Field campaign data in observation-driven AI weather prediction:

Opportunities and challenges

Peter Lean on behalf of the AI-DOP team:

Mihai Alexe, Ewan Pinnington, Eulalie Boucher, Patrick Laloyaux, Simon Lang, Tomas Kral, Tony McNally and many colleagues across ECMWF

Context: Multiple research paths at ECMWF

Development of operational **Analysis-driven ML** ensemble

incl. atmosphere, ocean, land, sea ice components

Research into

Foundation Models of the Earth System

(WeatherGenerator)

Hybrid Physics / ML

Augmenting the existing IFS with ML

Longer term research exploring

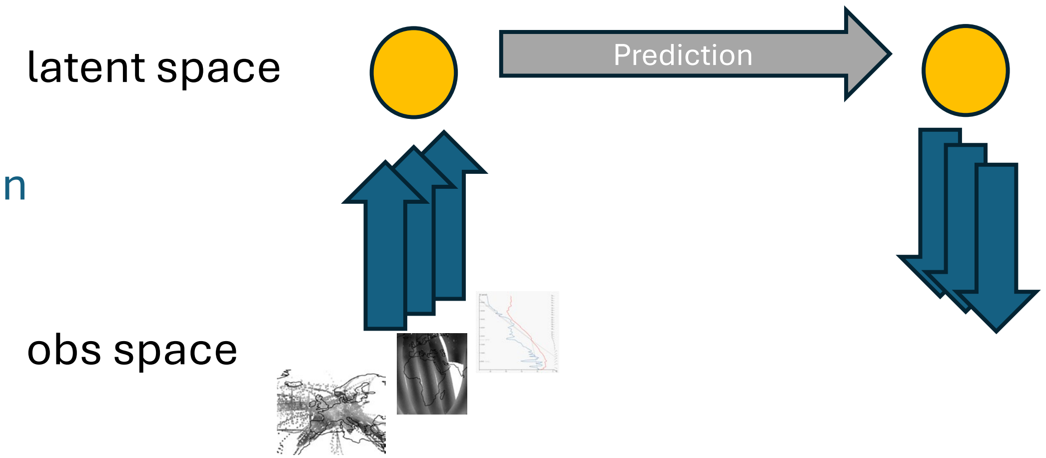
End-to end observation-driven forecasting systems

ECMWF is exploring observation-driven ML forecasting models: AI-DOP

End-to-end observations → forecast

A **learned forecast model** which evolves the latent state forward in time

A **learned "DA" function** which combines information from different sources into a (learned) set of unified latent state variables

