which identified three statistically significant shifts across three basins. To evaluate physical plausibility, the detected shift timings were compared against ENSO extremes using the Niño3.4 index. Because ENSO influences air–sea $CO_2$ exchange through atmospheric teleconnections that modulate winds, mixing, upwelling, and sea surface temperature—and thereby affect gas-transfer velocity, solubility, and surface carbon supply—ENSO was not treated as ground truth, but as a physical-coherence check. The observed alignment of several shift timings with strong El Niño/La Niña episodes increases confidence that these transitions reflect climate-relevant variability rather than isolated statistical artifacts. This coherence also motivates a forward-looking formulation in which ENSO events, alongside other climate-mode indicators and early-warning features (e.g., lagged anomalies, rolling variance, and lag-1 autocorrelation), could be used to flag elevated regime-shift risk with a lead time of 3–6 months. However, the present dataset yields only three confirmed "shift" labels and a larger number of non-shift or ambiguous cases; training a supervised binary classifier under such class imbalance would likely introduce bias and unreliable generalization. Therefore, developing robust early-warning prediction using ENSO-informed drivers and richer event-labeled datasets is identified as a key direction for future research.

# Chapter 6

# Conclusion

## 6.1 Summary of Findings

This thesis demonstrates that a time-aware, multi-source machine learning framework can reconstruct monthly global ocean-atmosphere $CO_2$ flux variability from indirect physical and biogeochemical drivers with meaningful generalization across unseen years. Using a leave-one-year-out evaluation design, the study prioritizes realistic temporal transfer rather than random splits that risk inflating skill in non-stationary climate data. Across the tested model families, the Random Forest model provides the strongest overall performance and the most consistent behavior across validation windows. Its best reconstruction skill occurs when trained on the longest available historical span, indicating that additional years of training data improve the model's capacity to learn stable, reusable relationships between drivers and flux patterns. This supports the broader interpretation that, for global $CO_2$ flux reconstruction, robustness under time-aware testing is more informative than peak performance under potentially optimistic sampling, and that tree-based nonlinear learners are well-suited to capturing the interacting, non-monotonic controls that govern air–sea exchange.

The thesis also finds that feature formulation materially influences what models learn, particularly in a climatological context where the seasonal cycle and mean-state structure dominate variance. Comparing an absolute-driver configuration against an anomaly-transformed representation reveals a dual behavior. Several models benefit from absolute variables because they can more easily learn the mean-state flux structure that co-varies with large-scale temperature gradients, wind forcing, and basin-scale physical setting. At the same time, the Random Forest model improves further when anomaly information is included, indicating that anomaly-augmentation can help it emphasize departures from typical seasonal conditions rather than implicitly relying on seasonal memorization. This result is scientifically important because the thesis goal is not only to reconstruct a smooth climatology, but also to preserve climate relevant variability that supports later regime-shift analysis. These finding shows a dual concept which frames that drivers capture baseline structure whereas anomaly features provide additional leverage for learning inter-annual variability signals that are more directly linked to climate anomalies and persistent shifts in basin scale behaviorchanges.

Utilizing machine learning for prediction, the thesis interprets a basin scale interpretation layer to assess whether the reconstructed flux dynamics exhibit persistent changes consistent with regime behavior. Regime-shift detection is intentionally not performed on raw flux magnitude, because raw fields are dominated by strong seasonality and longer-term trend components that can generate spurious change points. Instead, basin aggregation is used to improve interpretability and suppress grid-scale noise, and regime analysis is applied to deseasoned and detrended interannual-variability (IAV) signals. Candidate change points are then identified using Pettitt's test, followed by a persistence-window confirmation step that labels only robust, sustained level shifts rather than transient spikes or short-lived extremes. This two-stage design directly reflects the thesis definition of a regime shift as a persistent change in statistical behavior, and it reduces false alarms by requiring temporal stability around detected transitions.

