Learning Weather Forecasts from Observations

Ewan Pinnington and DOP Team

ECMWF

ewan.pinnington@ecmwf.int

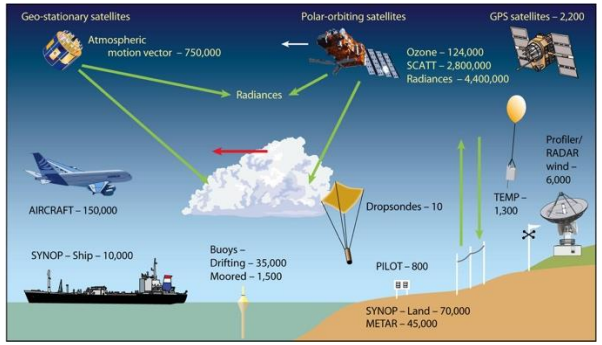


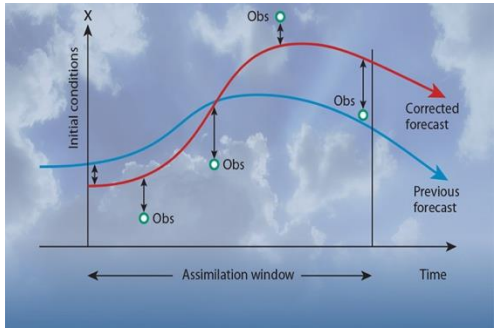
Overview

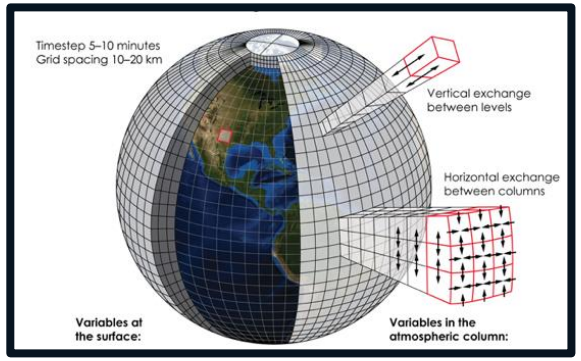
- Traditional Weather Forecasting Workflow:
  - How are Weather Forecasts made with the traditional system
- Artificial Intelligence Forecasting System (AIFS):
  - Where are we starting to use Machine Learning for weather forecasting?
- Direct Observation Prediction (AI-DOP):
  - Making Weather Forecasts directly from observations using ML
- Summary

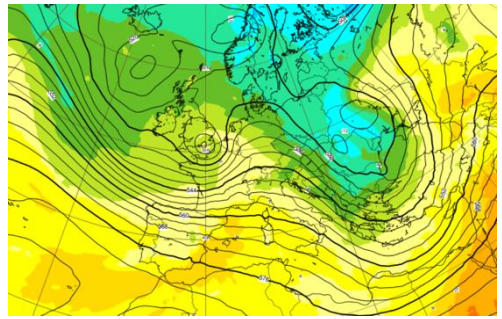


How do we make Weather Forecasts





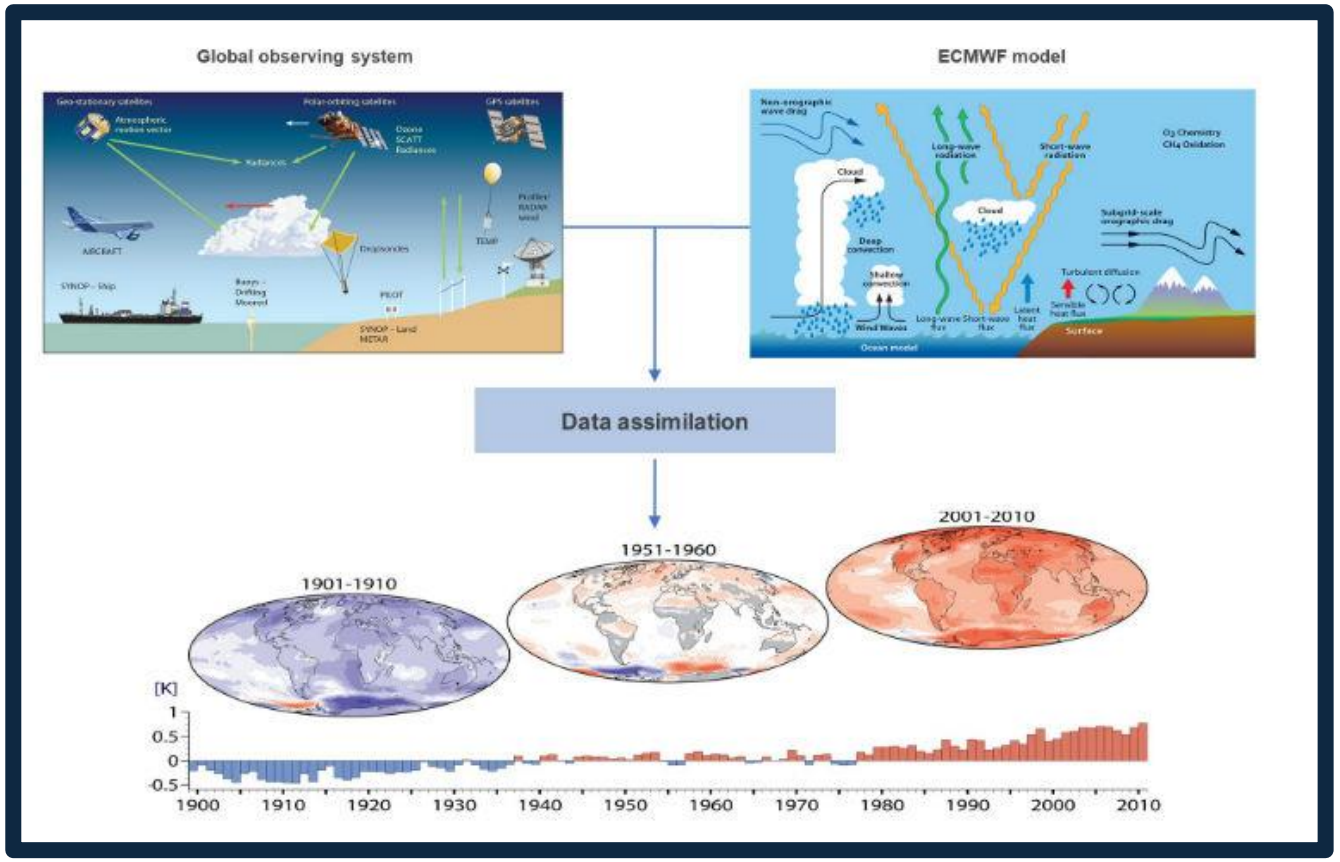




Observations → Data assimilation → Physics based forecast model → Post-processing and dissemination

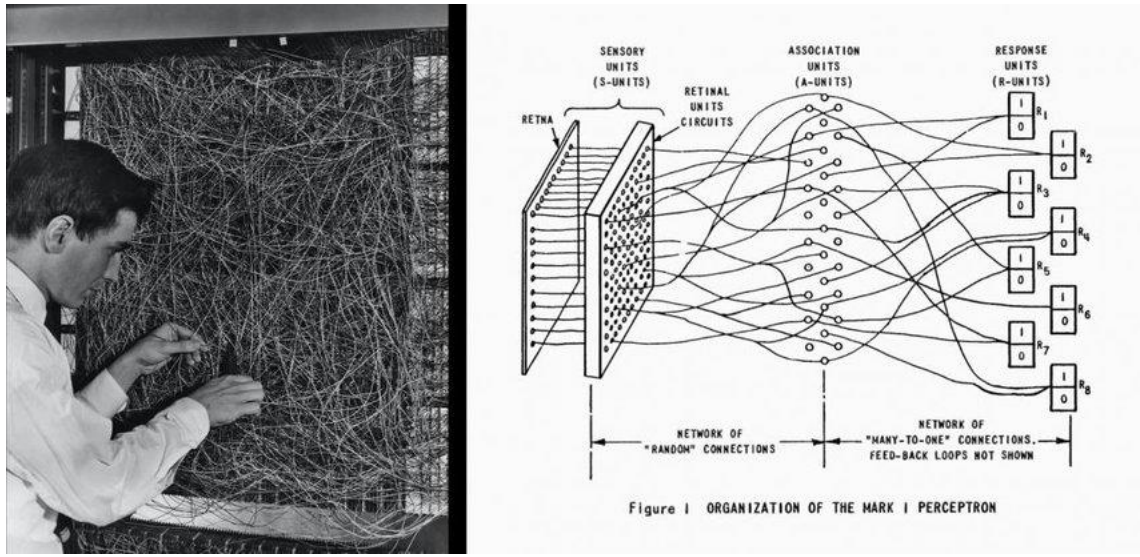
ERA5: blended observations and physics for over 50 years

- Run Data Assimilation system with available historical observations to get best estimates to the state of the atmosphere
- Available 1940 to present
- ~ 0.25 degree resolution
- Hourly estimates to state of the atmosphere combining observations and model physics
- Physically coherent and self-consistent
- Very large dataset! < 5 Petabytes



Step-back: Rise of Machine Learning…

- Machine Learning not a “new” idea
- Combination of access to ever increasing data volume and progress in compute (GPUs) makes its application possible and much more effective now
- Increased compute power (and available data) makes machine learning possible today
  - Explosion in data volumes following digitisation
  - “Domain specific compute architectures” e.g. GPUs
- Data produced by ECMWF and other centres presents significant opportunity for training a “Data-Driven” model



Frank Rosenblatt programming his Perceptron, 1970s

Step-back: Rise of Machine Learning…

Defining the dataset, split, headline fields and metrics

2020 WeatherBench

ECMWF staff
~500km ERA5 to predict future z500. Similar work from Rasp and Weyn.

Feb 2022
Full medium-range NWP

Keisler - GraphNN 1°, competitive with GFS

NVIDIA – FourCastNet
Fourier+ , 0.25°

O(10^4) faster & more energy efficient than IFS

Nov 2022
Tropical cyclones

Huawei – PanguWeather
0.25° hourly product

"More accurate tracks" than the IFS.

Dec 2022
Extensive predictions

Deepmind – GraphCast
0.25° 6-hour

Many variables and pressure levels with comparable skill to IFS.

Jan 2023
Global & Limited Area

Microsoft – ClimaX

Forecasting various lead-times at various resolutions, both globally and regionally

Apr 2023
7-day+ scores improve

FengWu – China academia + Shanghai Met Bureau
0.25° 6-hour product

Improves on GraphCast for longer leadtimes (still deterministic)

Spherical harmonics

NVIDIA – SFNO
0.25° 6-hour product

Extension of FourCastNet to Spherical harmonics, improved stability

Diffusion modelling

Alibaba – SwinRDM
0.25° 6-hour product

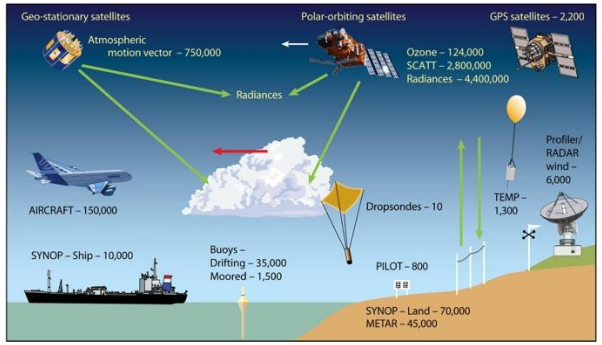
Sharp spatial features

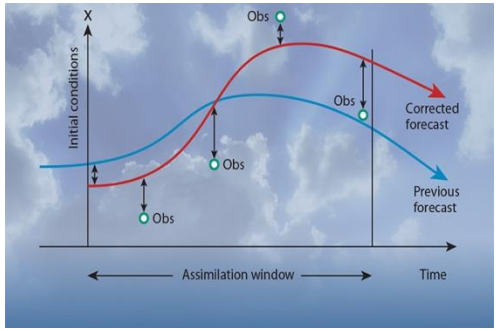
Jun 2023

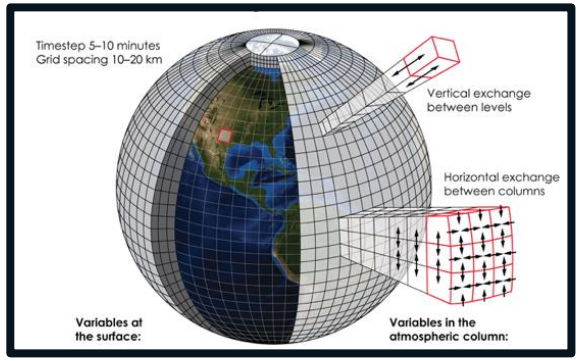
And so on…
AIFS
FuXi
AtmoRep
FuXi-extreme
NeuralGCM
GenCast
…

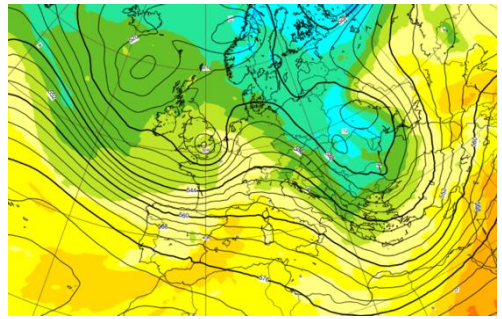
impossible to keep this figure updated!