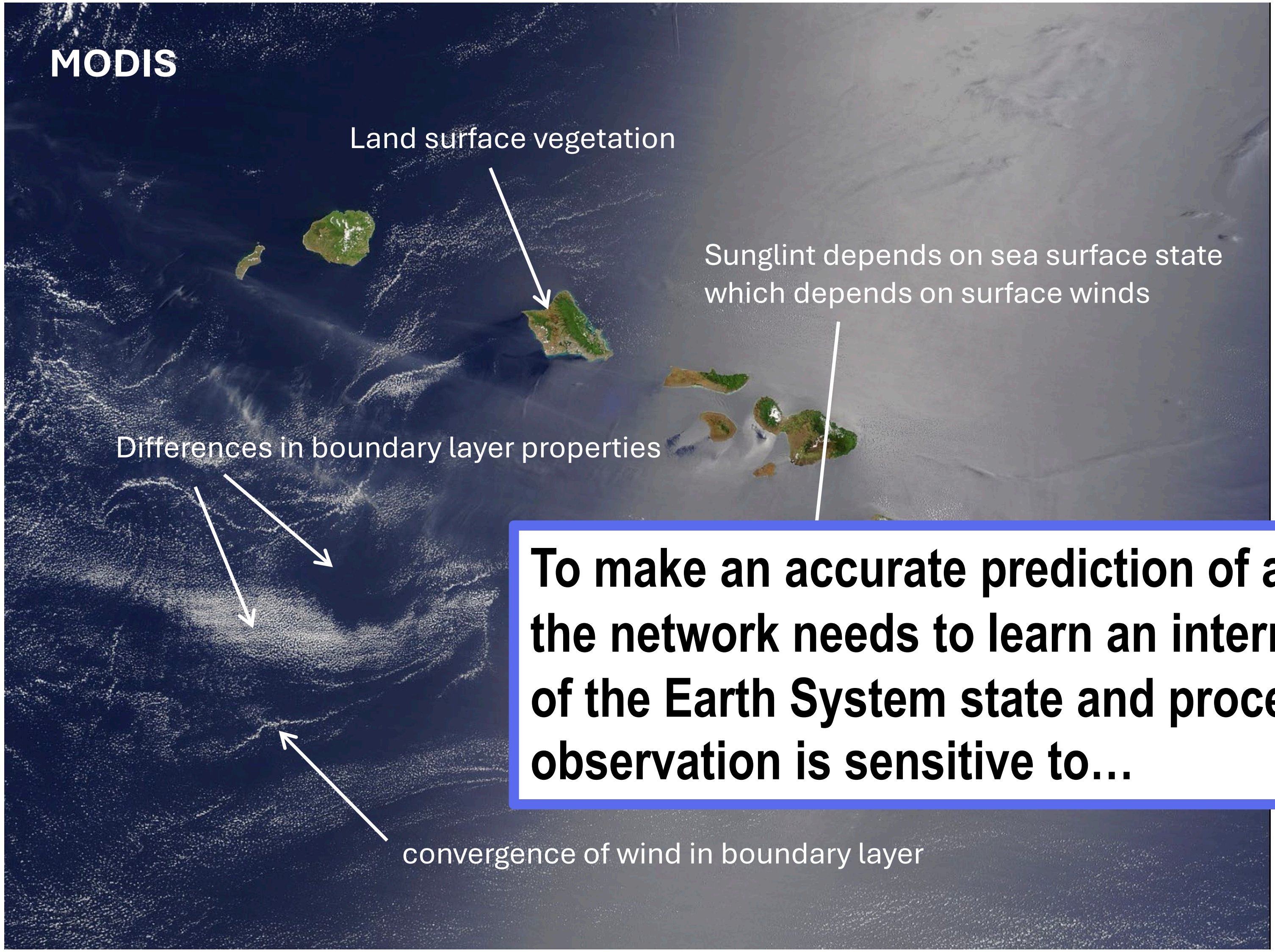


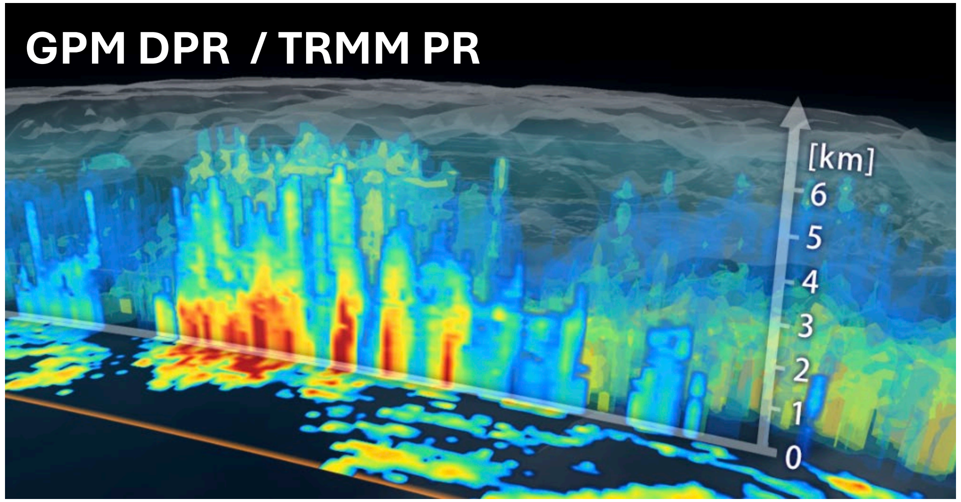
Learned observation operators which map from the network's internal model latent state to observation space

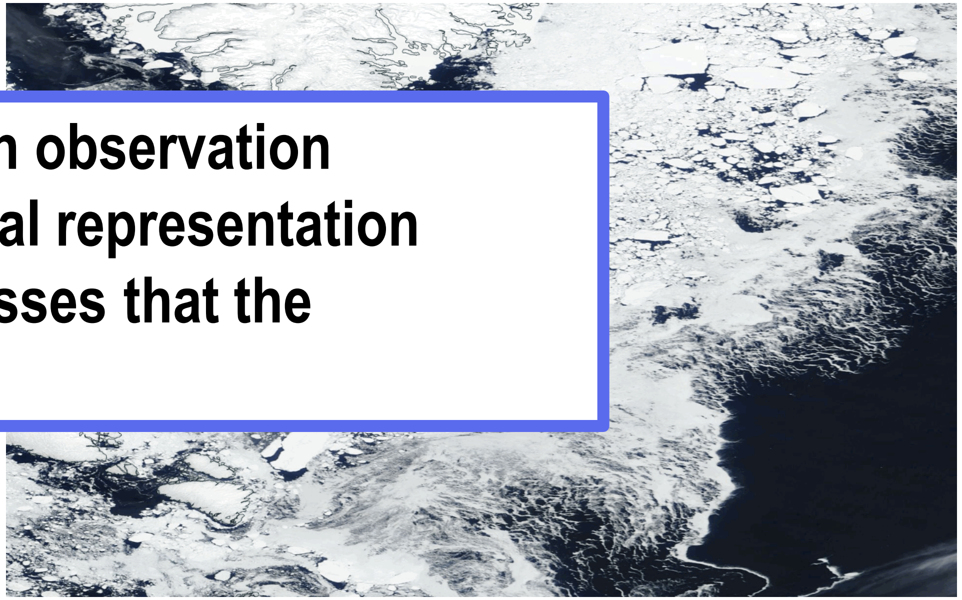
Given observations in the past, predict observations in the future

Learning a model from observational data

Learning from the rich information content of observational data



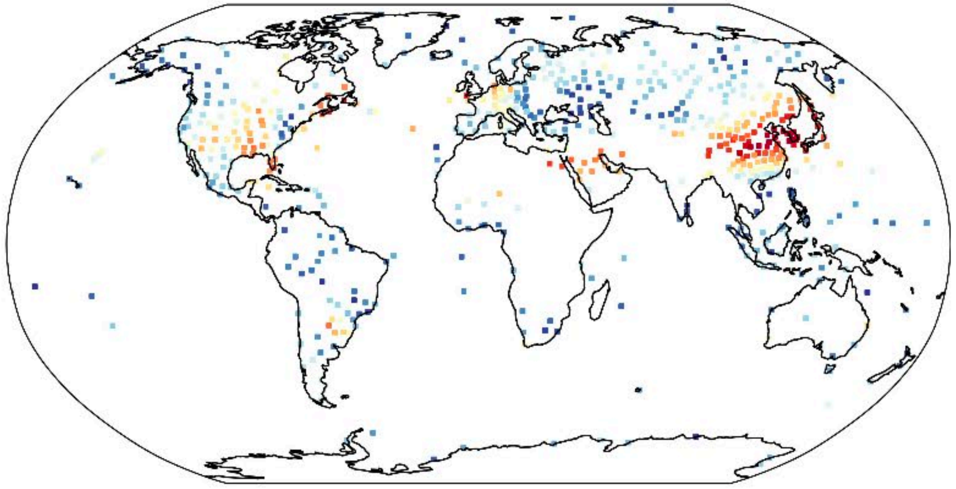




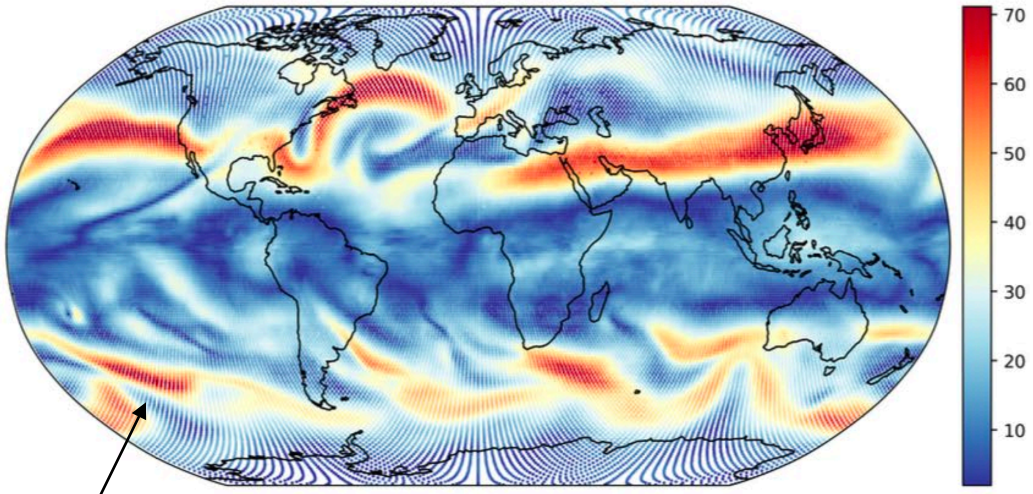
To make an accurate prediction of an observation the network needs to learn an internal representation of the Earth System state and processes that the observation is sensitive to…

