1st All Hands Meeting 31.01.2025

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE & Meta

ST498 Group 17: Modelling the World‘s Oceans

Agenda

1. Project Overview

2. Project Context

3. Methodology & Plan

4. Mandatory: Gantt-Chart

Agenda

1. Project Overview

- **Project Title**
- **Candidate Numbers**
- **LSE Supervisor & Industry Partner Names**

2. Project Context

3. Methodology & Plan

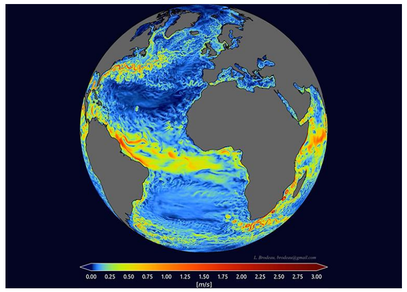
4. Mandatory: Gantt-Chart

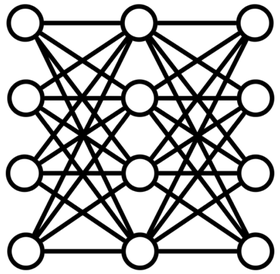
1. Project Overview

The project addresses the creation of a foundation model for modelling the world’s oceans

Project Title

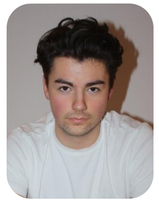
„Foundation model for modelling the world‘s oceans“





- **Thousands TB of open-source data collected from satellites and weather monitoring stations**
- **ESA alone third largest data provider in the world generating 20 TB/day**
- **20x more data than used to train modern AI lsystems ike ChatGPT**
- **Extracting useful insights requires application advancements in AI and BigData**
- **Focus on development of new algorithms beyond the state-of-the-art**

Research Team & Supervisors



Imar Colic

MSc Data Science

Can. Nr.: 50450

Bhavika Adapa

MSc Data Science

Can. Nr.: 50270

Prapti Pradhan

MSc Data Science

Can. Nr.: 51348

Shavya Tyagi

MSc Data Science

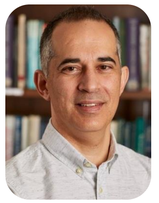
Can. Nr.: 45278

Prof. Piotr Fryzlewicz

Professor & Supervisor

Dr. Dima Karamshuk

Data Scientist & Industry Partner



Dr. Marcos Barreto

Capstone Coordinator & Supervisor

Agenda

1. Project Overview

2. Historical Context & Research

   Current Data sources, Ocean Characteristics, Modelling Approaches

   Current Challenges & Research Approach

   Further Insights Shaping Our Approach

3. Methodology & Plan

4. Mandatory: Gantt-Chart

2. Historical Context & Research Approach

Our research will build opon the relevant characteristics, models, and data

Ocean Characteristics

- **Secchi Depth (ZSD)**
- **Chlorophyll –a(CHL)**
- **Turbidity(TUR)**
- **Suspended Particulate Matter (SPM)**
- **Volume Attenution Coefficient (KD490)**
- **Dissolved Oxygen (DO)**
- **Sainity(SAL)**
- **Sea Surface Temperature (SST)**
- **Sea Wave Height (SWH)**

Data Sources so far

- **Copernicus Marine Service**
- **Sentinel 3 (ESA)**
- **EEA WISE-SoE Waterbase.**
- **In-Situ Buoy & Field Data**
- **ERA5 Reanalysis Data**
- **NOAA Wind & Wave Data**

Modelling Approaches

- **AR, ARIMA, SARIMA**
- **Random Forest (RF)**
- **Gradient Boosting Trees (GBT)**
- **Neural Networks (CNNs, RNNs, LSTMs)**
- **Transformer – Based AI (GraphCast, MetNet-3)**
- **Hybrid AI Physics Models**

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE Meta

Several challenges in the maritime industry form relevant research areas covered by our model

	1	2	3	4
Challenge	**Lack in Ocean Pollution control**	**No aligned Water Pollution Monitoring**	**Implementation of Water Framework Directive**	**No aligned ocean visibility prediction**
Research Area	Water Quality Monitoring: Leveraging Satellite Data for Agricultural Pollution Control	Using Remote Sensing to Improve Efficiency of Water Pollution Monitoring	Satellite-assisted monitoring of water quality to support the implementation of the Water Framework Directive	Ocean surface visibility prediction
Hypothesis	Enabling real time tracking of agricultural runoff to promote sustainable practices	Provides cost-effective, scalable solution for real time Water quality assesment	Supports compliance with global water quality standards and regulatory frameworks	Enhances maritime safety and operational efficiency in navigation and offshore activities

These are just some of the most recent topics in the area of modelling the world‘s oceans

We are using several insights of the current literature to create our foundational model