AIFS: Data-driven Weather Forecasting System



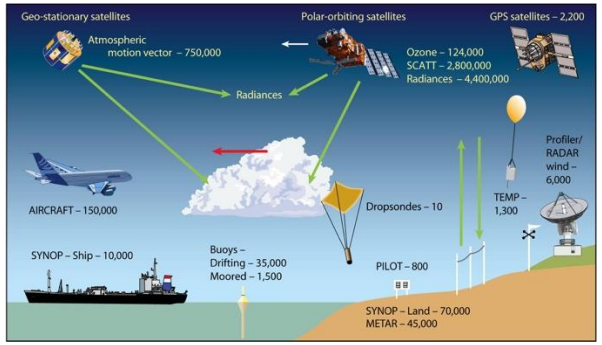


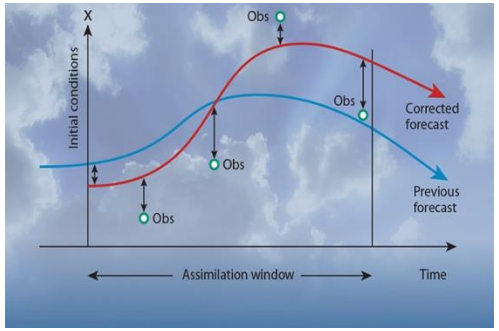


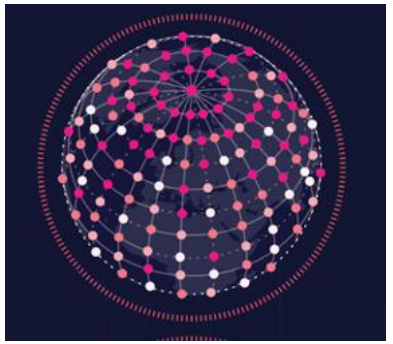


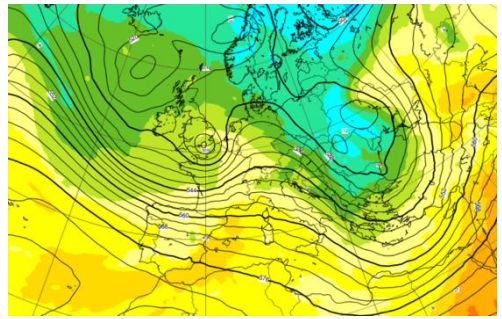
Observations → Data assimilation → Physics based forecast model → Post-processing and dissemination

AIFS: Data-driven Weather Forecasting System



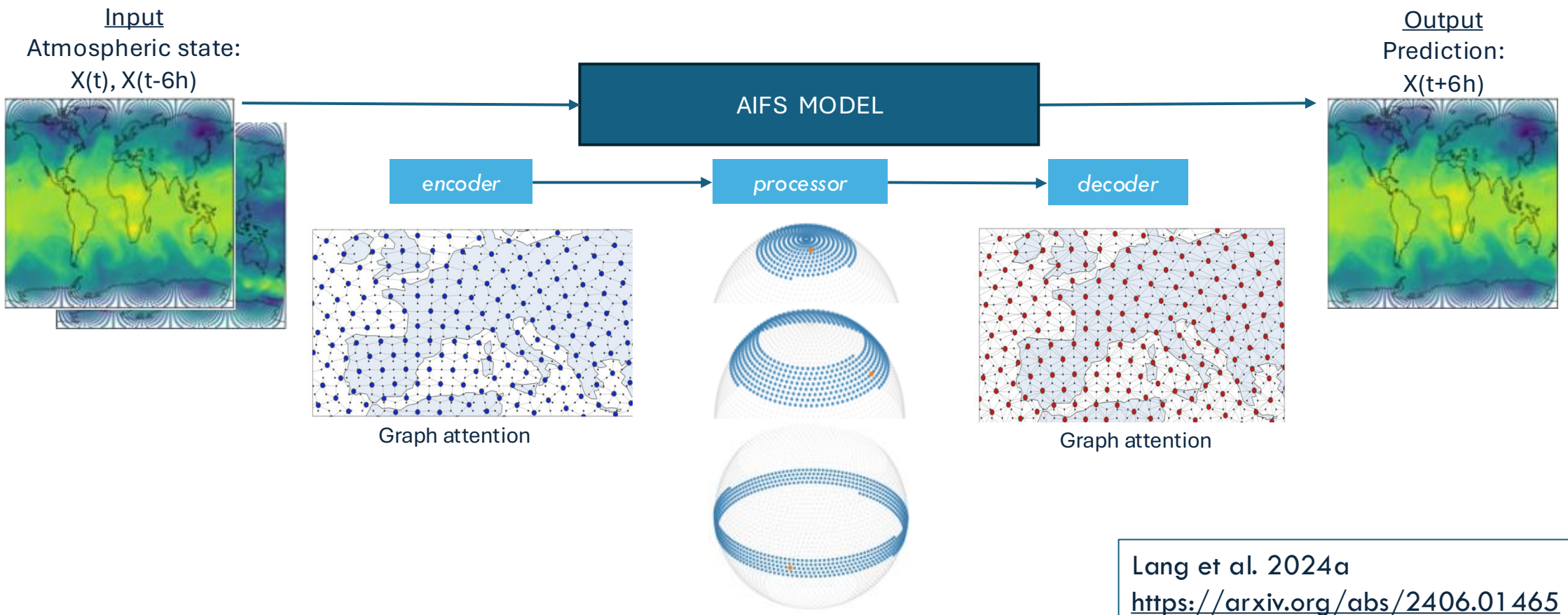






Observations → Data assimilation → "Data-driven" forecasting model → Post-processing and dissemination

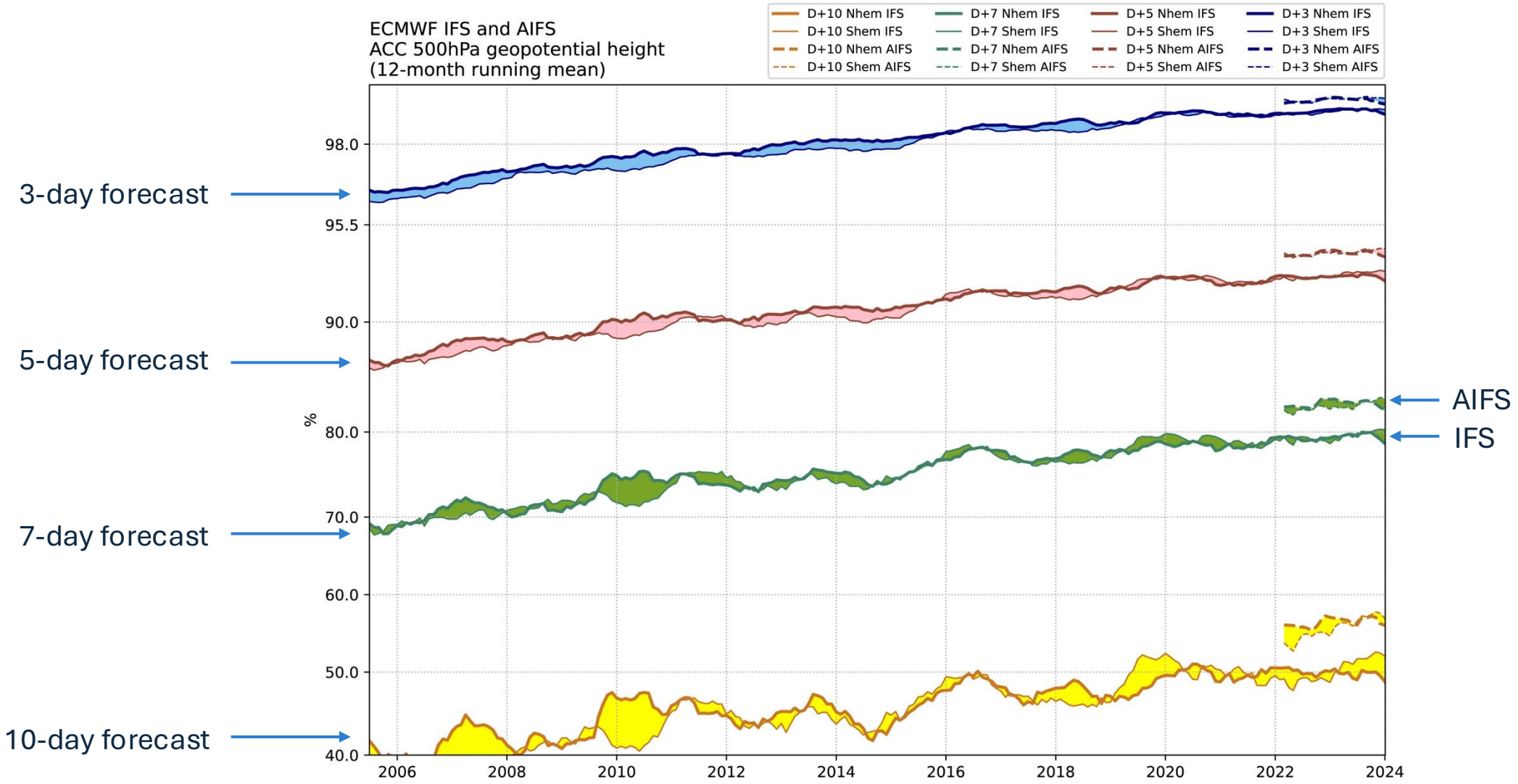
AIFS: Data-driven Weather Forecasting System



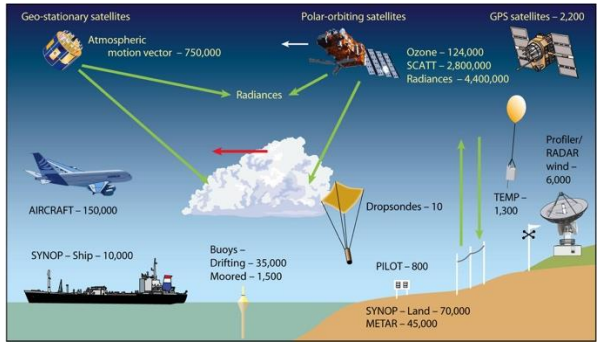
Transformer blocks and windowed attention (attention across regional bands).

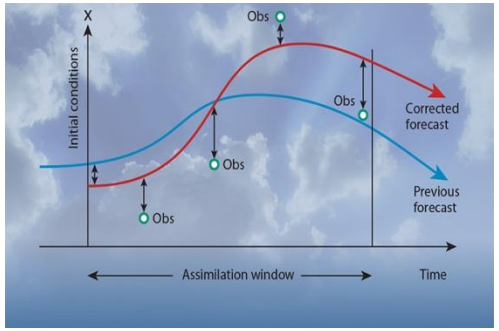
Lang et al. 2024a
https://arxiv.org/abs/2406.01465

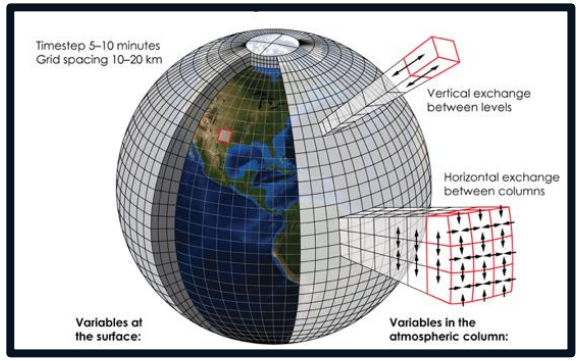
AIFS: Forecast Performance

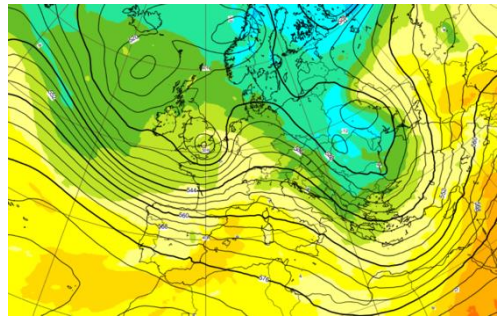


AI-Direct Observation Prediction (AI-DOP)