Learning a data assimilation function to combine information from multiple sources: GraphDOP can make predictions on a regular grid

Input 200hPa wind obs



Predicted 200hPa wind at t+24h on O96 grid

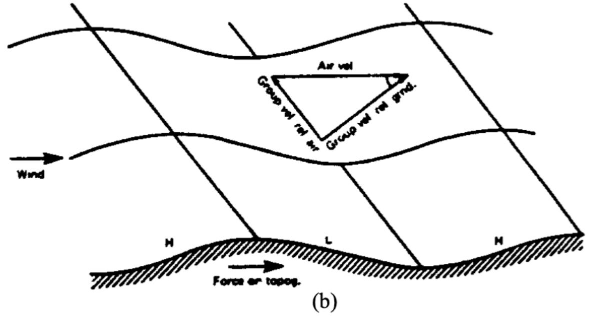


Even in areas where there are no direct observations for that variable

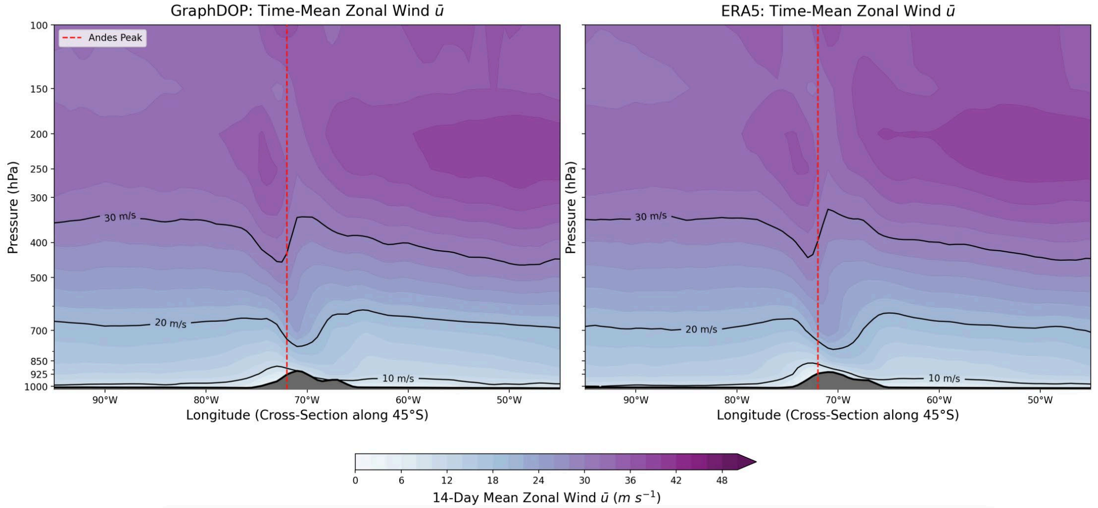
- evidence that it has developed a model of the relationship between different observed variables

Learning a model of the Earth System from sparse, indirect observations

Flow over orography



GraphDOP | ERA5



Ageostrophic flow

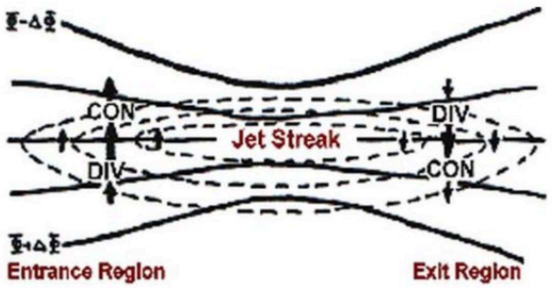
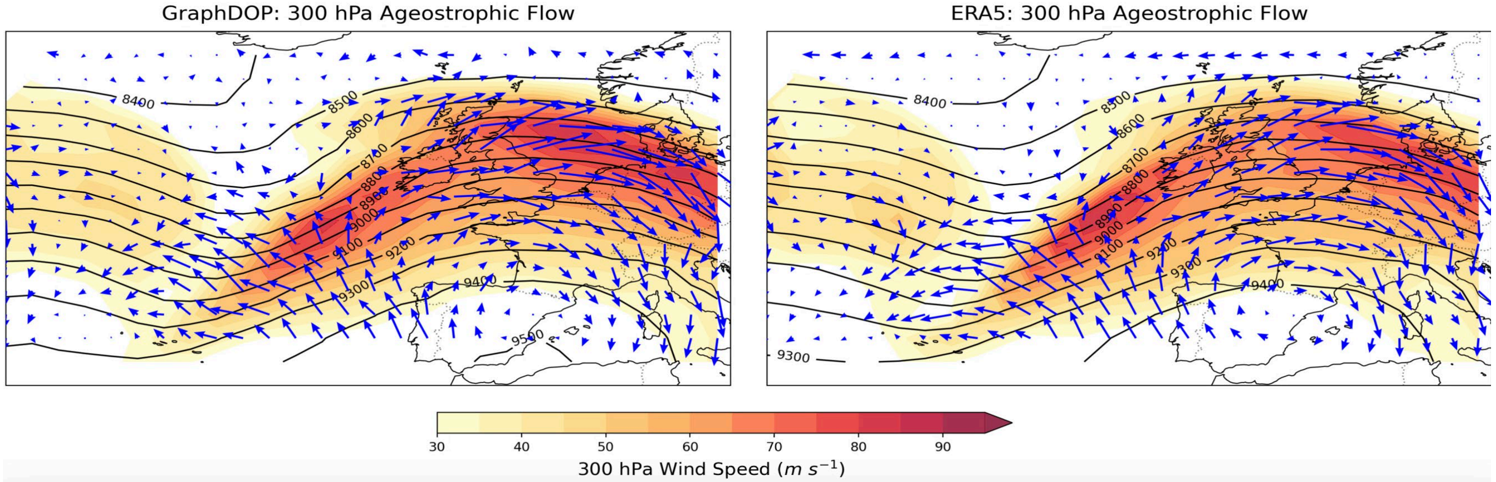


Figure 2: Location of convergence (CON) and divergence (DIV) areas in the entrance and exit region of a jet streak, © NOAA

