Travis Dawson

THE UNIVERSITY OF WESTERN AUSTRALIA

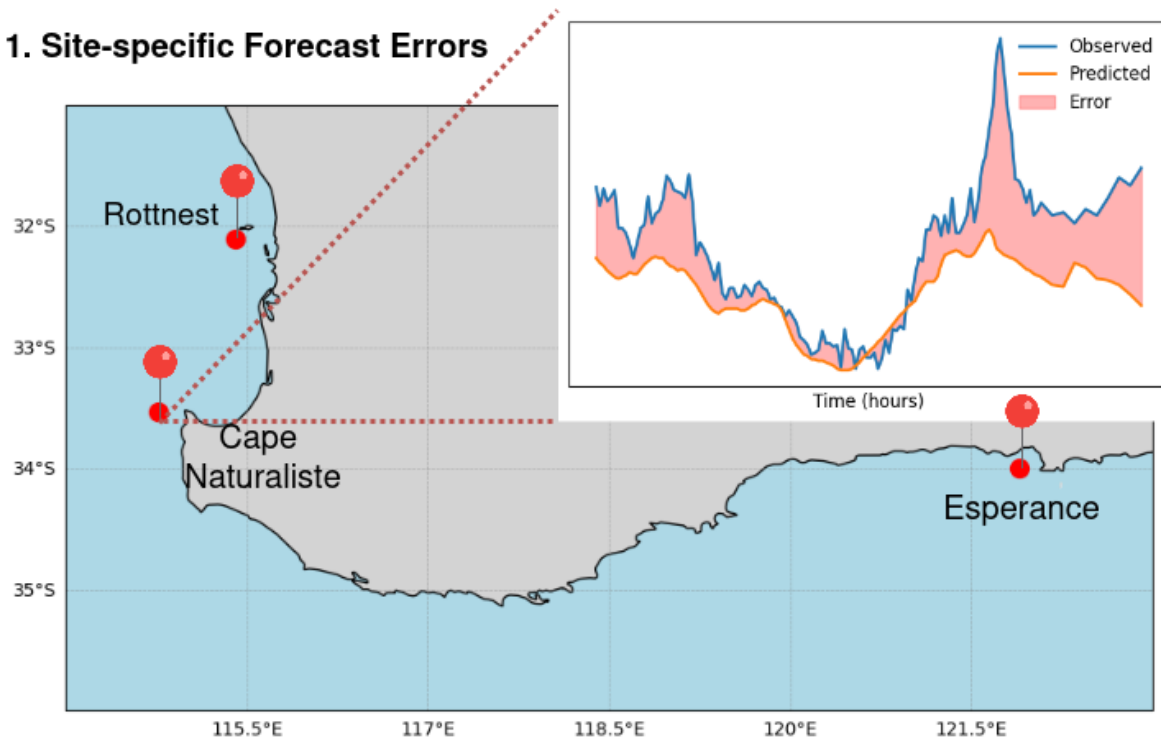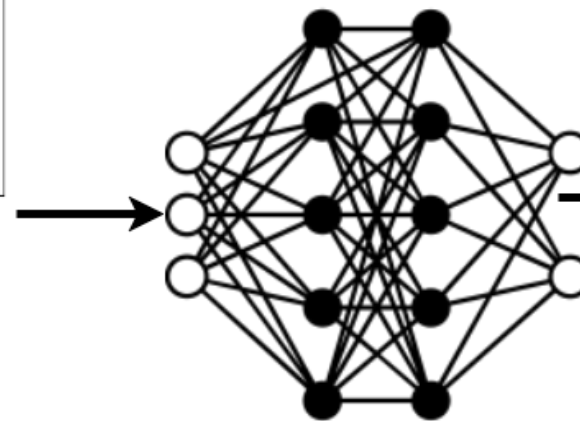## A Machine Learning Framework to Improve Site-specific Ocean Wave Forecasts
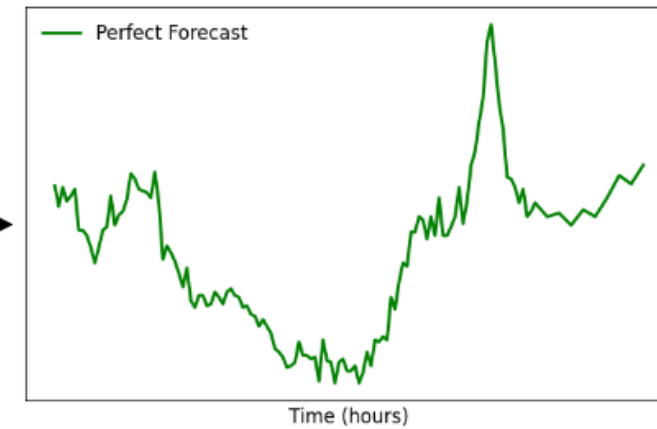
Supervisor Team: Jeff Hansen and Arthur Filoche
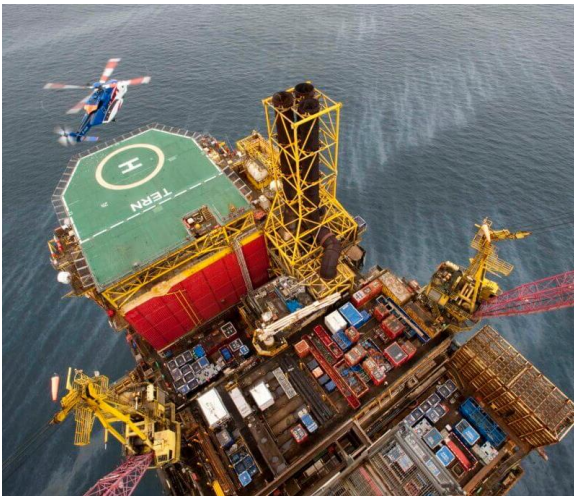
Why this is important?