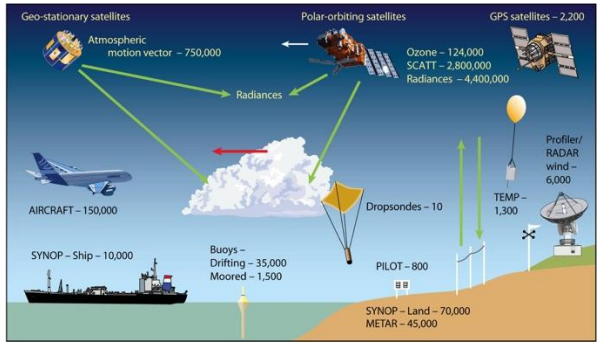


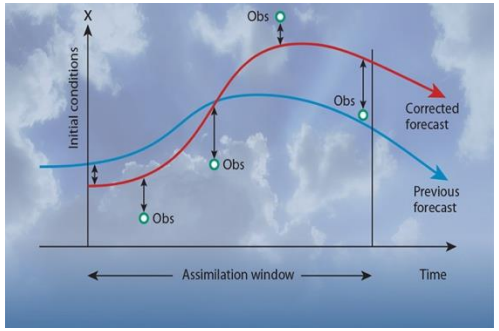


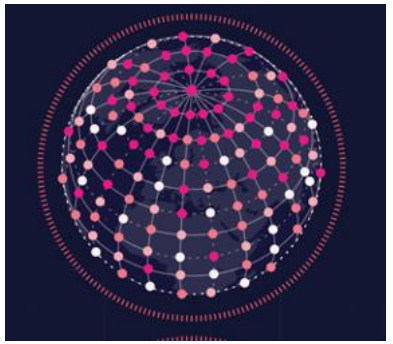


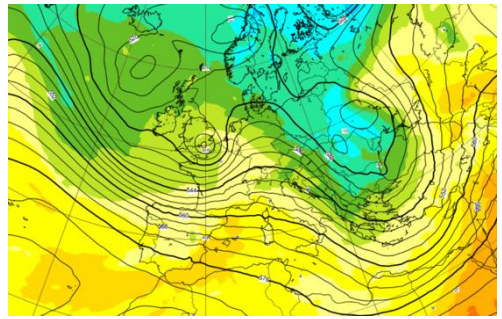
Observations → Data assimilation → Physics based forecast model → Post-processing and dissemination

AI-Direct Observation Prediction (AI-DOP)



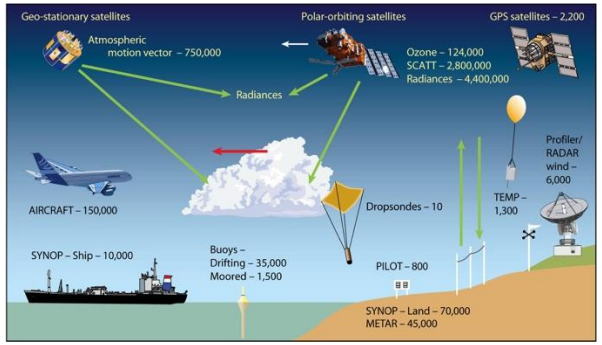


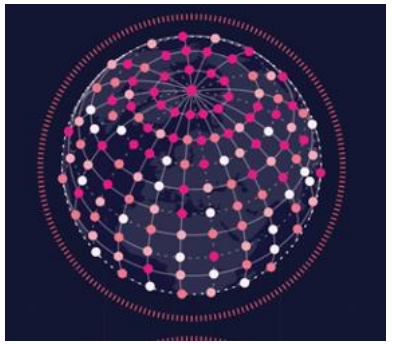


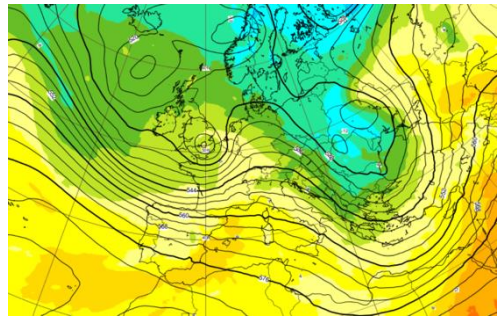


Observations → Data assimilation → "Data-driven" forecasting model → Post-processing and dissemination

AI-Direct Observation Prediction (AI-DOP)



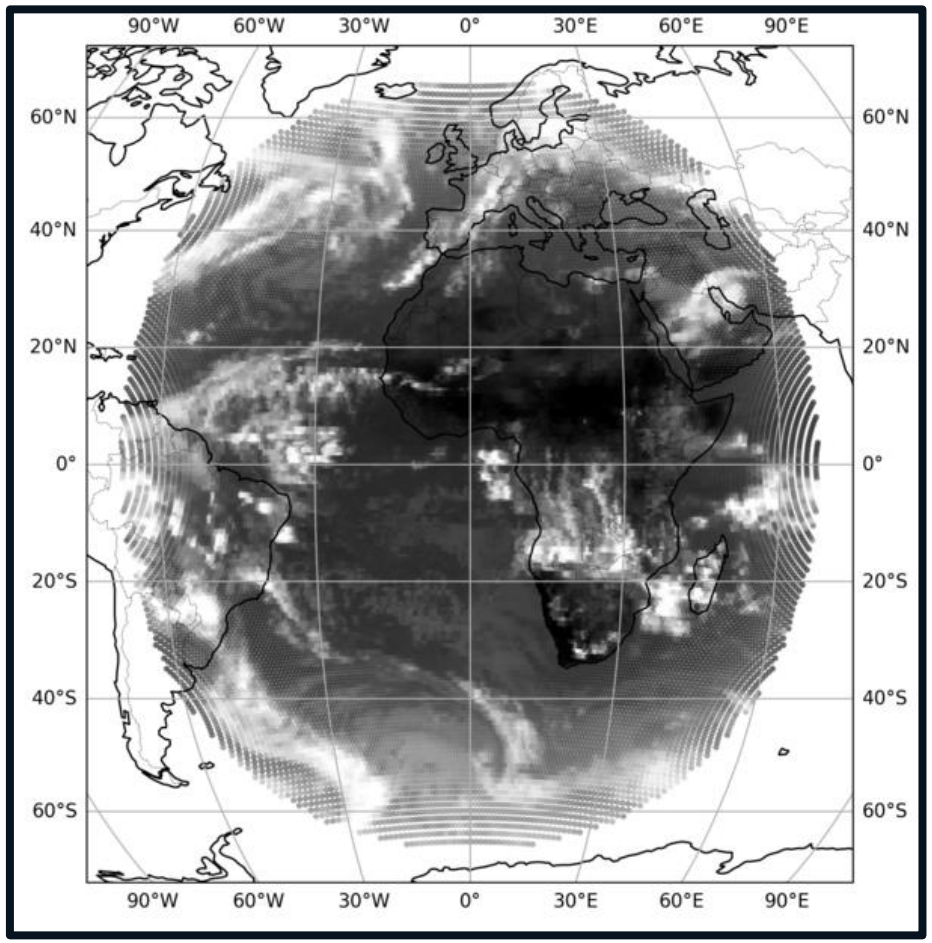




Observations → "Data-driven" forecasting model → Post-processing and dissemination

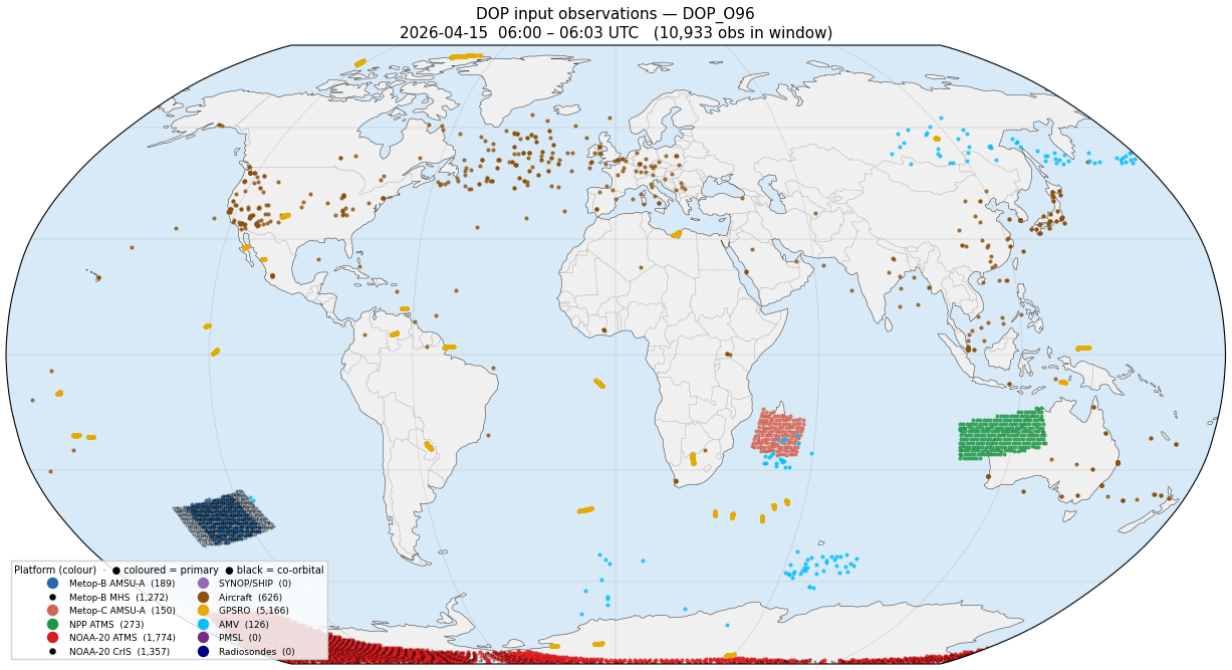
AI-Direct Observation Prediction (AI-DOP)

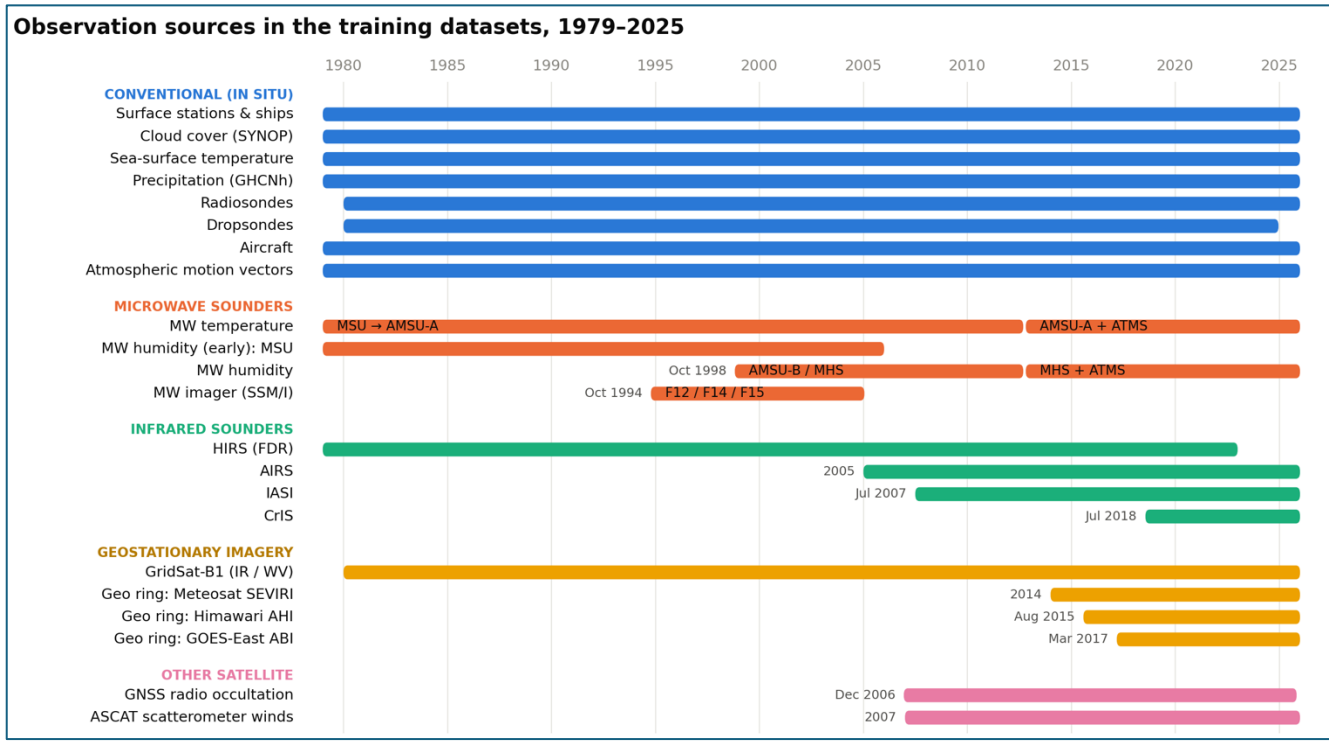
- We use historical observations to train a Neural Network to forecast future observations
- Include observations of the full Earth system (atmosphere, ocean, land) simultaneously
- Use all observations, without demanding a detailed physical model of the measurement
- Initialize model directly from observations:
  - Lower latency for forecasts
  - Faster access to warnings



