Agenda

1. Overview of Dataset

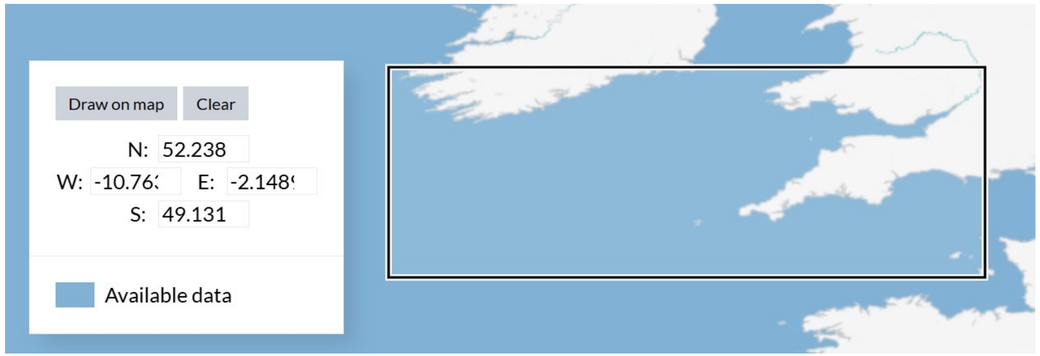
- Geographic Focus: South-West Coast of England
- About L3 and L4 Dataset
- Choosing L3 over L4 Dataset
- Challenges Faced while using L3 Data
- Ocean Variables Selected
- Handling Missing Data

2. Modelling Approaches

- Factor Models for Tensor Time Series
- CNN / RNN

1. Overview of Dataset

Geographic Focus: South-West Coast of England



About L3 and L4 Dataset:

- Level-3 data is processed satellite data that has been mapped to a uniform time and space grid, such as daily averages over a fixed grid (e.g., 4km x 4km).
- Level-4 data is gap-filled, interpolated, and model-assimilated satellite data. It provides complete coverage, even where raw satellite observations are missing.

Choosing L3 over L4 Dataset:

We selected L3 (Level-3) data instead of L4 (Level-4) due to the following core reason:

- L4 data doesn't offer daily-level optics and reflectance variables.
- L3 Data provides daily satellite-derived observations at a higher resolution and less smoothing.
- L4 Data, while more interpolated and gap-filled, is available mostly in monthly means or modeled estimates and often does not include optical parameters like RRS490 or KD490, which are vital for our research.
- Since the goal is to observe fine-scale temporal patterns over 2 years (Jan 2023- Dec 2024), daily granularity was non-negotiable.

Challenges Faced while using L3 Data:

- Missing Values- Gaps due to sensor limitations, causing spatial- temporal discontinuity.
- High Memory Usage- Unprocessed and Ungridded data causing processing failures.
- Data Size and Storage Issues- Storing and Processing large data leading to operational inefficiency

Ocean Variables Selected:

We selected the following 8 variables from the available satellite-derived oceanographic parameters, focusing on optical and biogeochemical properties:

Variable	Description	Why Selected
KD490	Diffuse attenuation coefficient at 490 nm	Represents **light penetration** in water. Key for studying turbidity, depth of light availability, and productivity zones.
ZSD	Secchi Disk Depth	A traditional and intuitive measure of **water clarity**. Complements KD490 and is relevant for biological activity monitoring.
RRS490	Remote sensing reflectance at 490 nm	Critical for **optical water property detection**, sensitive to suspended particles and chlorophyll.
RRS443	Remote sensing reflectance at 443 nm	Helps in detecting **chlorophyll-a** and other pigments. Essential for **primary productivity** estimation.
CHL	Chlorophyll-a concentration	Direct proxy for **phytoplankton biomass**, crucial for marine ecosystem and carbon cycle studies.
MICRO	Microplankton proportion	Indicates **plankton size structure**, affecting the food web and carbon export.
BBP	Backscattering coefficient	Used for **suspended particulate matter** quantification, important for turbidity and pollution monitoring.
CDM	Coloured dissolved organic matter	Reflects **decomposition** and **terrestrial input**, crucial near coasts like south-west England.