1. Uncertainty in Offshore Oil and Gas Operations

Operational disruptions:
- Delays, downtime, missed windows, conservative risk tolerances

Potential impacts:
- Increasing operational costs
- Loss of revenue

2. Uncertainty in Offshore Wind

Operational disruptions:
- Missed maintenance windows, turbine downtime, etc.

Potential impacts:
- Increased maintenance costs
- Missed energy production targets

The Dataset?

NOAA Global Forecast System (GFS) for Ocean Waves

Dataset Info:
- Spatial Extent: Global
- Resolution 0.25 deg
- Horizon: 15 days
- Variables: 19
- Time period: 2021/03/23 - current
- Source: Publicly available on AWS Bucket

GFS Significant Wave Height 0.25 Resolution
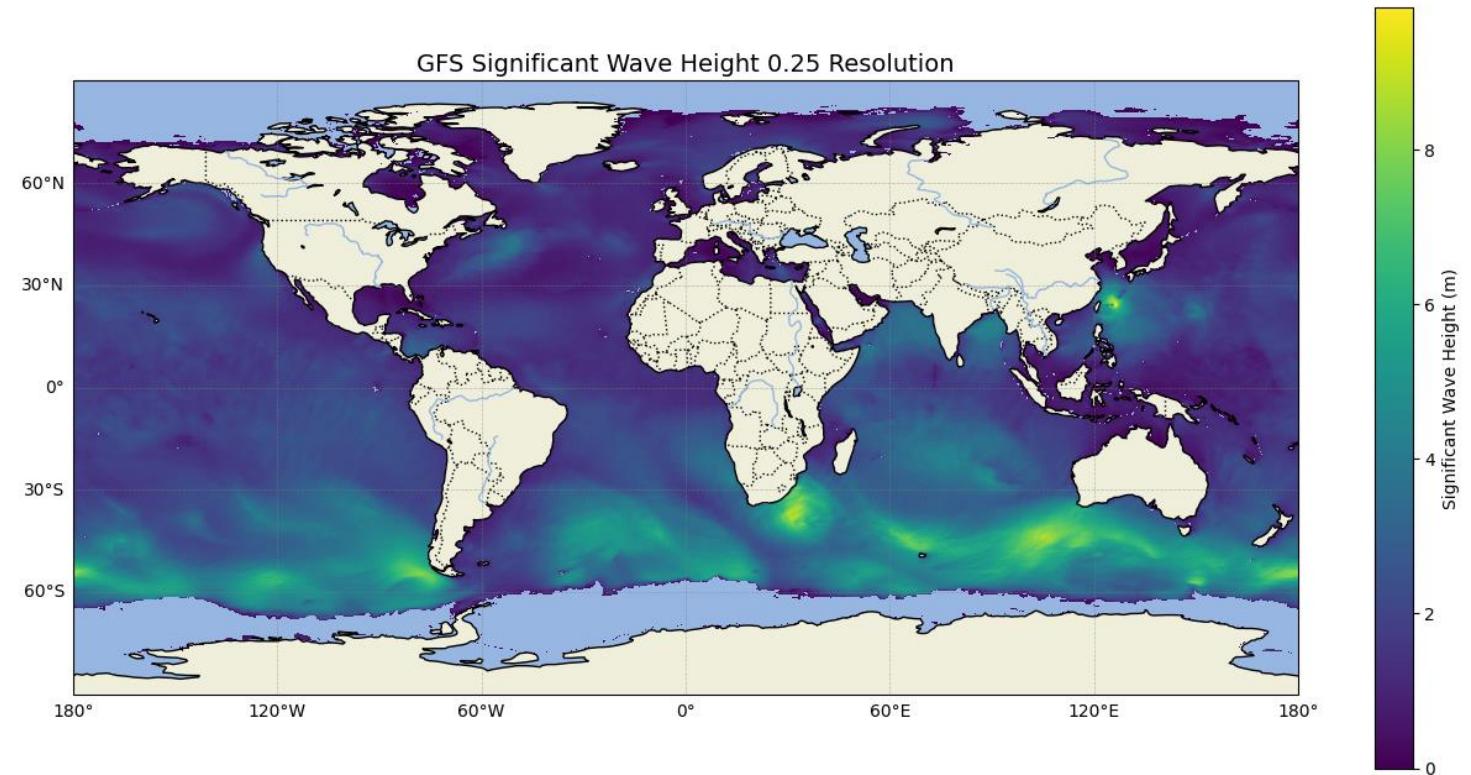
The Problem?