Based on large observation datasets

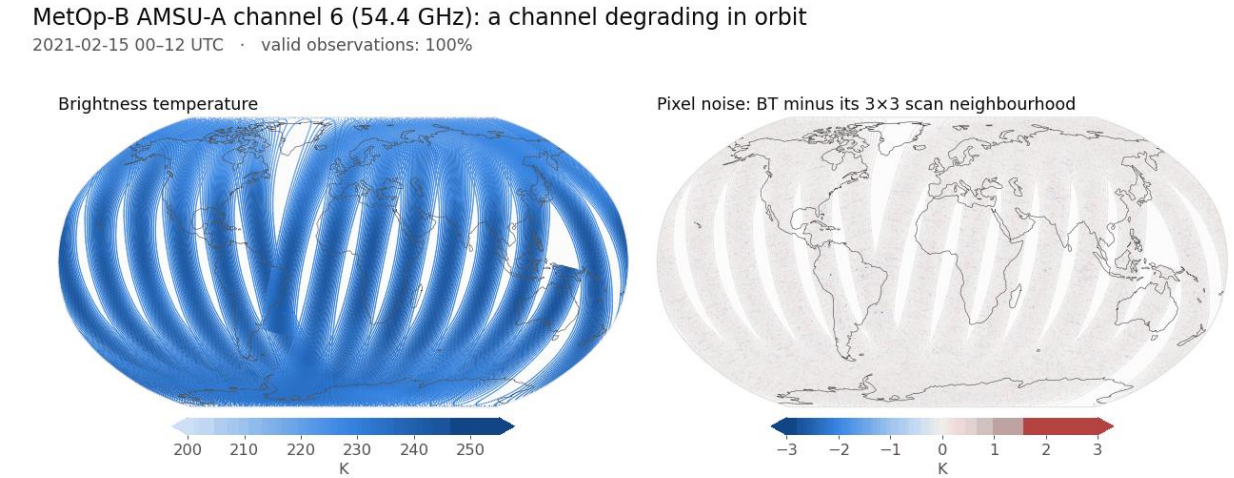
- Building observation datasets for ML training
- QC remains crucial (noisy observations can have a negative impact on model training… more on that next!)
- Challenge is then to learn spatial and temporal correlations between all these measurements

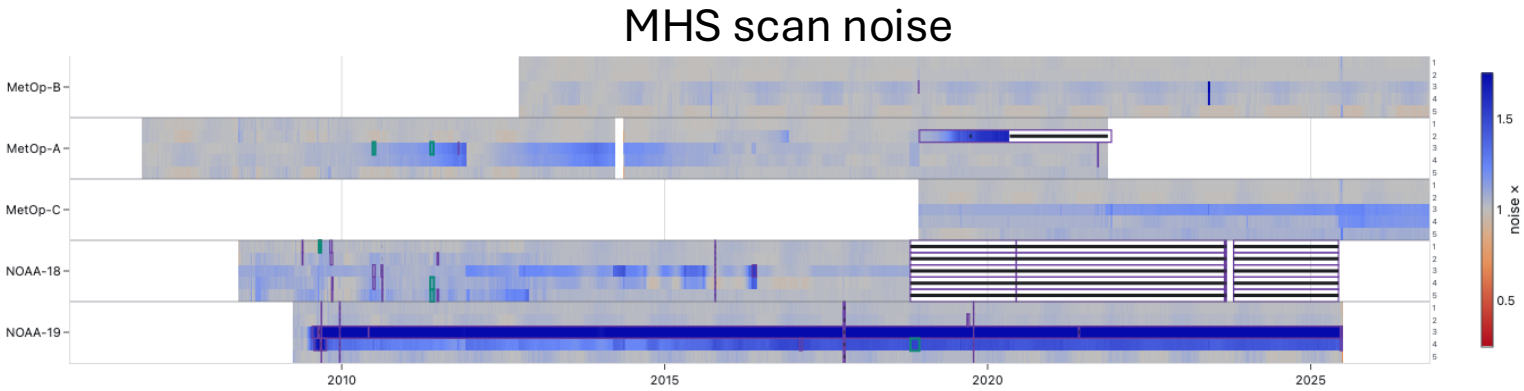




Data, Data, Data

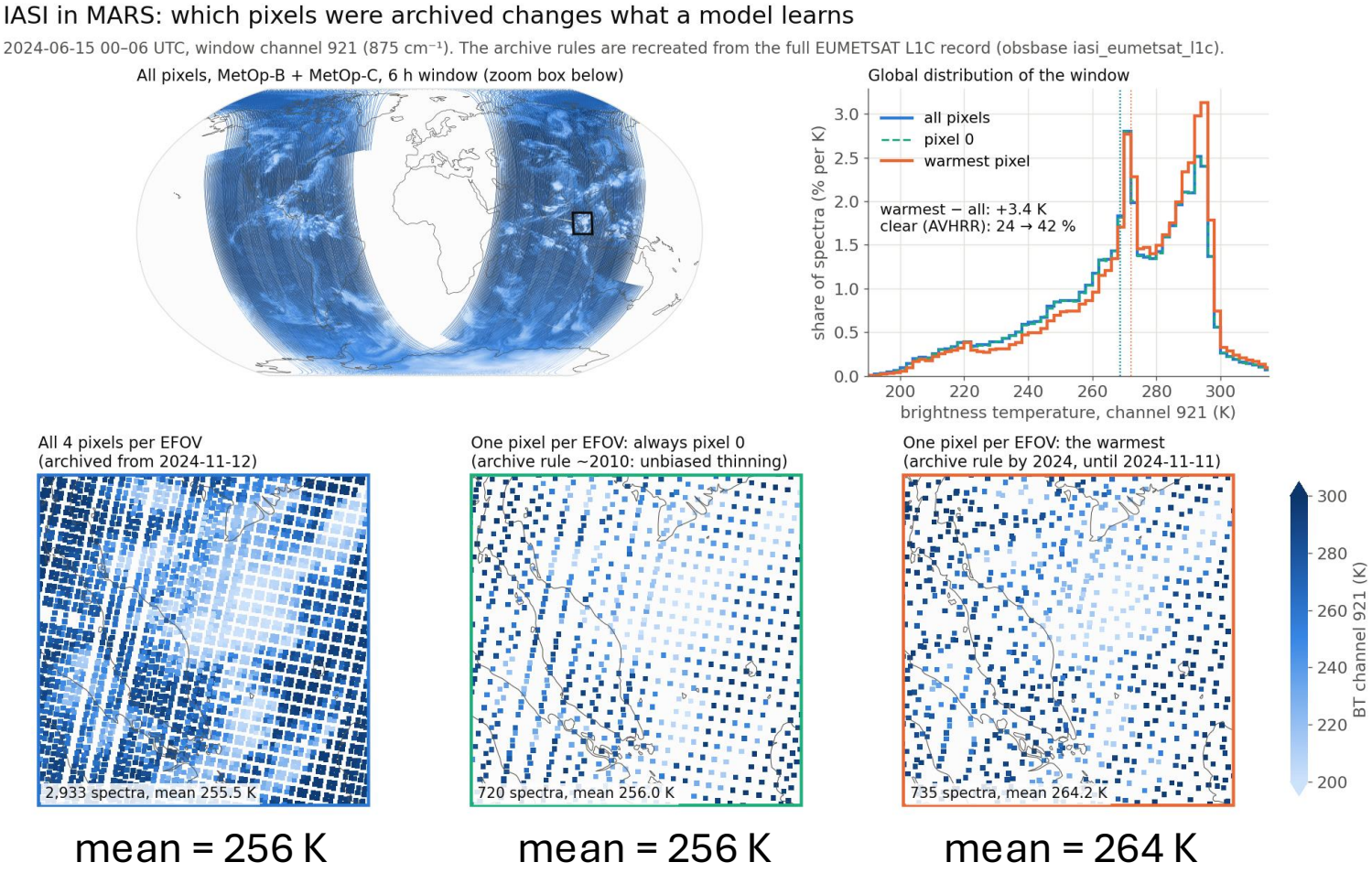
- Problem changes from a set “clean” dataset with focus on model development to equal effort needing to go into producing training datasets from observations
- Issues with observations:
  - Satellite channels degrading/dying/noising (don’t want to teach model to predict noise)
  - Mislabeled dates for gathered observations
  - Mislabeled coordinates (latitude vs longitude)
  - Distribution shifts in archiving practices
  - Skewed statistics for “conventional” observations (Hadley cell example)
  - Combining multiple different sources through time, different sensors, different measurement strategies





Data, Data, Data

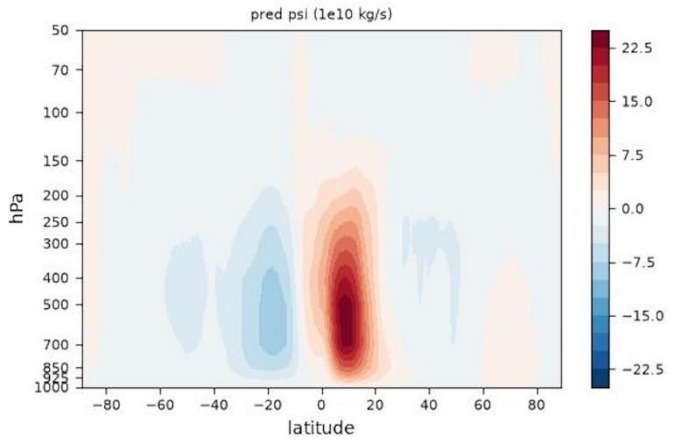
- Problem changes from a set “clean” dataset with focus on model development to equal effort needing to go into producing training datasets from observations
- Issues with observations:
  - Satellite channels degrading/dying/noising (don’t want to teach model to predict noise)
  - Mislabeled dates for gathered observations
  - Mislabeled coordinates (latitude vs longitude)
  - Distribution shifts in archiving practices
  - Skewed statistics for “conventional” observations (Hadley cell example)
  - Combining multiple different sources through time, different sensors, different measurement strategies



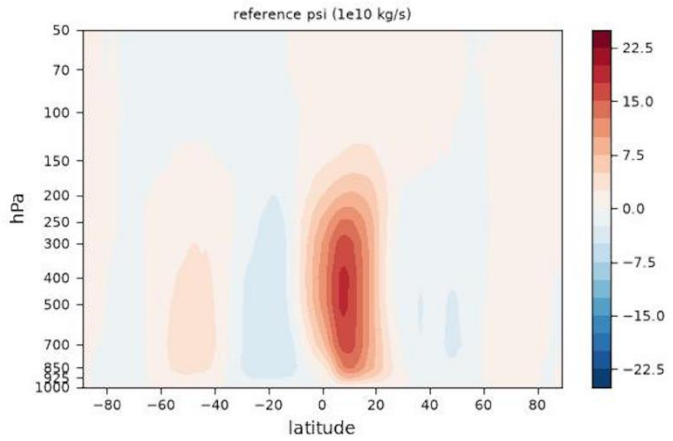
Data, Data, Data

- Problem changes from a set “clean” dataset with focus on model development to equal effort needing to go into producing training datasets from observations
- Issues with observations:
  - Satellite channels degrading/dying/noising (don’t want to teach model to predict noise)
  - Mislabeled dates for gathered observations
  - Mislabeled coordinates (latitude vs longitude)
  - Distribution shifts in archiving practices
  - Skewed statistics for “conventional” observations (Hadley cell example)
  - Combining multiple different sources through time, different sensors, different measurement strategies