GraphDOP | ERA5



ECMWF

Learning a model of the Earth System from sparse, indirect observations

Semi-diurnal solar atmospheric tides

$$S_2^* = \frac{(p_0 + p_{12}) - (p_6 + p_{18})}{4}$$

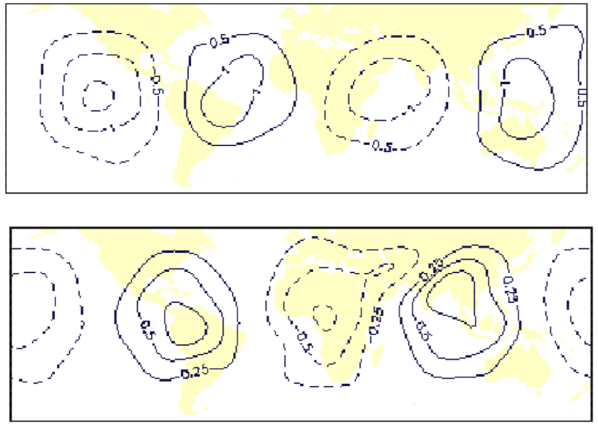
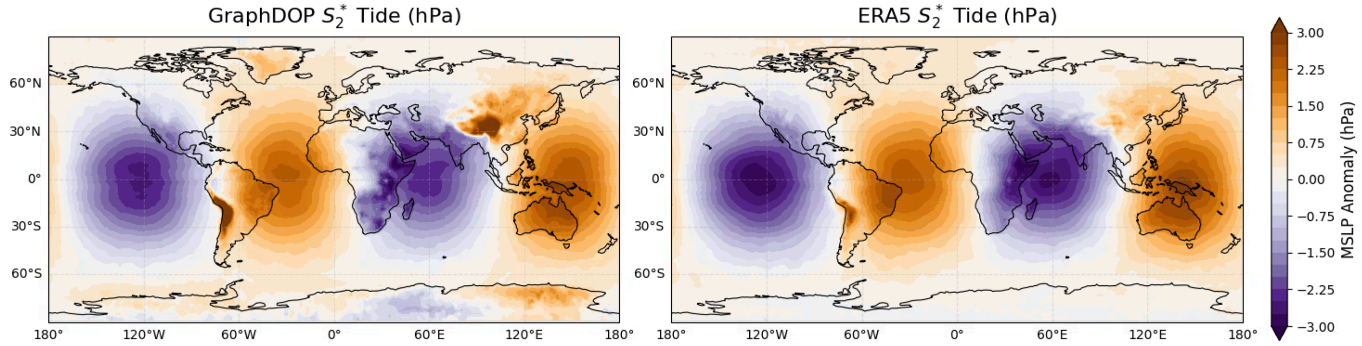


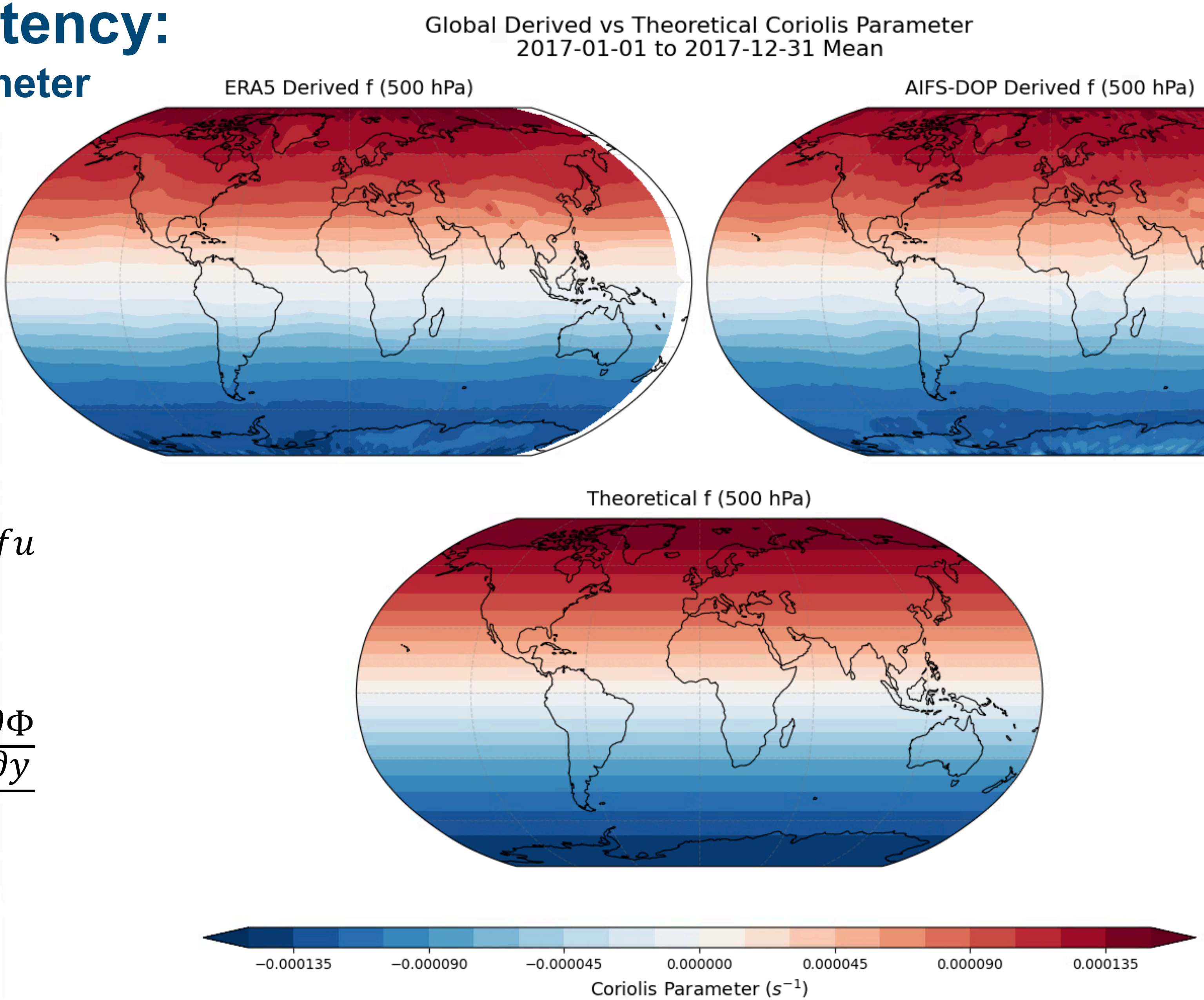
Figure 5. The semi-diurnal tide. The upper plot shows surface pressure (hPa) and the lower plot the tendency of pressure (hPa/hour) at the lowest model level. The quantity plotted is $(1/4)(A_{00} - A_{06} + A_{12} - A_{18})$ where A_{xx} denotes the mean ECMWF analysis at xx UTC for January 1997, truncated spectrally at T10 to remove local orographic and station-specific features. Solid contours denote positive values and dashed contours negative values. The nature of atmospheric tides is described in Chapman and Lindzen (1970).