Finally, detected shifts are interpreted within an explicit physical climate context using ENSO indicators, particularly the Niño3.4 index. Importantly, ENSO is treated as a physical-coherence check rather than a ground-truth validator: agreement with major El Niño/La Niña extremes or consistent lead–lag associations increases confidence that detected basin transitions are climate-linked and plausible, while lack of ENSO alignment does not invalidate a shift because other modes or regional dynamics may dominate in different basins. This framing empowers the scientific justifiability of the regime shift element based on rooting statistical discoveries in identified natural causes of climate variation without excessive attribution of cause. Overall, it's results indicate that time aware nonlinear reconstruction, anomaly sensitive representations conditioning by basin scale IAV and interpretation in a physically context in combination creates a consistent flow of transition between global flux prediction and climate relevant identification of sustained ocean carbon variations.

## 6.2 Contributions to the Field

The study has some contribution to the environmental and climate science by showing that monthly global ocean-atmosphere CO2 flux can be recreated by machine learning indirect global patterns with indirect physical and biogeochemical drivers. There are scanty and irregular observations. The framework enables scalable by using flux variability learning on predictors of flux variability that are readily available via satellites and reanalyses. Observation of carbon exchange between the oceans and assists in filling gaps between ship-based global flux fields and measurements. The significance of this is as follows: carbon-budget accounting relies on reliable flux reconstructions, as do the evaluation of whether or not the ocean is behaving like a net sink or it is periodically shifting to weaker uptake or source-like behavior. In its turn, better sink-source characterization makes it possible. Further study of the vulnerability of carbon cycles, such as to circulation. changes, wind-driven gas exchange, stratification and mixing, and ecosystem-driven biological uptake processes which are directly related to long-term questions of ocean buffering ability and the dangers of ocean acidification and the associated stressors.In this sense the work provides a practical framework for synthesizing multi source climate information into interpretable flux fields that can complement established inversion and model products and support climate assessment, mitigation evaluation, and regional mechanism diagnosis.

Another methodological contribution of the thesis is the development of a climate-appropriate statistical and machine-learning research design of non-stationary spatiotemporal systems. It regularises a strict time-conscious generalisation approach that minimises the threat of overfitting skill to climatological data, where seasonal structure may dominate variance, and where relationships may vary between years under varying forcing. It also adds an explicit distinction between mean-state reconstruction and climate-relevant departures that considers both absolute and anomaly-transformed representations, which upholds the fact that interannual variability can only be captured with different information than climatology. In addition to prediction, the thesis presents a statistically conservative regime-analysis workflow that focuses on regime shifts as lasting changes in behaviour of variability as opposed to transient extremes, making heavy use of signal conditioning and confirmation logic in order to minimise false detections. Finally, it adds a physically grounded interpretation layer by using large-scale climate variability context as a coherence check rather than as ground-truth labeling, strengthening the linkage between statistical outputs and climate mechanisms such as temperature-driven solubility effects and circulation/upwelling-driven disequilibrium changes. Collectively, these contributions position machine learning not as a replacement for process understanding, but as an integrative framework for producing climate-relevant flux reconstructions and variability diagnostics that are scientifically interpretable, methodologically defensible, and directly useful for carbon-cycle and ocean-change research

## 6.3 Recommendations for Future Work

A natural next step is to move from post hoc regime-shift detection toward supervised early-warning prediction. After constructing basin-month labels from Pettitt-detected and persistence-confirmed events (e.g., “shift” vs “no shift”), a binary classification model can be trained to learn precursor patterns that precede confirmed transitions. Inputs should include lagged basin predictors derived from anomaly/IAV features, rolling statistics commonly used as early-warning indicators (e.g., changes in variance, lag-1 autocorrelation, persistence), and external climate indices such as Niño3.4. The goal would be to estimate an elevated probability of an upcoming shift within a defined lead-time window, enabling a probabilistic “risk” perspective rather than a purely retrospective identification of change points.

Future work should also extend the physical-context layer beyond ENSO to better support interpretation and predictive features in basins where ENSO is not the dominant mode. Because ENSO teleconnections vary strongly by region, additional large-scale climate modes can be incorporated to improve coherence checks and driver attribution outside ENSO-sensitive domains. For example, NAO is particularly relevant for the North Atlantic, while SAM exerts strong influence over Southern Ocean circulation and air–sea exchange pathways. Including these indices as contextual variables and potentially as predictive features can strengthen both the explanatory narrative of detected shifts and the robustness of any early-warning classifier across basins.