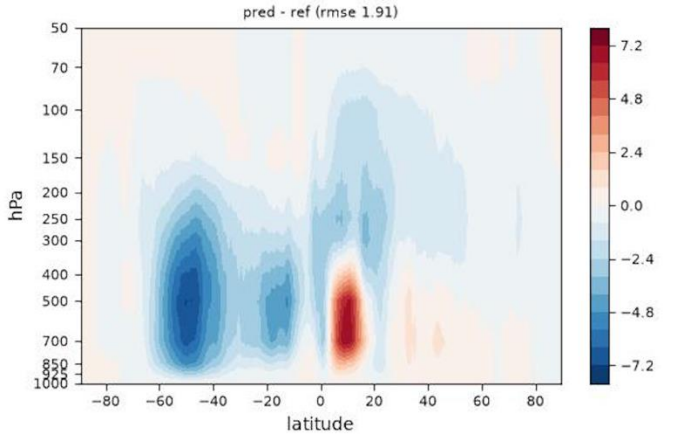
DOP

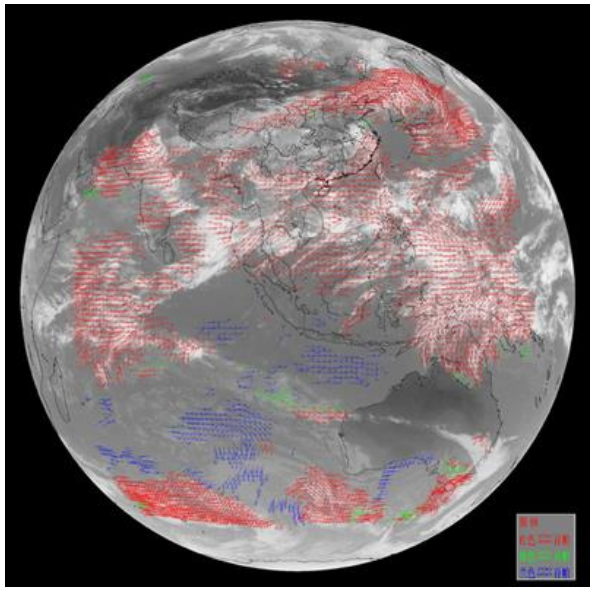


IFS



Error = 1.9

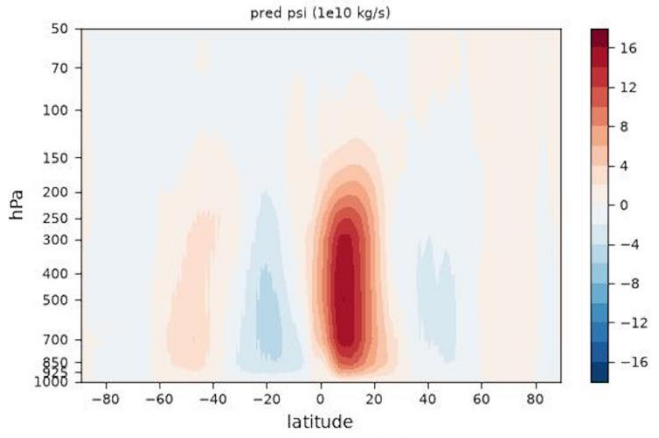




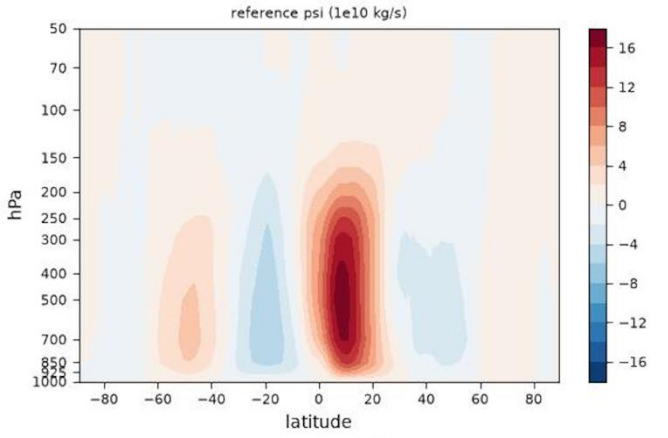
Data, Data, Data

- Problem changes from a set “clean” dataset with focus on model development to equal effort needing to go into producing training datasets from observations
- Issues with observations:
  - Satellite channels degrading/dying/noising (don’t want to teach model to predict noise)
  - Mislabeled dates for gathered observations
  - Mislabeled coordinates (latitude vs longitude)
  - Distribution shifts in archiving practices
  - Skewed statistics for “conventional” observations (Hadley cell example)
  - Combining multiple different sources through time, different sensors, different measurement strategies

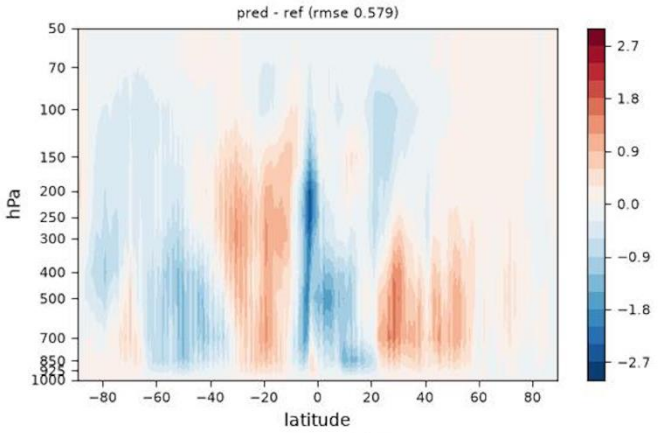
DOP

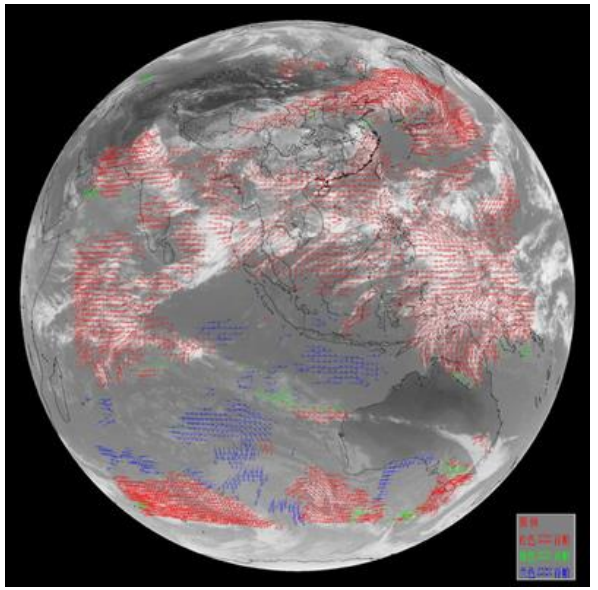


IFS



Error = 0.6



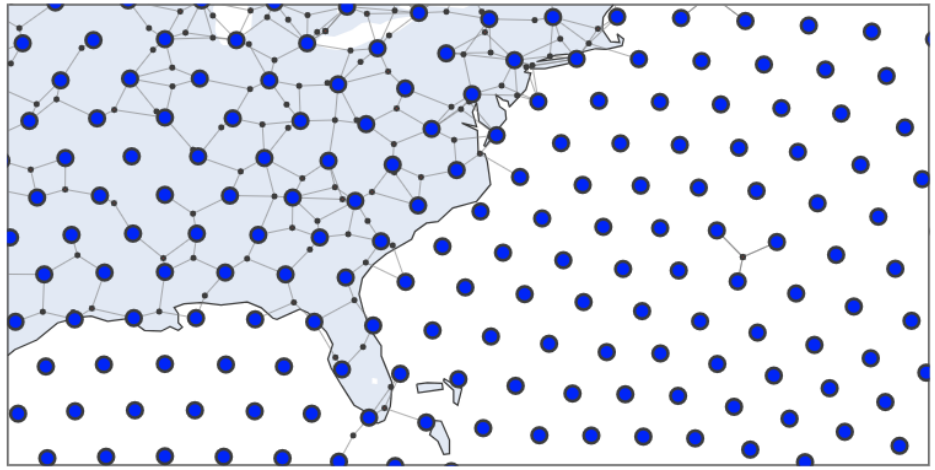


Encoding and decoding the observations with graph neural networks

Unstructured observations and dynamic graphs

anemoi

Under development



GRAPHDOP: TOWARDS SKILFUL DATA-DRIVEN MEDIUM-RANGE WEATHER FORECASTS LEARNT AND INITIALISED DIRECTLY FROM OBSERVATIONS

A PREPRINT

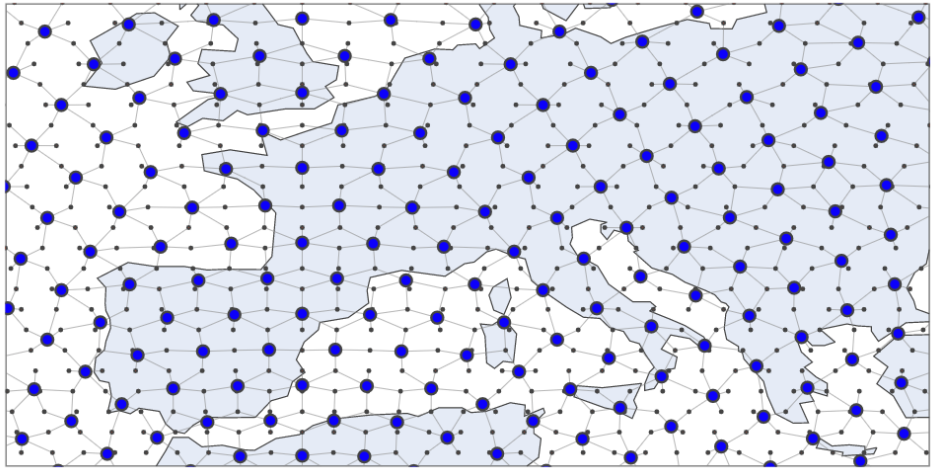
Mihai Alexe · Eulalie Boucher · Peter Lean · Ewan Pinnington · Patrick Laloyaux · Anthony McNally · Simon Lang · Matthew Chantry · Chris Burrows · Marcin Chrust · Florian Pinault · Ethel Villeneuve · Niels Bormann · Sean Healy

European Centre for Medium-Range Weather Forecasts (ECMWF)

https://arxiv.org/pdf/2412.15687

Gridded observations and static graphs

anemoi



AIFS-DOP: END-TO-END MEDIUM-RANGE WEATHER PREDICTION FROM OBSERVATIONS ALONE WITH MACHINE LEARNING

Ewan Pinnington · Peter Lean · Mihai Alexe · Eulalie Boucher · Simon Lang · Patrick Laloyaux · Gert Mertes · Tomas Kral · Patricia de Rosnay · Matthew Chantry · Anthony McNally

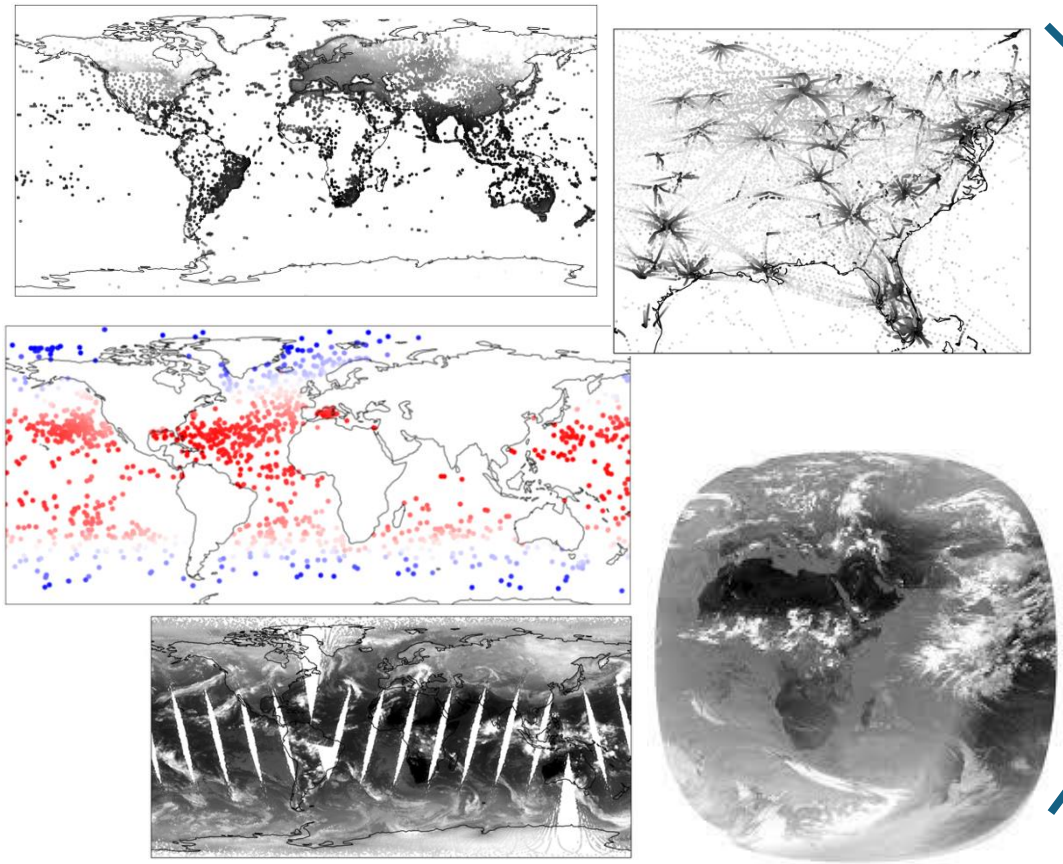
European Centre for Medium-Range Weather Forecasts (ECMWF)

https://arxiv.org/pdf/2606.19093

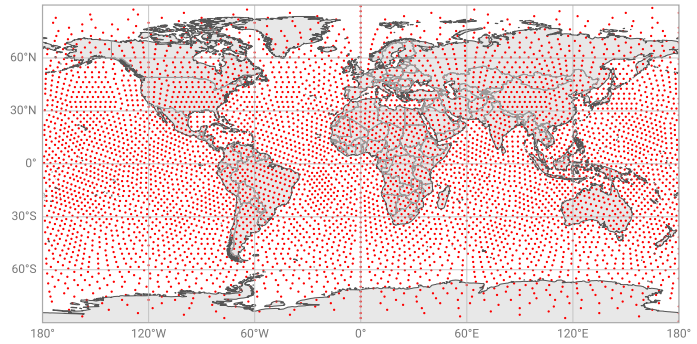
The “optimal” projection for each obs type is still an open research question

First Anemoi Examples of Training with Observations

- Putting observations on the same resolution grid in 6-hour slices
- Treat them as additional variables under the data loader
- Slight updates to NaN-handling to allow training/generalising in space