Satellite Data Gaps: Caused by Clouds, Winter Sun Angles, and Atmospheric/Sensor Constraints

Step-by-Step guide

Data Preparation

- Removed land pixels (flags=1)
- Sorted by (latitude, longitude, time) for spatiotemporal analysis

Gap Analysis & Cleaning

- For each location identified and removed periods where a parameter had ≥15 consecutive missing values

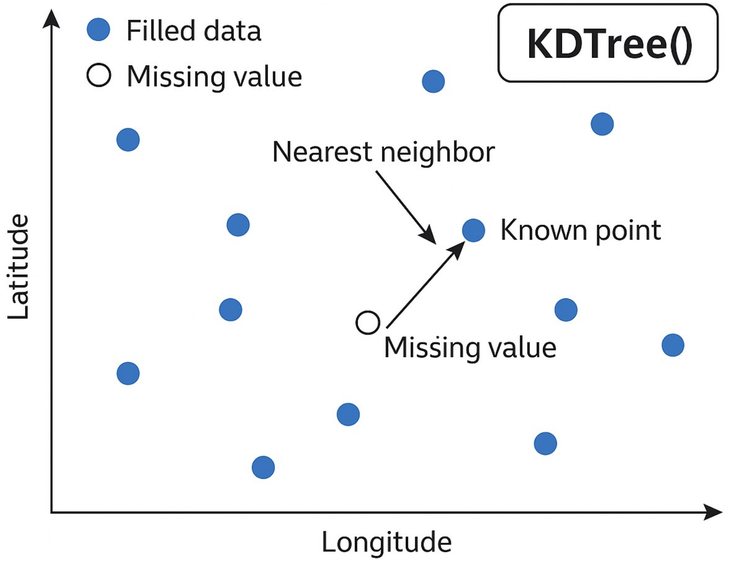
Temporal Imputation

- Applied forward-fill within each (lat,lon) group for temporal continuity in oceanographic parameters (slow-changing variables)

Spatial Imputation

- Applied daily k-NN (k=1) using KDTree.
- Uses closest real observation
- coastal vs. open-ocean differences remain intact

Dates with data missing for 100% of the location, primarily during winter months due to low solar angles



Agenda

1. Overview of Dataset

- Geographic Focus: South-West Coast of England
- About L3 and L4 Dataset
- Choosing L3 over L4 Dataset
- Challenges Faced while using L3 Data
- Ocean Variables Selected
- Handling Missing Data

2. Modelling Approaches

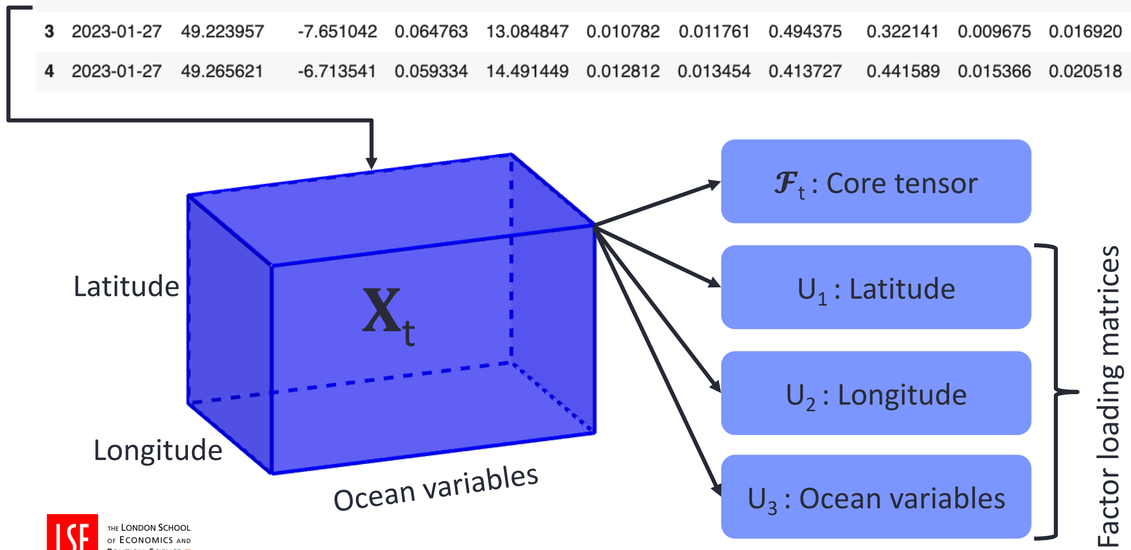
- Factor Models for Tensor Time Series
- CNN / RNN