Another promising direction is spatiotemporal learning and uncertainty quantification. While tabular models are effective baselines, architectures that explicitly exploit spatial structure and temporal dependence (e.g., convolutional models for gridded patterns, sequence models for lagged dynamics) may better represent coherent propagation and persistence in climate signals. In parallel, uncertainty quantification should be added to express confidence in both reconstructed flux fields and detected shifts using approaches such as model ensembles, block bootstrapping, or probabilistic prediction frameworks—so that conclusions can be communicated with calibrated confidence levels rather than point estimates alone.

Finally, to reduce sensitivity to any single reference product and strengthen the generality of conclusions, multi-product robustness testing is recommended. Training and evaluation can be repeated using alternative flux products, or by defining multi-product targets (e.g., consensus or weighted blends) to quantify how much the reconstruction and regime-shift results depend on inversion choice. Demonstrating agreement across multiple products would substantially increase confidence that the learned relationships and detected variability changes reflect robust features of the ocean carbon system rather than artifacts of one reconstruction pathway.

# Bibliography


[1] N. O. B. P. Group, *Nasa Ocean Color web*, https://oceancolor.gsfc.nasa.gov/, Accessed: 2026-01-30, NASA Goddard Space Flight Center, Ocean Biology Processing Group, 1996.

[2] S. N. Rodionov, "A brief overview of the regime shift detection methods," in *Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability*, J. Alheit, Ed., UNESCO-IOC, ICES Annual Science Conference CM 2005/N:15, Paris: UNESCO-IOC, 2006, pp. 17–24.

[3] H. Xie, D. Li, and L. Xiong, "Exploring the ability of the pettitt method for detecting change point by monte carlo simulation," *Stochastic Environmental Research and Risk Assessment*, vol. 28, no. 7, pp. 1643–1655, 2014. DOI: 10.1007/s00477-013-0814-y

[4] Q. Liu et al., "A review of the detection methods for climate regime shifts," *Discrete Dynamics in Nature and Society*, vol. 2016, p. 3 536 183, 2016. DOI: 10.1155/2016/3536183

[5] Q. Liu, S. Wan, and B. Gu, "A review of the detection methods for climate regime shifts," *Discrete Dynamics in Nature and Society*, vol. 2016, p. 3 536 183, 2016. DOI: 10.1155/2016/3536183

[6] M. C. Gregg et al., "Assessment of the response of surface-ocean co2 fluxes to tropical pacific climate modes," *Journal of Climate*, vol. 30, no. 21, pp. 8595–8613, 2017. DOI: 10.1175/JCLI-D-16-0543.1

[7] Océania Consortium, "OcéanIA: AI, Data, and Models for Understanding the Ocean and Climate Change," Inria, Tech. Rep. hal-03274323, 2021.

[8] S. Cooley et al., "Ocean and coastal ecosystems and their services," in *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, H.-O. Pörtner et al., Eds., Cambridge, UK and New York, NY, USA: Cambridge University Press, 2022, pp. 673–1030. DOI: 10.1017/9781009325844.009

[9] L. Dunkl et al., "Contrasting projections of the enso-driven co2 flux variability in the equatorial pacific under high-warming scenario," *Earth System Dynamics*, vol. 13, no. 4, pp. 1097–1119, 2022. DOI: 10.5194/esd-13-1097-2022

[10] W. Li et al., "Global air–sea co2 flux from 1957 to present: A data-based estimate using the mpi-som-ffn neural network," *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 5, e2021MS002960, 2022. DOI: 10.1029/2021MS002960

[11] M. Qiu et al., “Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions,” *Atmospheric Chemistry and Physics*, vol. 22, no. 16, pp. 10 551–10 566, 2022. DOI: 10.5194/acp-22-10551-2022