Put obs onto grid
o96 (~100 km)
Add NaNs for missing data



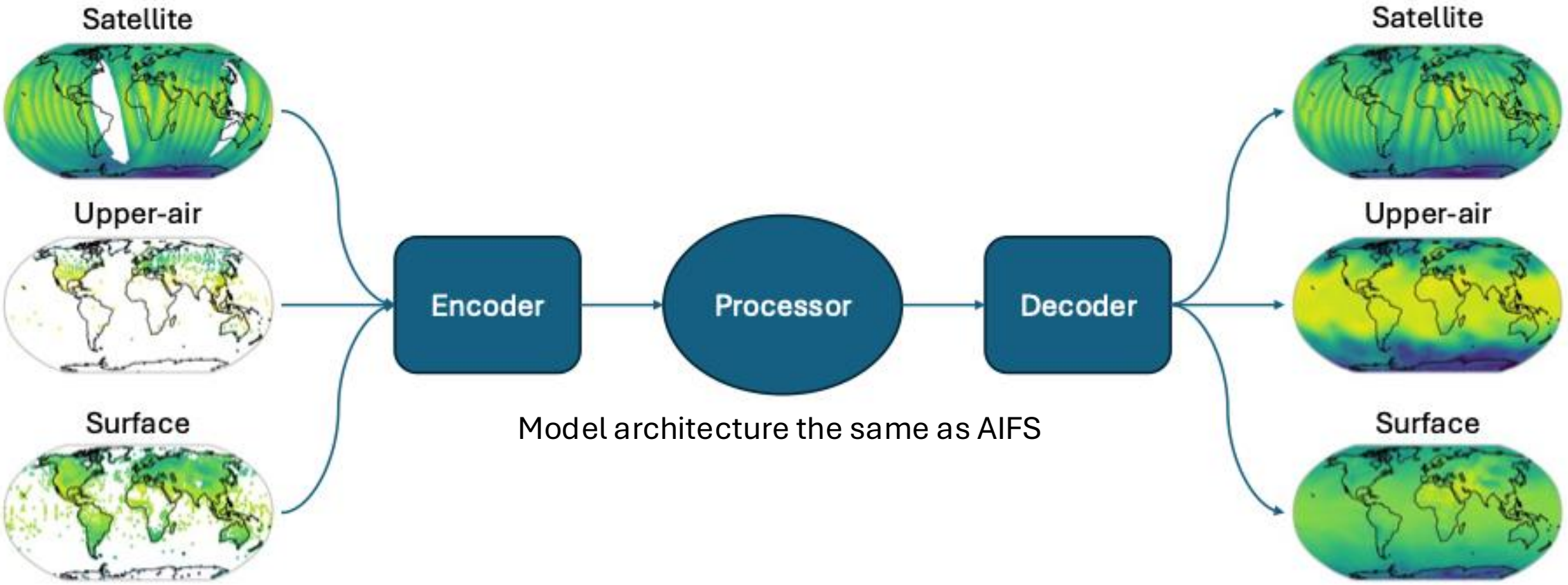
Make anemoi-zarr from
Gridded observations

anemoi datasets

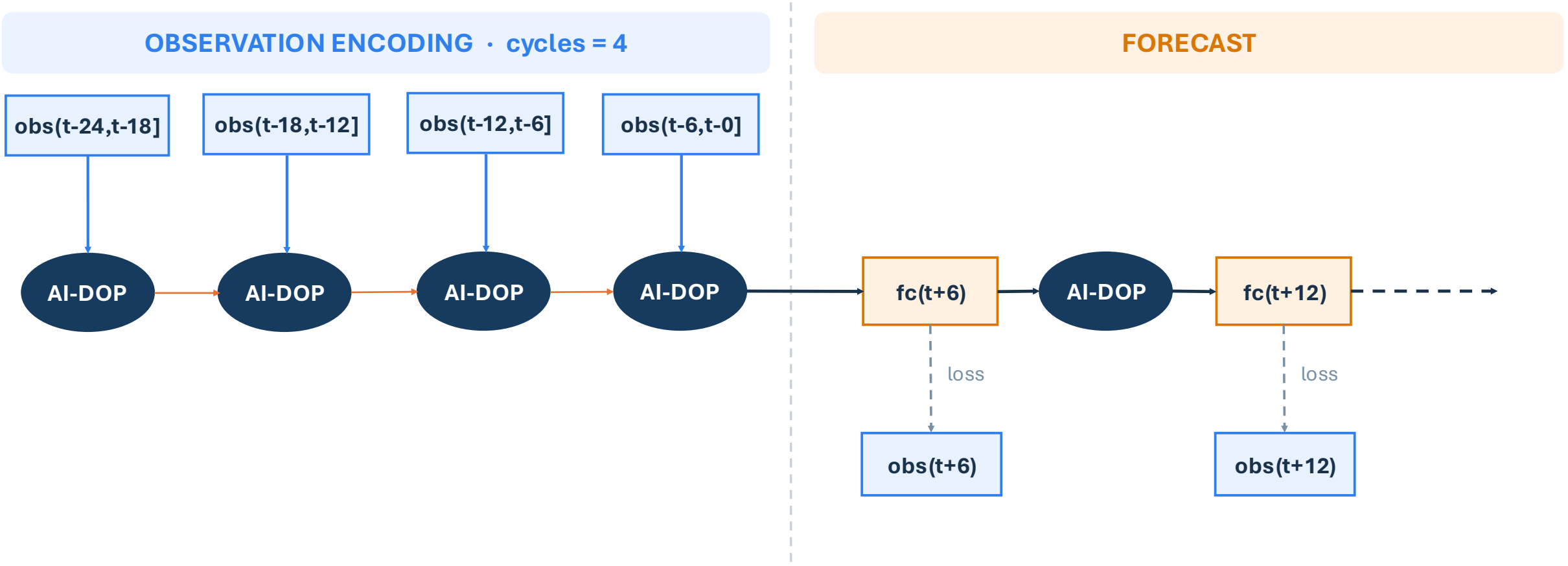
First Anemoi Examples of Training with Observations

- Training ML models directly from observations
- Using Open-Source Anemoi packages: https://github.com/ecmwf/anemoi-core
- Architecture for this example very similar to the AIFS with some tweaks