Factor Models for Tensor Time Series

When handling multiple time series simultaneously, dimension reduction is a critical tactic to extract common information from the data without being overwhelmed by idiosyncratic variations. One such approach is a Dynamic Factor Model, which takes a tensor (high-dimensional arrays) at each time point and summarises the common information stored in these tensors into a small number of factors. The assumption is that the co-movement of the multiple time series is driven by these factors and their inherited dynamic structures. Therefore, the factor model allows us to systematically study the dynamics of tensor systems by jointly modelling the entire tensor simultaneously, while preserving the tensor and time series structure.

We first convert our following interpolated dataset into a 3-dimensional tensor for every time point, as shown below, and then conduct tucker decomposition to extract the core tensor and factor matrices:

	time	latitude	longitude	KD490	ZSD	RRS490	RRS443	CHL	MICRO	BBP	CDM
0	2023-01-27	49.255207	-7.223958	0.055129	15.796281	0.010847	0.012907	0.354588	0.262248	0.011544	0.015350
1	2023-01-27	49.161457	-7.598958	0.049587	17.909706	0.010573	0.011508	0.281386	0.317375	0.010022	0.017370
2	2023-01-27	49.223957	-7.640625	0.064803	13.075287	0.010620	0.011355	0.494998	0.329566	0.009359	0.017536
3	2023-01-27	49.223957	-7.651042	0.064763	13.084847	0.010782	0.011761	0.494375	0.322141	0.009675	0.016920
4	2023-01-27	49.265621	-6.713541	0.059334	14.491449	0.012812	0.013454	0.413727	0.441589	0.015366	0.020518



Perform Tucker Decomposition on the tensor $\mathbf{X}_t$

$\mathcal{M}_t$ and $\mathcal{E}_t$ are the corresponding signal and noise components of X_t, respectively. We assume that $\mathcal{E}_t$ are uncorrelated across time. Tucker decomposition approximates the signal $\mathcal{M}_t$ as follows:

$$X_t = \mathcal{M}_t + \mathcal{E}_t = \mathcal{F}_t \times_1 U_1 \times_2 U_2 \times_3 U_3 + \mathcal{E}_t \tag{5}$$

- $\mathcal{F}_t \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the time-dependent core tensor (compressed representation) capturing latent ocean dynamics, which we would like to model over time.
- U_k are the factor loading matrices, fixed over time:
  - $U_1 \in \mathbb{R}^{d_1 \times r_1}$ is the latitude factor matrix
  - $U_2 \in \mathbb{R}^{d_2 \times r_2}$ is the longitude factor matrix
  - $U_3 \in \mathbb{R}^{d_3 \times r_3}$ is the ocean characteristic factor matrix

Vector AutoRegressive Model (VAR) on $\mathcal{F}_t$

Since $\mathcal{F}_t$ is much smaller than X_t, we model its temporal evolution using a Vector AutoRegressive (VAR) model. First, we vectorise $\mathcal{F}_t$:

$$f_t = \text{vec}(\mathcal{F}_t) \in \mathbb{R}^{r_1 r_2 r_3}$$

We then apply a **VAR(p) model**, to capture the temporal dependencies in the core tensor $\mathcal{F}_t$:

$$f_t = A_1 f_{t-1} + A_2 f_{t-2} + \cdots + A_p f_{t-p} + \epsilon_t \tag{6}$$

- A_i are the VAR coefficient matrices.
- $\epsilon_t \sim N(0, \Sigma)$ is white noise.