Scope:
- A single daily forecast has 161 grib2 files [0-240 hours horizon] ~ 2.0 GB (compressed)
  - Individual file size 12.3 MB (small)
- 5 years of data ~ 4 TB dataset (805 825 small files)

Identified Challenges:
- Grib2 files are not friendly to Machine Learning or HPC
- Small files present an I/O bottleneck
- I will experiment over selections of variables (I don't need the full dataset every time)

GFS Significant Wave Height 0.25 Resolution
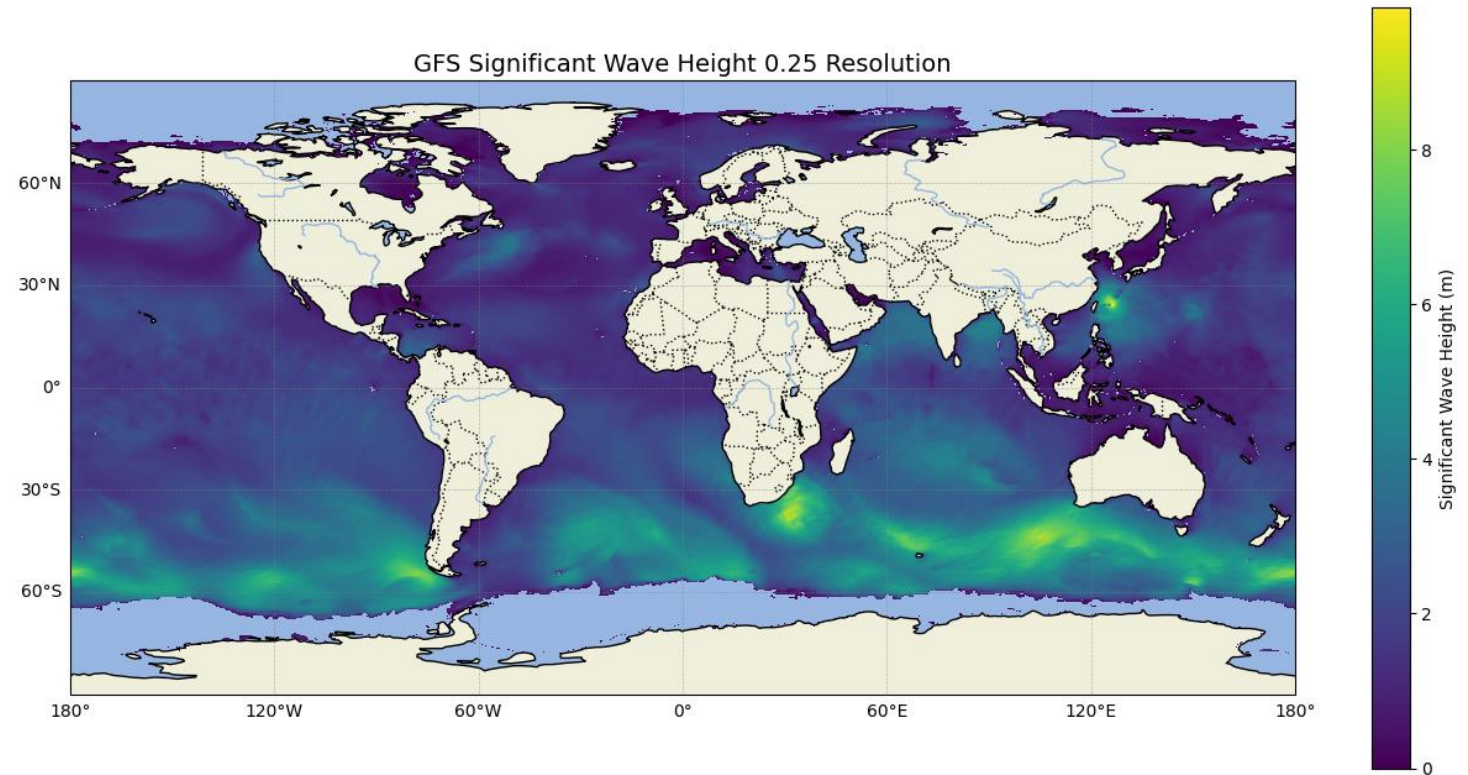
The Idea
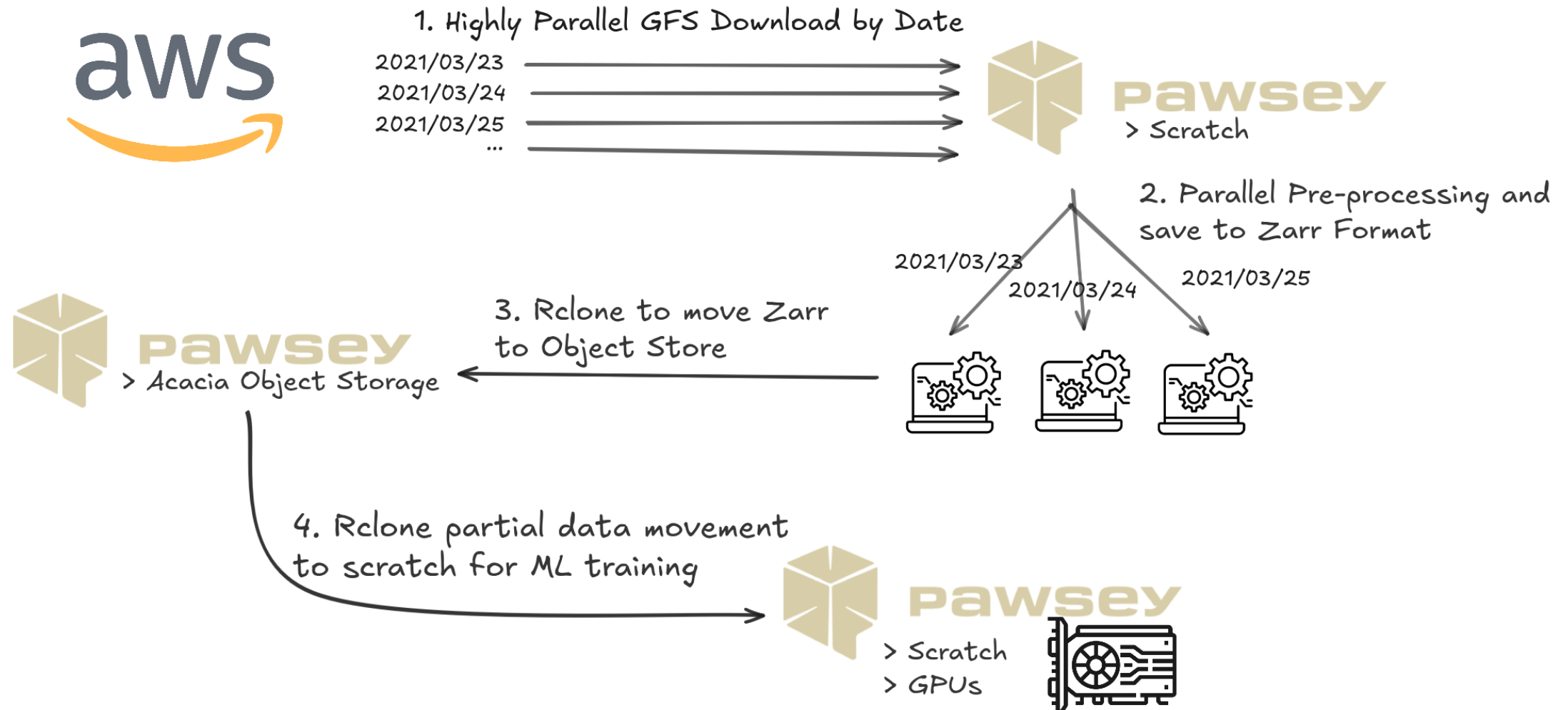
Zarr Format?

Grib2 File Context:
- Grib2 is the WMO standard for meteorological data
- Highly compact with mature tooling

Zarr File Context:
- Chunking enabling parallel on HPC
- Partial data movement through chunks
- ML friendly

How it helped:
161 files -> 91 files (44% file count reduction)
Each file is ~30 MB (3x larger)
Select variable access example:
- I need a single variable at 5 days horizon for ML training
- Requires 15 files per forecast compared to 161 (91% reduction)
- In data storage this is 30 MB vs ~2.0 GB

This still needs to be tested and experimented over.