anemoi training



Observation Cycling



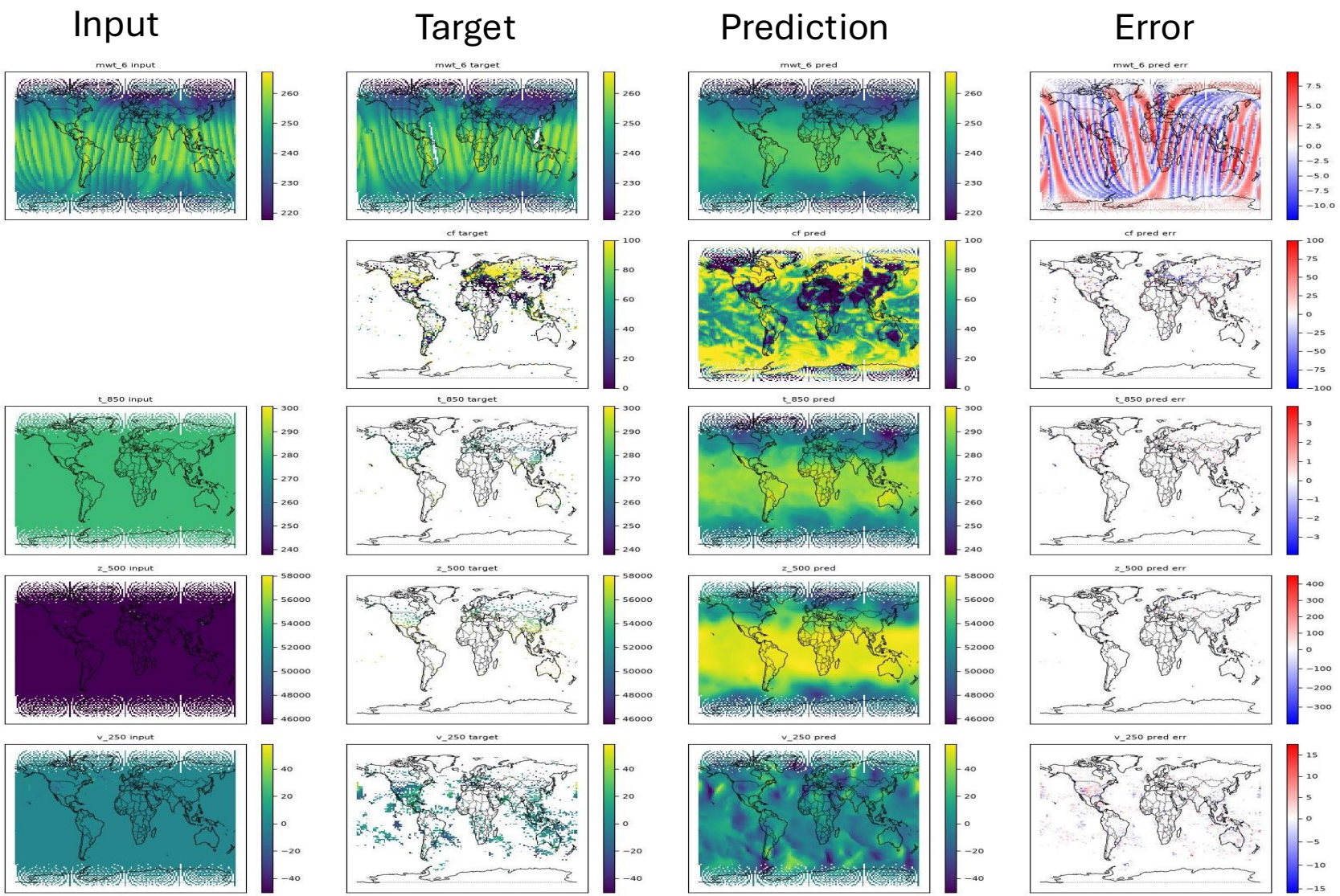
obs = observation

fc = prediction

AI-DOP
sliding-window-transformer

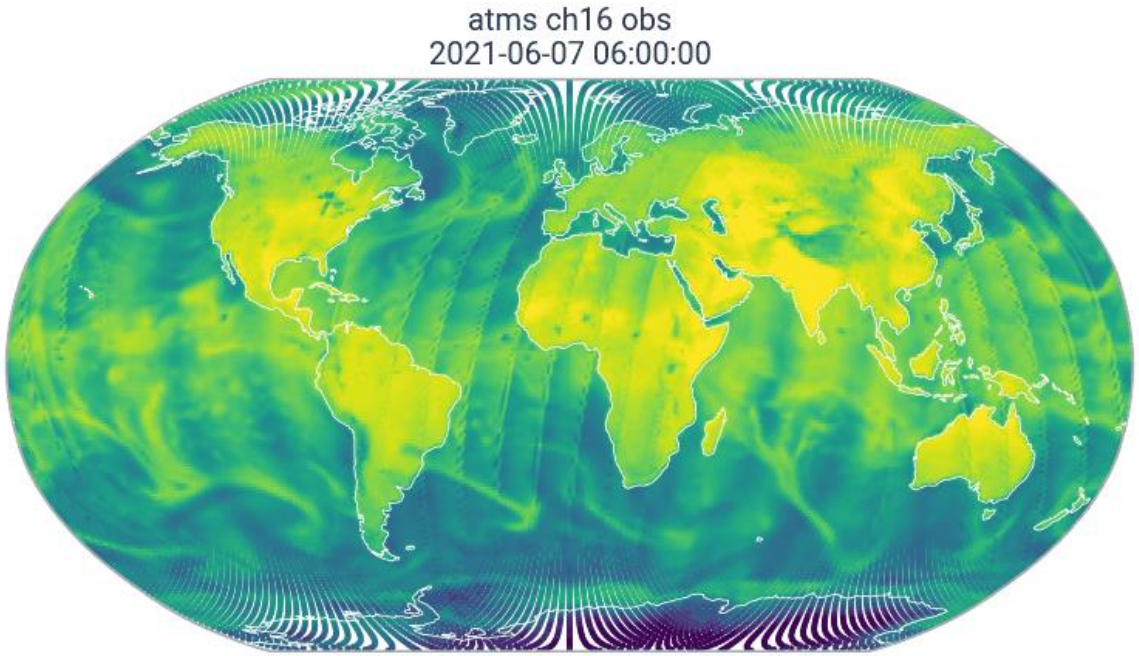
Dashed = training loss

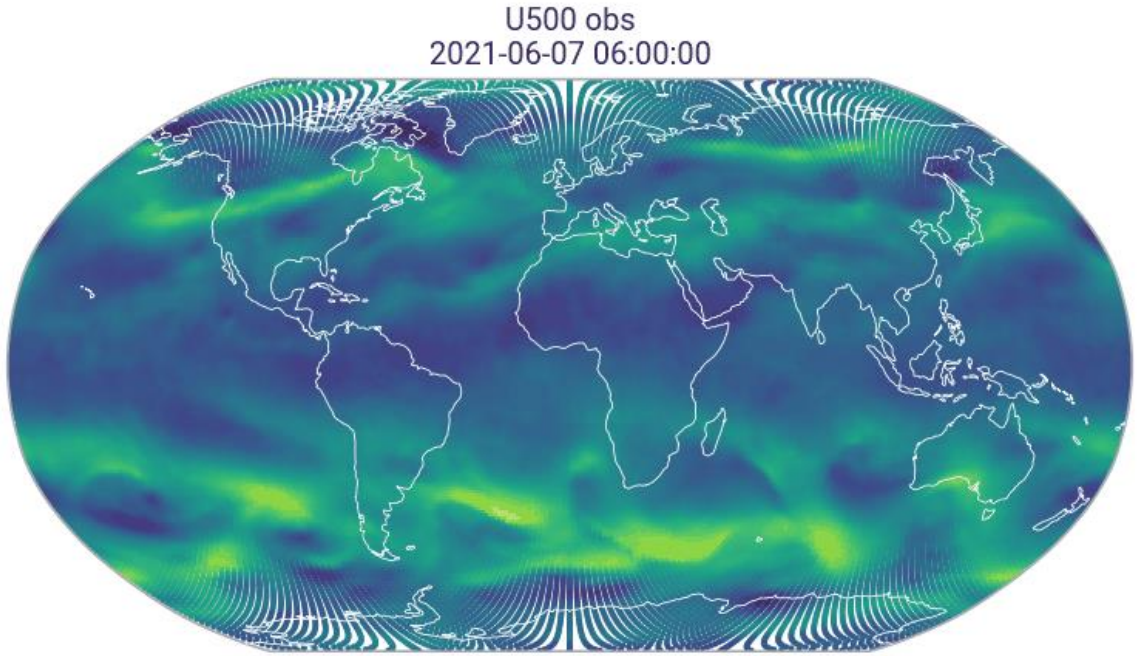
Training Example



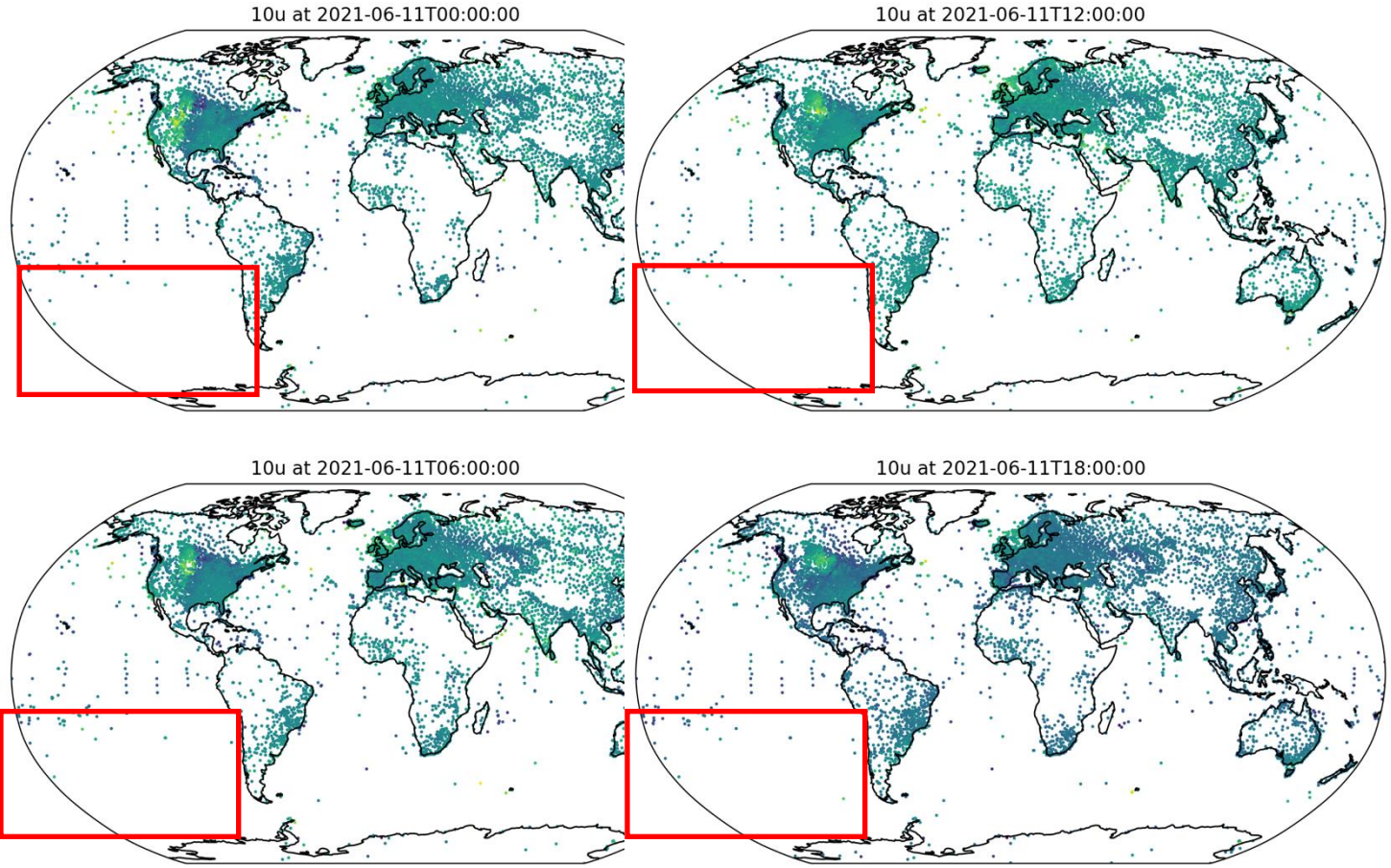
- Despite sparse observations the model forecasts globally
- Loss (MSE) is computed at locations with an observation in the target grid
- For satellites, model predicts a “flat” brightness temperature (at nadir), on top of which we apply a correction (observation operator)

Training Example

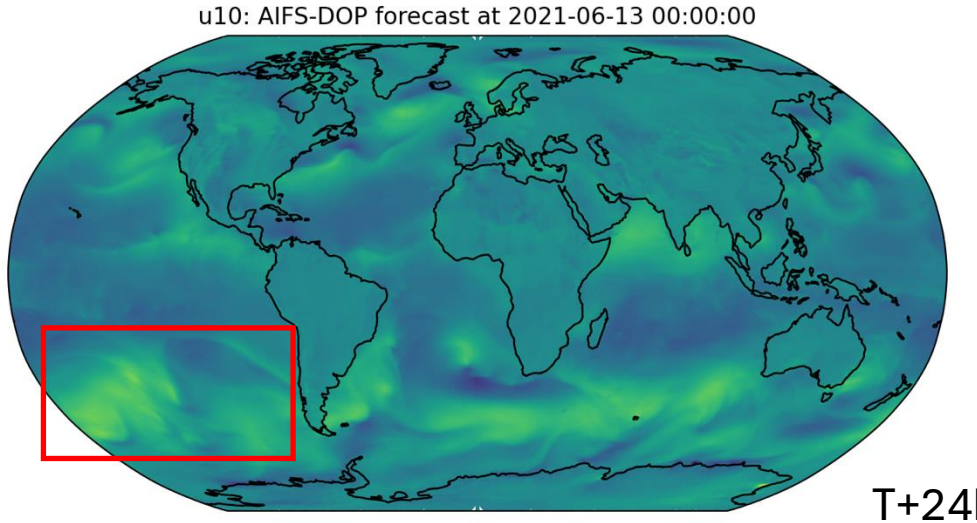




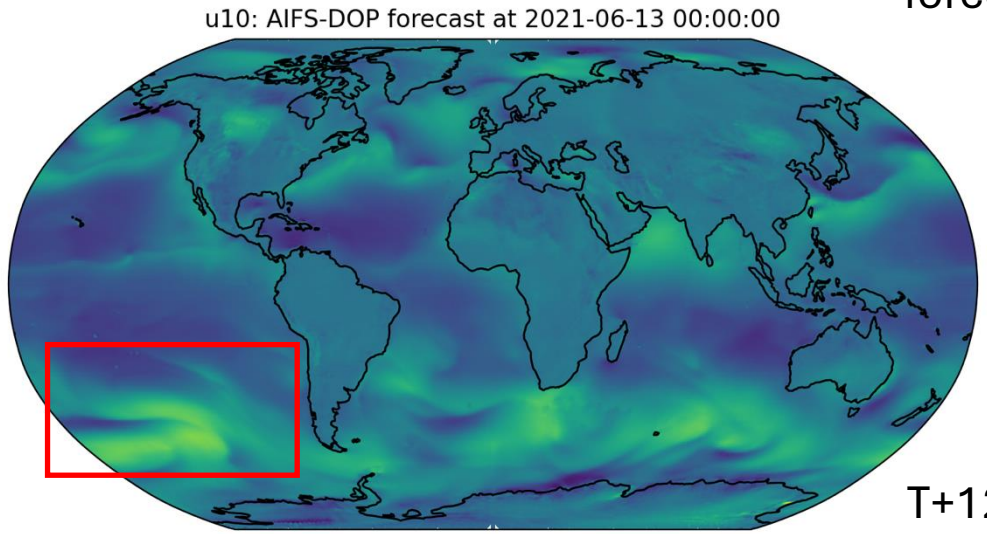
AIFS-DOP forecasts are produced from very sparse observations



Input u10 observations



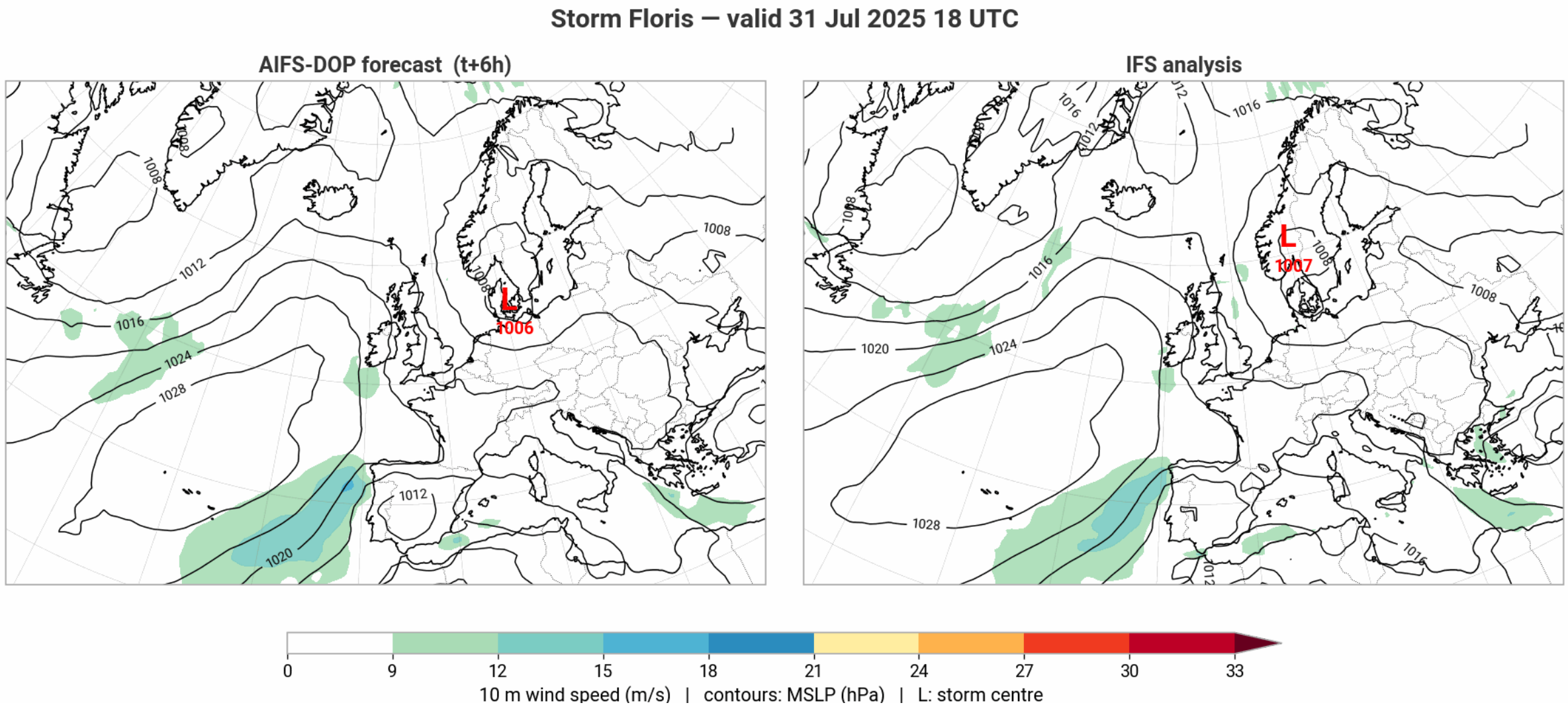
T+24h forecast



T+120h forecast

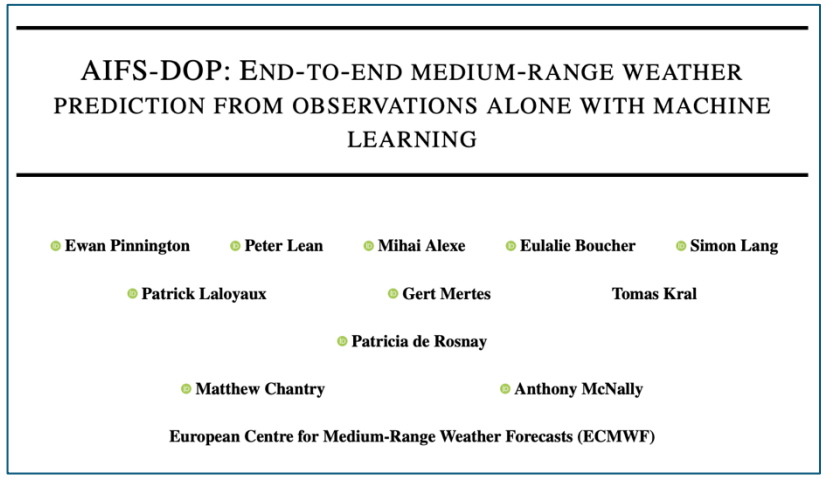
→ **Uses correlations with other input data to produce a forecast even where there are no direct observations for that variable**