GraphDOP

ERA5

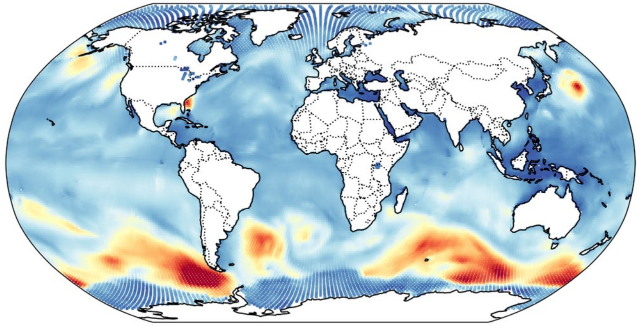
Physical consistency:
derived Coriolis parameter



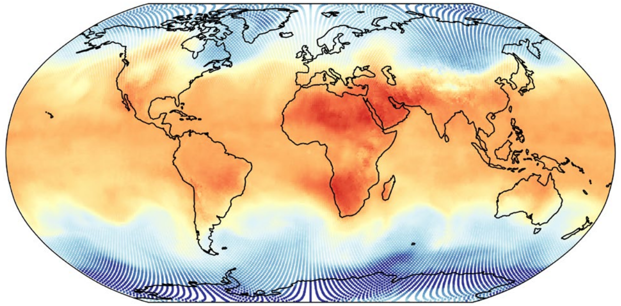
$$\frac{\partial \Phi}{\partial x} = fv \qquad \frac{\partial \Phi}{\partial y} = -fu$$

$$f_{derived} = \frac{v\frac{\partial \Phi}{\partial x} - u\frac{\partial \Phi}{\partial y}}{u^2 + v^2}$$

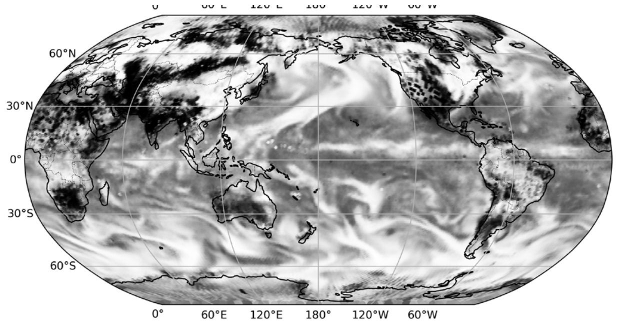
Learning observation operators that map back to observed variables