Note that the size of the training dataset is 10 years (Jan 2014 - Dec 2023), while the test dataset is 1 year (Jan 2024 - Dec 2024). After training the VAR(p) model, we can run predictions for $\hat{f}_{t+1}$ and reshape it to a tensor.

$$\begin{aligned} \hat{f}_{t+1} &= A_1 f_t + A_2 f_{t-1} + \cdots + A_p f_{t-p+1} + \epsilon_{t+1}, \\ \hat{\mathcal{F}}_{t+1} &= \text{reshape}(\hat{f}_{t+1}, (r_1, r_2, r_3)). \end{aligned} \tag{7}$$

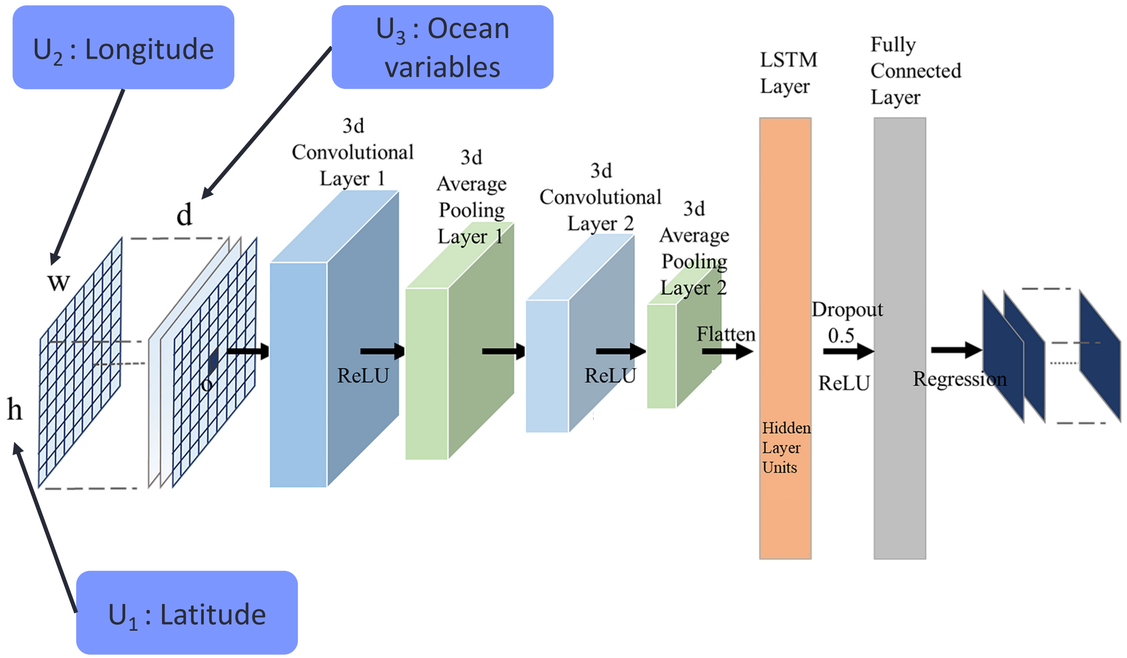
We derive the **forecasted ocean characteristics tensor** $\hat{X}_{t+1}$ by using mode-k multiplication:

$$\hat{X}_{t+1} = \hat{\mathcal{F}}_{t+1} \times_1 U_1 \times_2 U_2 \times_3 U_3 \tag{8}$$

ConvLSTM Architecture for Oceanographic Forecasting

Architecture Summary

The ConvLSTM (Convolutional LSTM) model combines spatial and temporal processing in a single integrated architecture. It processes sequences of spatial grids to predict future spatial patterns of oceanographic variables like chlorophyll and microbial concentrations.



$(U_1, U_2, U_3) = \mathcal{F}_t$: Core tensor

Why ConvLSTM Works for Spatio-temporal Prediction?

1.) Spatial Understanding:

- Unlike standard LSTMs that flatten spatial data, ConvLSTM preserves the 2D structure
- Convolutional operations identify local spatial patterns such as algal blooms, currents, and concentration gradients
- Spatial filters detect relevant features at different scales across the ocean surface

2.) Temporal Learning:

- LSTM gates (input, forget, output) retain information about previous days' patterns
- The cell state acts as a "memory" that accumulates important long-term trends
- This allows the model to learn recurring patterns, seasonal effects, and progressive changes

3.) Combined Benefits:

- Spatial relationships and temporal evolution are learned simultaneously
- The model can predict how patterns will move, grow, or dissipate across the ocean
- Perfect for oceanographic data where both location and time history matter

This unified approach allows the model to understand how spatial patterns of ocean variables change over time, making it superior to separate CNN and RNN models for forecasting future ocean conditions.