[12] C. Rödenbeck, S. Zaehle, R. Ralay, and R. F. Keeling, “Data-based estimates of interannual sea-air co2 flux variations 1957–2020 and their relation to environmental drivers,” *Biogeosciences*, vol. 19, no. 10, pp. 2627–2652, 2022. DOI: 10.5194/bg-19-2627-2022

[13] J. Zeng, T. Matsunaga, and T. Shirai, *A new estimate of oceanic co2 fluxes by machine learning reveals the impact of co2 trends in different methods*, Preprint, not accepted for publication, 2022. DOI: 10.5194/essd-2022-71

[14] W. Li, J. Zhang, and X. Chen, “Ocean co2 flux estimation using satellite data and machine learning,” in *2023 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 2023, pp. 1–4. DOI: 10.1109/IGARSS52108.2023.10366513

[15] Authors to be determined, *Title to be determined*, EGUsphere preprint, Preprint under review, 2024. DOI: 10.5194/egusphere-2024-1369

[16] Authors unavailable, “Limnology and Oceanography: Methods article,” *Limnology and Oceanography: Methods*, 2024. DOI: 10.1002/lom3.10665

[17] K.-H. Cohrs, G. Varando, N. Carvalhais, M. Reichstein, and G. Camps-Valls, “Causal hybrid modeling with double machine learning,” *arXiv preprint arXiv:2402.13332*, 2024. DOI: 10.48550/arXiv.2402.13332

[18] E. Ellison, M. Mazloff, and A. Mashayek, “The rapid response of southern ocean biological productivity to changes in background small-scale mixing,” *npj Climate and Atmospheric Science*, vol. 7, p. 184, 2024. DOI: 10.1038/s41612-024-00824-w

[19] S. Iyer, P. Landschützer, and J. Zscheischler, *Spatiotemporal upscaling of sparse air-sea pco2 data via physics-informed transfer learning*, Preprint, 2024.

[20] Y. Liu, R. Thompson, S. Cohen, et al., “Spatiotemporal variations of the air–sea co2 flux over global oceans,” *Communications Earth & Environment*, vol. 5, no. 1, p. 271, 2024. DOI: 10.1038/s43247-024-01257-2

[21] Y. Liu, R. L. Thompson, P. Landschützer, J. Zeng, Y. Deng, and R. F. Keeling, “Spatiotemporal variations of the air–sea co2 flux over global oceans,” *Communications Earth & Environment*, vol. 5, no. 1, p. 271, 2024. DOI: 10.1038/s43247-024-01257-2

[22] Y. Liu, R. L. Thompson, P. Landschützer, J. Zeng, Y. Deng, and R. F. Keeling, “Spatiotemporal variations of the air–sea co2 flux over global oceans,” *Communications Earth & Environment*, vol. 5, p. 271, 2024. DOI: 10.1038/s43247-024-01257-2

[23] Y. Liu, R. L. Thompson, P. Landschützer, J. Zeng, Y. Deng, and R. F. Keeling, “Spatiotemporal variations of the air–sea co2 flux over global oceans,” *Communications Earth & Environment*, vol. 5, p. 271, Jun. 2024. DOI: 10.1038/s43247-024-01257-2

[24] World Ocean Review, *How the ocean absorbs carbon dioxide*, https://worldoceanreview.com/en/wor-8/, World Ocean Review 8: The Ocean – A Climate Champion? Accessed: 2026-01-28, 2024.

[25] S. Cohen, Y. Liu, and R. Thompson, *Quantitative assessment of satellite-observed atmospheric co2 concentrations over oceanic regions*, Preprint, 2025. DOI: 10.13140/RG.2.2.12345.6789

[26] P. Friedlingstein et al., "Global carbon budget 2024," *Earth System Science Data*, vol. 17, no. 3, pp. 965–1035, 2025. DOI: 10.5194/essd-17-965-2025

[27] I. Higgs et al., *Hybrid machine learning data assimilation for marine biogeochemical models*, Preprint under review, 2025. DOI: 10.5194/egusphere-2025-1676

[28] H. Wei et al., "Scientific advances and future trends in ocean carbon sink: An interdisciplinary review," *Frontiers in Marine Science*, vol. 12, 2025. DOI: 10.3389/fmars.2025.1658207