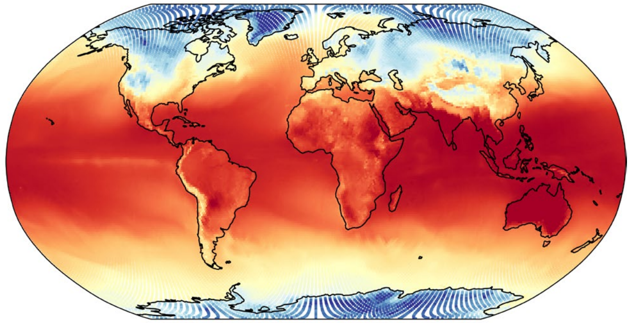
Significant wave height



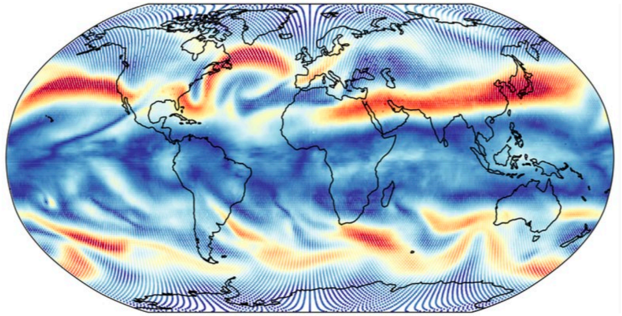
850hPa temperature



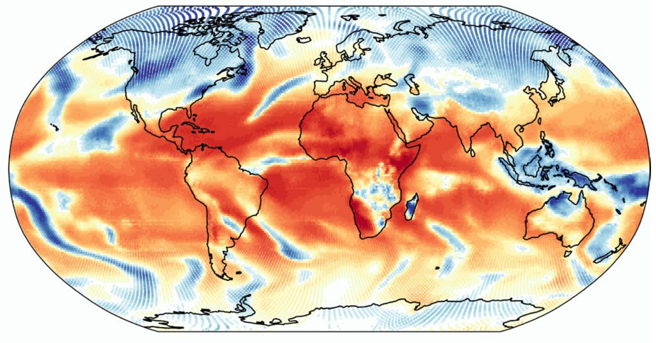
Cloud fraction



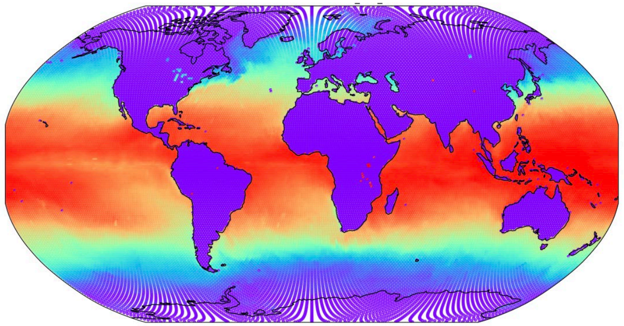
2-meter temperature



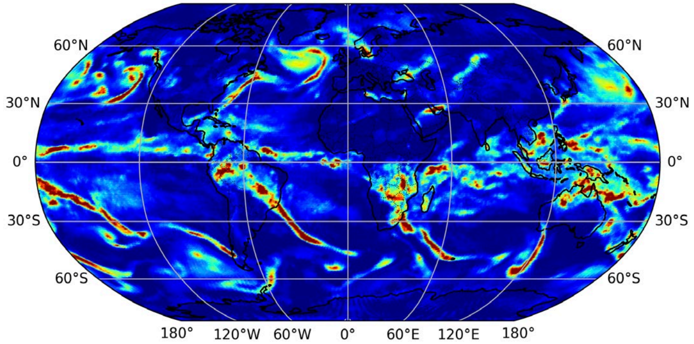
200hPa winds



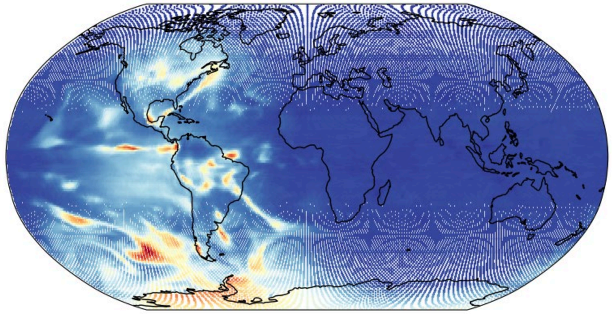
SEVIRI infrared window channel



SST

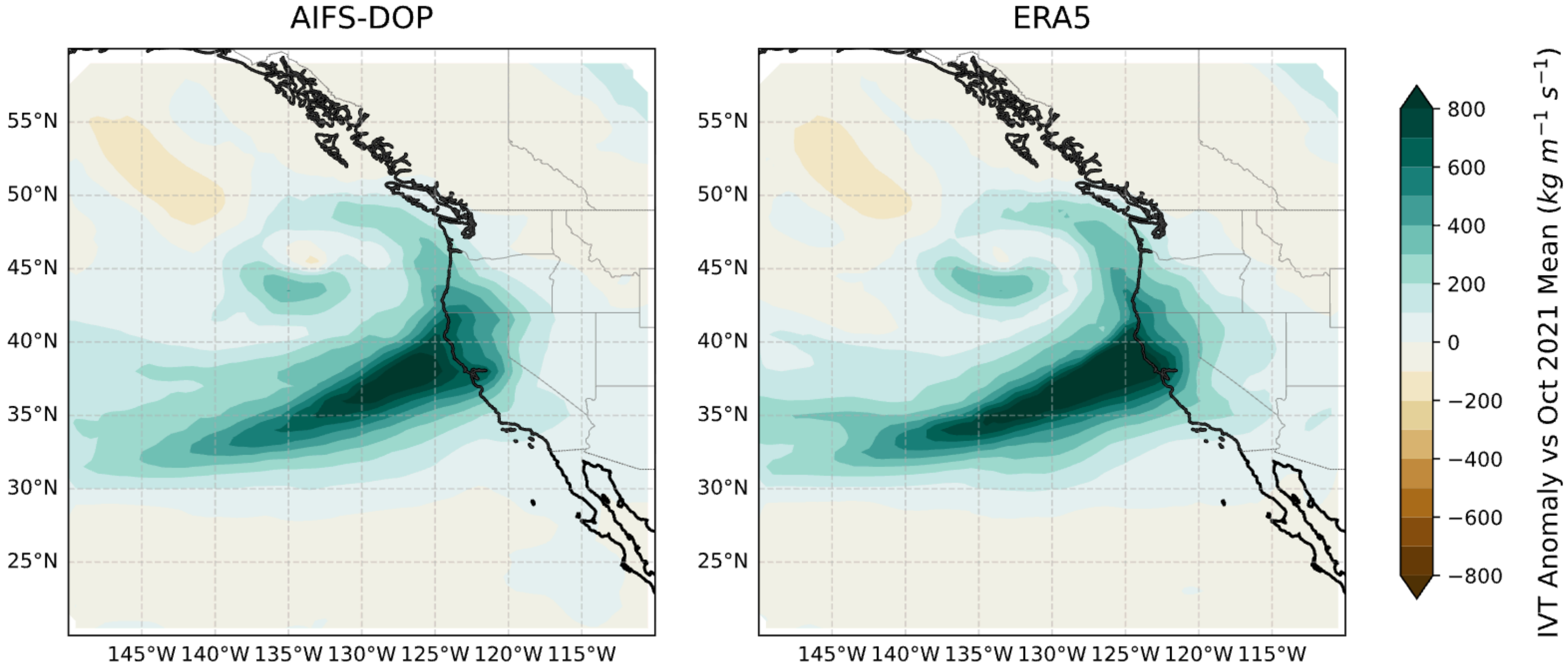


1 hour precipitation accumulation