Most important isnights shaping our research

- **Integrating weather conditions and land characteristics improved the accuracy of machine learning models for water pollution detection.**
- **Satellite data (Sentinel-2, Copernicus) enable tracking and predicting phytoplankton blooms, detecting Harmful Algal Blooms (HABs) via chlorophyll-a estimations to assess eutrophication impact**
- **Missing satellite data, often caused by cloud cover or infrequent observations, has been managed using Grouped Mean Substitution (filling with the mean of available values) and Negative/Zero Substitution (-10 as a placeholder or 0 as neutral)**
- **Spatiotemporal deep learning models like SimVP can be more effective than traditional statistical methods (ARIMA, SES) for ocean visibility prediction by capturing both spatial and temporal dependencies, potentially reducing RMSE and improving inference speed significantly.**

Agenda

1. Project Overview

2. Project Context

3. Methodology & Plan

   - Datasets
   - Candidate Models
   - Other important aspects

4. Mandatroy: Gantt-Chart

These are the datasets we want to use to develop the foundation model for modelling the world's oceans

1. Satellite Data

Primary source of data, providing most general and accurate data for ocean monitoring

Sentinel Satellites (ESA – Copernicus Programme)

- **Sentinel-1**: Synthetic Aperture Radar (SAR) for all-weather, day/night ocean monitoring, useful for wave height and oil spill detection.
- **Sentinel-3**: Most relevant for Sea Surface Temperature, Ocean Color, Sea Height
- **Sentinel-6 (Michael Freilich)**: focused on precise sea level rise measurements

NASA Satellites

- **MODIS (Aqua & Terra)**: Measures ocean color, SST, and biological productivity
- **Suomi NPP & NOAA-20 (JPSS Program)**: Atmospheric and oceanic data
- **Surface Water and Ocean Topography**: Ocean currents and surface water bodies
- **Landsat-8 & 9 (Joint NASA-USGS)**: Multispectral coastal and near shore water data

Other international statellites

- **Himawari-8 (Japan - JAXA)**: Geostationary satellite providing SST measurements
- **GOES-16/GOES-17 (NOAA - US)**: Monitors SST, ocean storms, and atmospheric
- **CryoSat-2 (ESA)**: Measures sea ice thickness, useful for Arctic and Antarctic

2. Non-Satellite Data

Aimming to diversify with air, in-water, etc.

NASA Air-based Platforms

- **CASEI**: Catalog of Archived Suborbital Earth Science Investigations, airborne and field data

NASA Water-based Platforms

- Data from ships, boats, and buoys advance our knowledge of how changes in Earth's systems affect water quality, water supply, and sea level rise.

Foundation Model

Output

Pedict target variable for specific time

Potential variables: Chlorophyll (Plankton Bloom), Sea Surface Temperatures, Turbidity

Pollution Management

- Providing polluting industrial companies an overview of their pollution
- Regulation can make use of pollution monitoring and increasing compliance for polluting companies

Fishery Forecasting

- Using set of relevant variables to predict fish population in the future
- Supporting fishery companies for sustainable and predictable fishery

We work alongside current research to optimize our approach to build our foundational model

Foundation Model

Personalized Adapter for Large Meteorology Model on Devices (PLM) (Chen et. al, 2024)

- **NLP-Based Weather Forecasting:** Uses NLP architectures to model weather data as sequential measurements (e.g., temperature, humidity, wind speed) with dependencies across time and space.
- **Pretrained Language Model (PLM) for Weather:** Trained on a large corpus of multivariate weather variables to predict the next event in a sequence, enabling it to handle sequential weather patterns beyond seen data.
- **Input-Output Representation*:*** modeling dependencies for improved forecasting accuracy. *Maps [Temperature, Humidity, Wind Speed] → [Future Weather Variables]*
- **Adapters for Weather-Specific Knowledge:** Lightweight adapters fine-tune the PLM for weather forecasting, making it aware of trends, seasonality, and residuals using domain-specific adjustments.
- **Weather Data Decomposition:** Splits input into trend, seasonal, and residual components, with each processed separately via task-specific adapters to capture unique behaviors.
- **Long-Term Pattern Recognition:** Excels at identifying long-term trends, seasonal variations, and anomalies in weather sequences, such as how temperature affects wind speed.

Key Takeaways

1. Consider NLP / transformer-based architecture
2. Ensure availability of a wider dataset for different set of spatial & temporal variables that extend beyond ocean characteristics
3. Decompose weather data into trend, seasonal, residual components and train the model on these individually

Extending our data volume and width will help us to approach the generalization of our foundational model

Foundation Model

Scaling transformer neural networks for skilful and reliable medium-range weather forecasting (Nyugen et. al, 2023)

- **Transformer-Based Weather Prediction:** Stormer is a transformer model optimized for weather forecasting, using a randomized forecasting objective to predict weather dynamics over different time intervals.
- **Weather-Specific Embeddings:** Converts multi-dimensional spatial grids into tokens, capturing climate variable interactions (e.g., pressure-temperature effects) and spatial dependencies (e.g., wind patterns).
- **Multi-Scale Training:** Trained on 6, 12, and 24-hour intervals, allowing the model to learn both short-term high-frequency patterns and long-term low-frequency trends for more accurate predictions.
- **Ensemble Forecasting Approach:** Generates multiple forecasts iteratively over different time intervals and combines them via averaging or weighted aggregation, reducing prediction errors and improving robustness.
- **Pressure-Weighted Loss Function:** Assigns higher penalties for errors in pressure-sensitive regions (e.g., cyclone centers), ensuring the model prioritizes accurate forecasting in critical weather conditions.