AVHRR visible channel

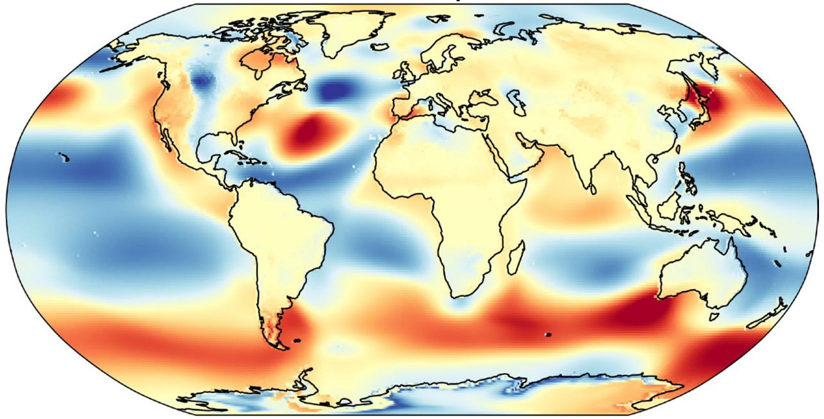
Case studies: atmospheric river event



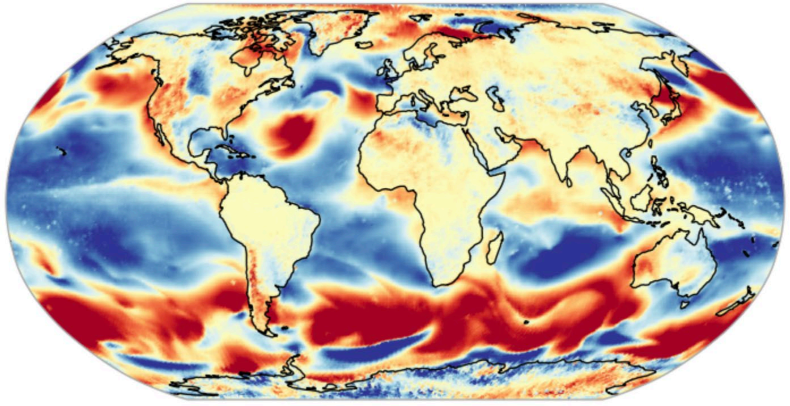
Oct 24th, 2021
12 UTC

AI-DOP: Progress during 2025

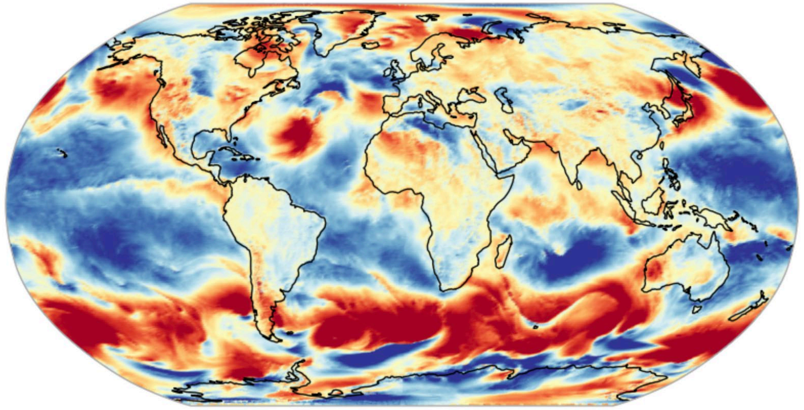
10m wind speed



Results from Sept 2024

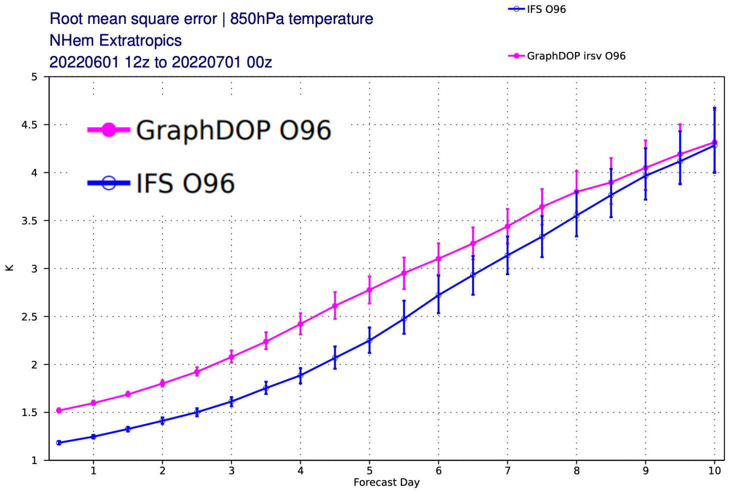


Sept 2025

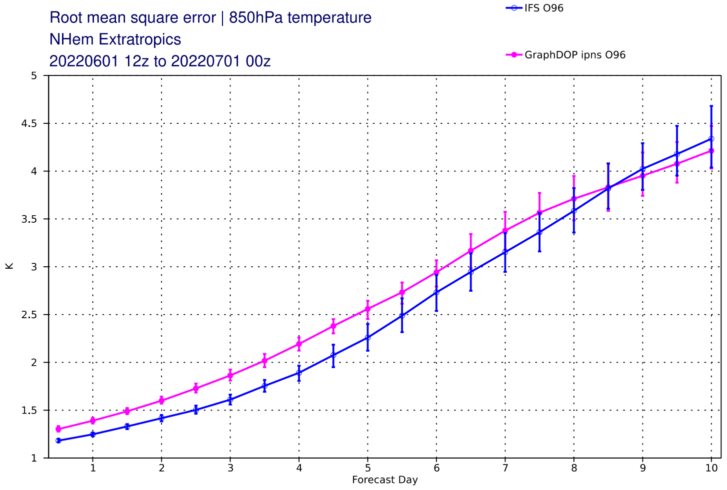


IFS analysis

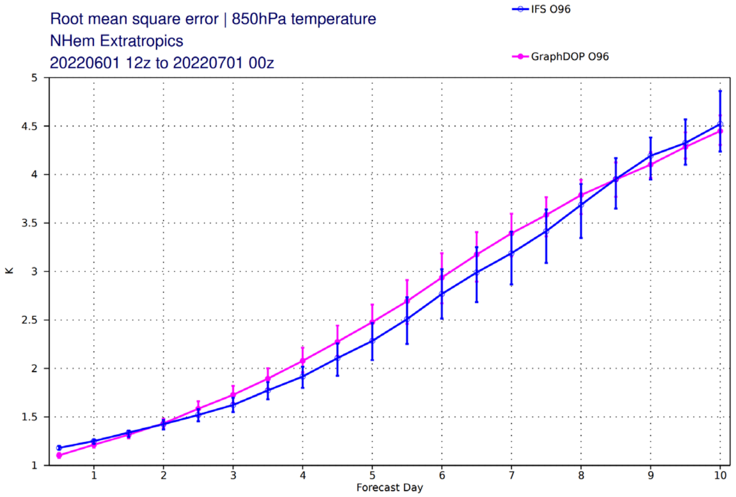
T850 N.Hem RMSE:



May 2025



June 2025



August 2025

ECMWF

Results

850hPa Temp

Results from
1-year evaluation

500hPa geopotential

Results from
1-year evaluation

250hPa wind

Results from
1-year evaluation

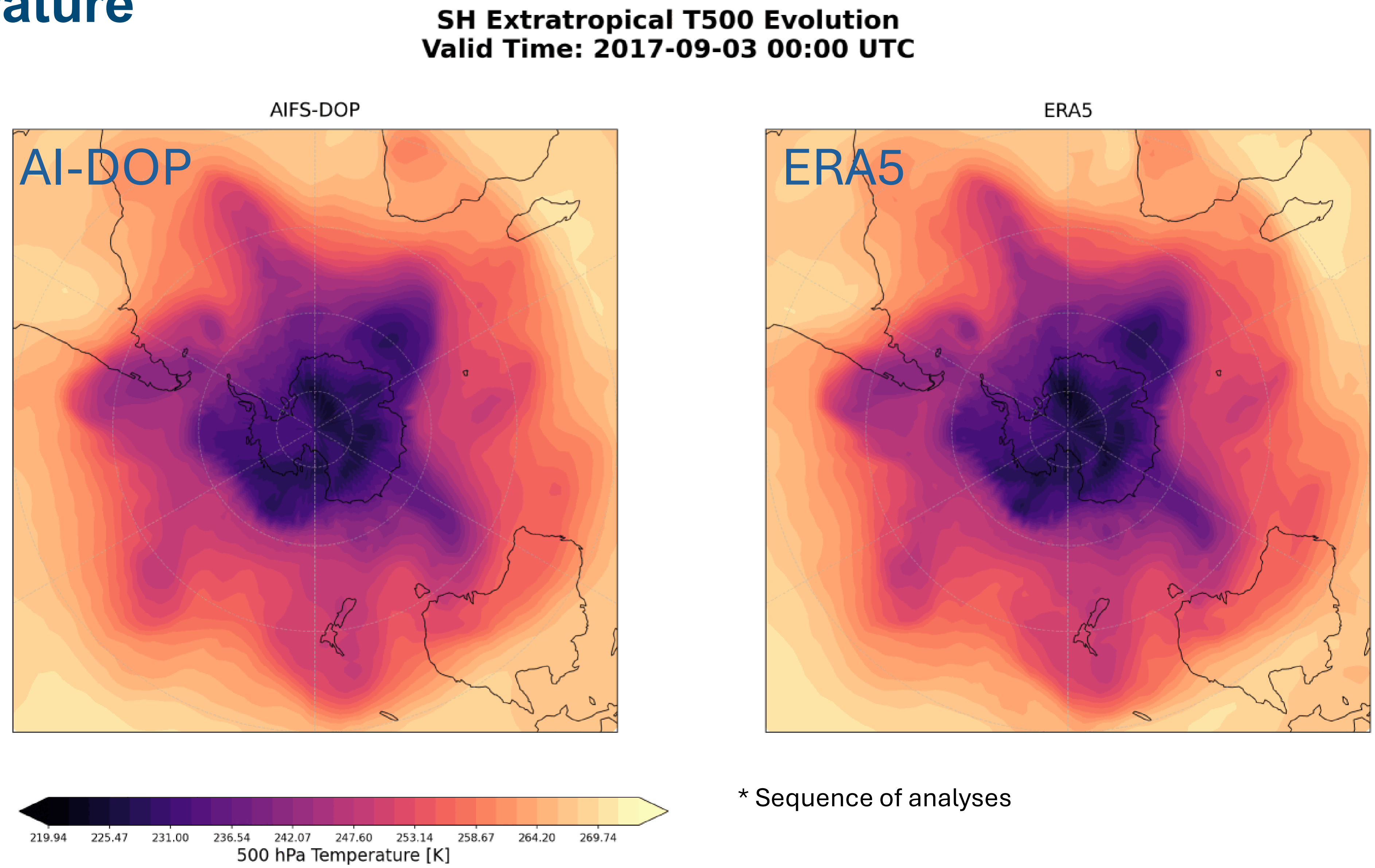
2m temperature

Results from
1-year evaluation

10m wind

Results from
1-year evaluation

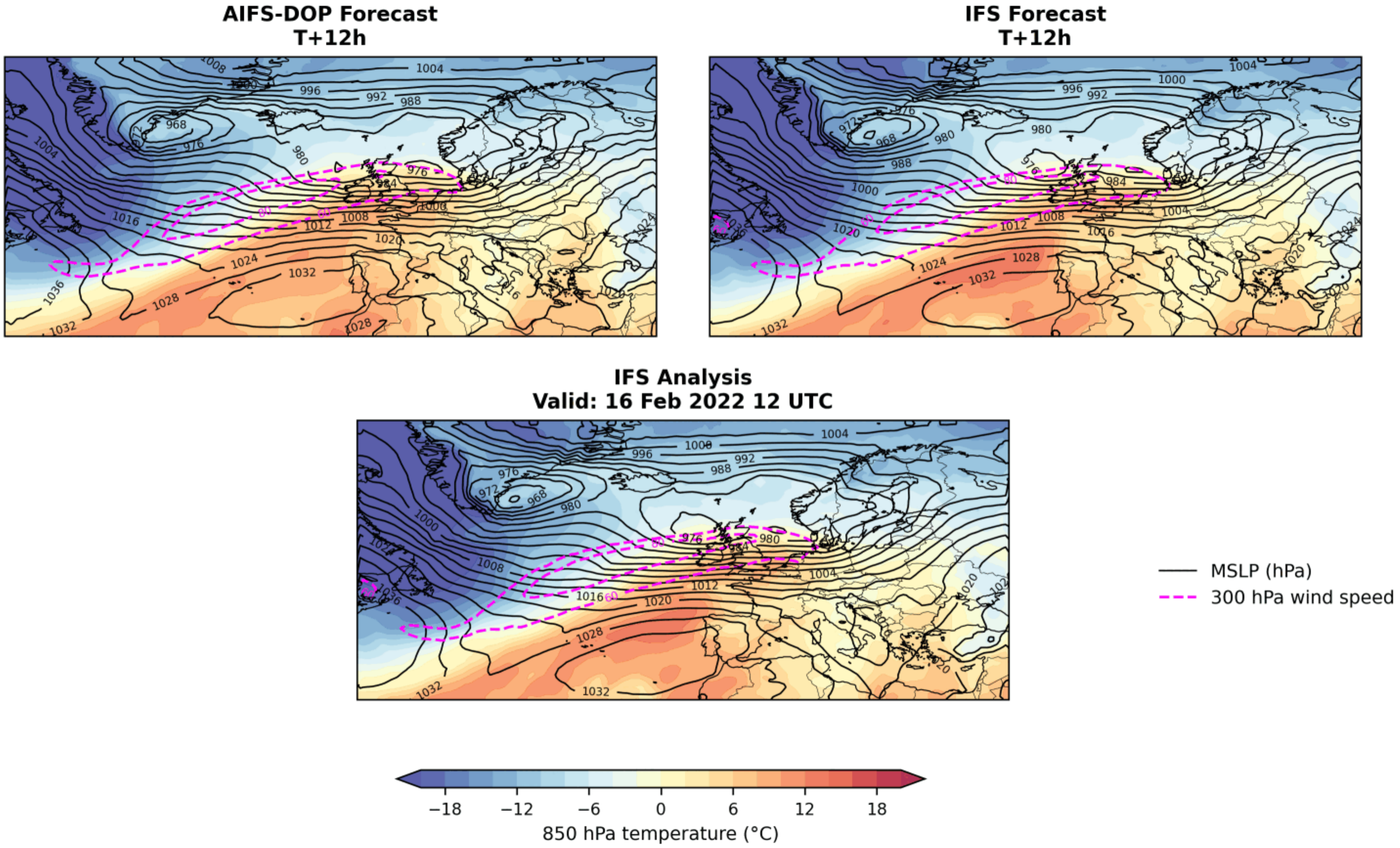
500hPa Temperature



* Sequence of analyses

Storm Eunice, February 2022

Case study: Storm Eunice



Implications for field campaign data

- For observation-driven ML models, **observational data is everything**
- So far, we have been training on data that is available in real-time
- NWP model development has been informed by field campaign data that AI-DOP has never seen
- **Q. Can we add field campaign data into the training dataset to help it improve its model of processes that are not well captured by the traditional observing system?**
- **Q. Can the network learn relationships which generalize globally from spatially and temporally limited field campaign datasets?**

If observation-driven ML takes off could it herald
a new **Golden Age** for field campaigns?

Summary

- ECMWF is exploring end-to-end machine learning weather prediction systems alongside physics-based and reanalysis-driven ML systems
- Encouraging results with medium-range forecast scores continuing to improve
- Signs of physical consistency of the predictions produced
- Observation-driven ML models learn a model from observational training data
- Plan to start including field campaign datasets in the training dataset
- Open questions regarding ability of the network to learn generalized model improvements from spatially and temporally restricted campaign datasets