Key Takeaways

1. Weather-specific embeddings to process this data
2. Train with different time intervals to capture short & long-term trends
3. Generate multiple forecasts to create a combined forecast that minimises prediction error

Agenda

1. Project Overview

2. Project Context

3. Methodology & Plan

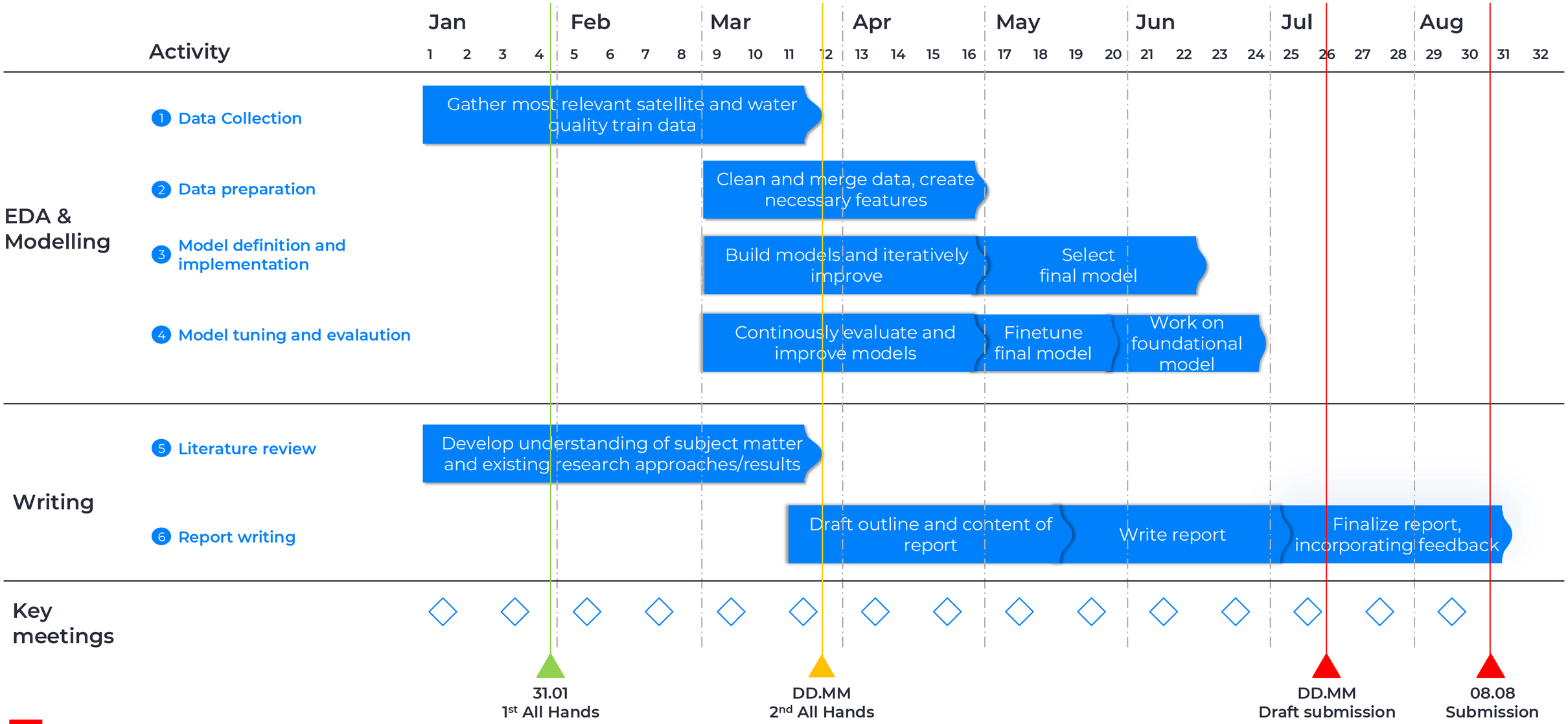
4. Mandatory: Gantt-Chart

Gantt-Chart (Nov 2024 - Aug 2025)

List of tasks

4. Gantt-Chart

The next step in our project is to gather more data and structure it to be able to train our model



4. Gantt-Chart

We split our tasks into pairs, however we sync always on our most critical tasks to work together

List of tasks

Task No.	Tasks	Deadline	Project Member
1	Literature review: Develop understanding of subject matter	January	Shavya, Prapti
2	Gather relevant satellite and water quality train data	January	Imar, Bhavika
3	Clean and merge data, create necessary features	February	Shavya, Bhavika
4	Build models and iteratively improve	March - April	Imar, Prapti
5	Continuously evaluate and improve models	April - May	All Members
6	Select final model	May	Shavya, Imar
7	Draft report outline and content	May	Imar, Bhavika
8	Fine-tune the final model	June	Shavya, Prapti
9	Work on creating the foundational model	June	All members
10	Finalize report and incorporate feedback	July - August	All members

Q&A

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE & Meta

ST498 Group 17: Modelling the World‘s Oceans