Storm Floris, August 2025



Moving to real time

- Paper trained until end of 2020 and validated 2022
- Retraining model up to end of 2024
- Transitioning from older sensors to newer ones
  - *e.g.* HIRS missing – pickup CrIS
- Initial results look promising



AIFS-DOP: END-TO-END MEDIUM-RANGE WEATHER PREDICTION FROM OBSERVATIONS ALONE WITH MACHINE LEARNING

Ewan Pinnington Peter Lean Mihai Alexe Eulalie Boucher Simon Lang
Patrick Laloyaux Gert Mertes Tomas Kral
Patricia de Rosnay
Matthew Chantry Anthony McNally
European Centre for Medium-Range Weather Forecasts (ECMWF)

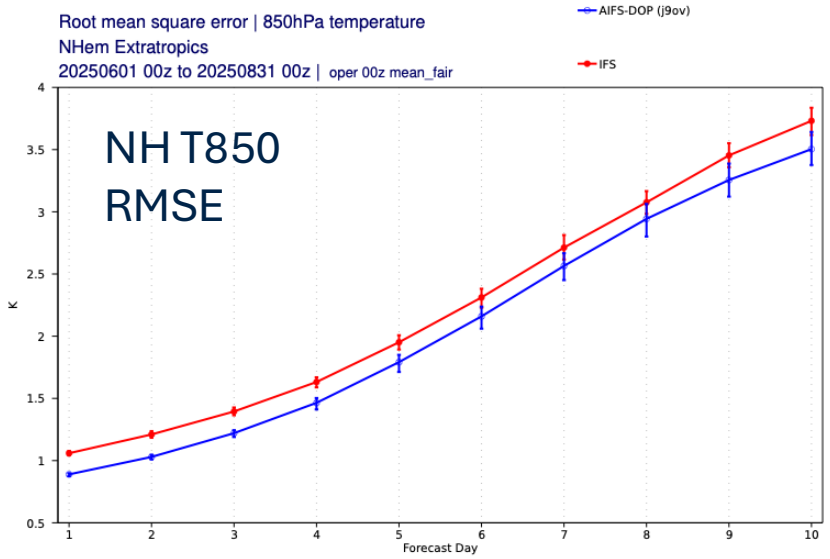
https://arxiv.org/pdf/2606.19093

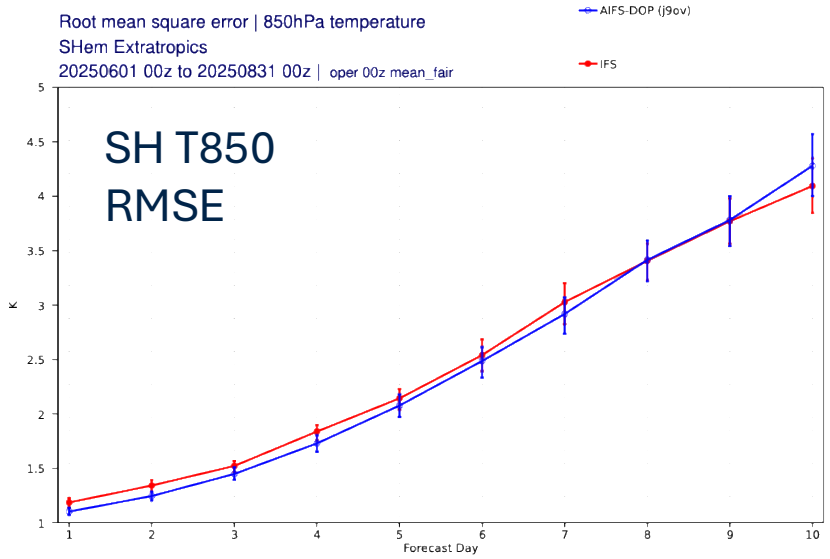
Table 1: Description of curated observation dataset

Category	Instruments	Period	Variables	Details
~~Infrared Sounder~~	~~HIRS~~	~~1980–2022~~	~~Brightness~~ Temperatures	EUMETSAT Fundamental Data Record [EUMETSAT, 2024]
Microwave Sounders	MSU SSM/T-2 AMSU-A AMSU-B MHS ATMS	1980–2005 1994–2005 1998–2021 1998–2014 2005–2022 2012–2022	Brightness Temperatures	MSU taken from NOAA Climate Data Record [Zou et al., 2013], SSM/T2 taken from EUMETSAT Fundamental Data Record [EUMETSAT, 2020]
Surface Observations	SYNOP, Buoys, Ships	1980–2022	2t, 2d, msl, 10u, 10v	Existing ECMWF data archive
Upper-air Observations	~~Radiosonde~~, Aircraft, ~~AMV~~	1980–2022	t, u, v, z on pressure levels	Existing ECMWF data archive
~~Geostationary Satellite~~	~~GridSat~~	~~1980–2022~~	~~Brightness~~ Temperatures	NOAA Climate Data Record [Knapp et al., 2011]

Moving to real time

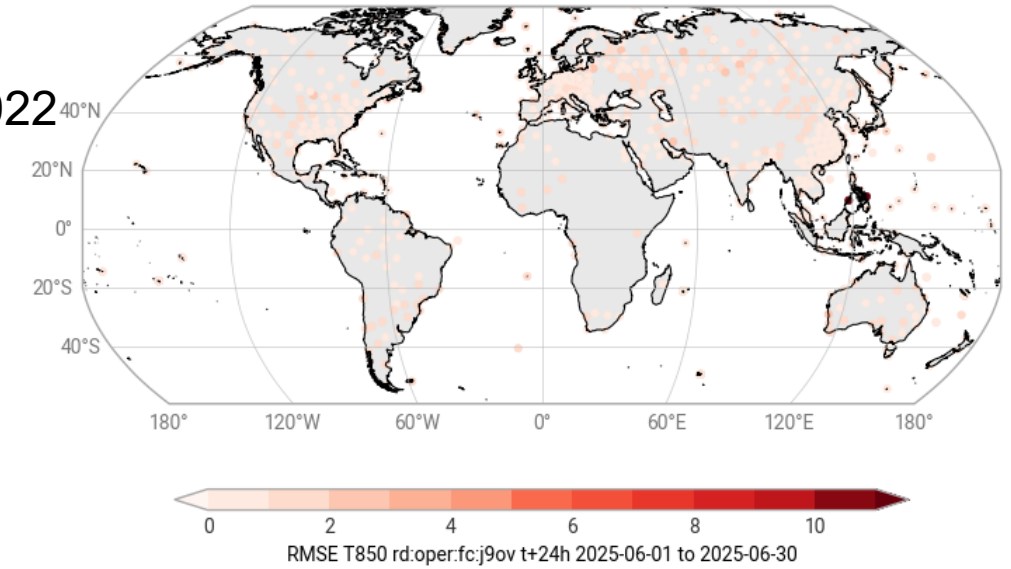
- Paper trained until end of 2020 and validated 2022
- Retraining model up to end of 2024
- Transitioning from older sensors to newer ones
  - *e.g.* HIRS missing – pickup CrIS
- Initial results look promising

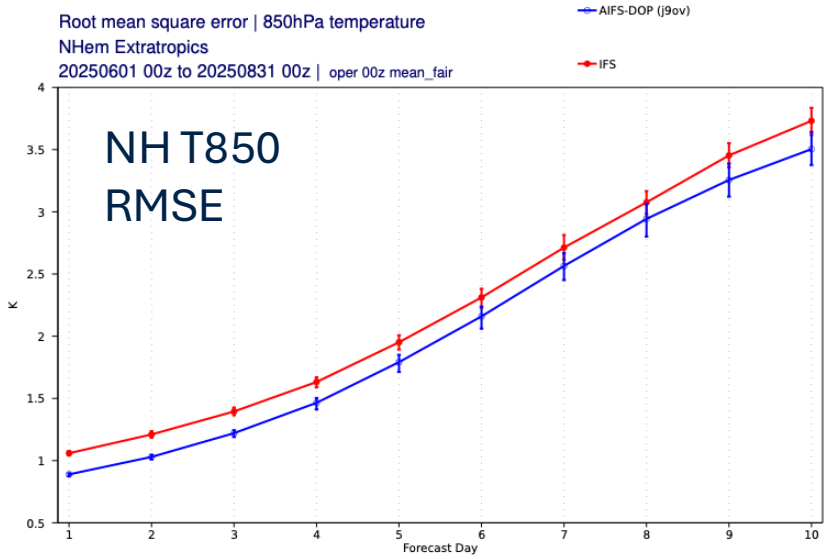


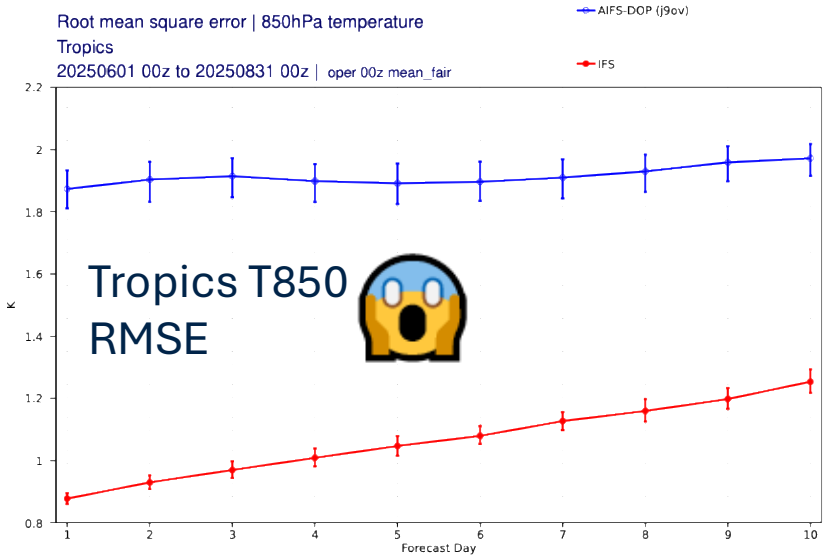


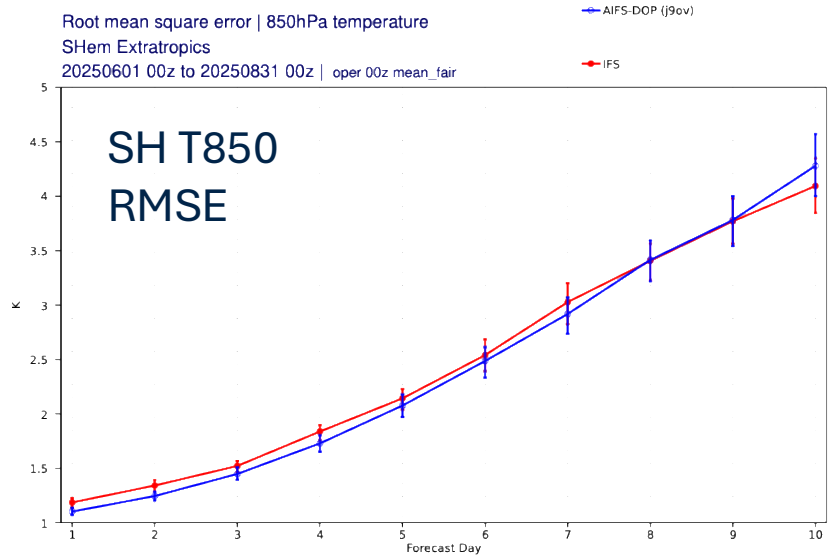
Moving to real time

- Paper trained until end of 2020 and validated 2022
- Retraining model up to end of 2024
- Transitioning from older sensors to newer ones
  - *e.g.* HIRS missing – pickup CrIS
- Initial results look promising



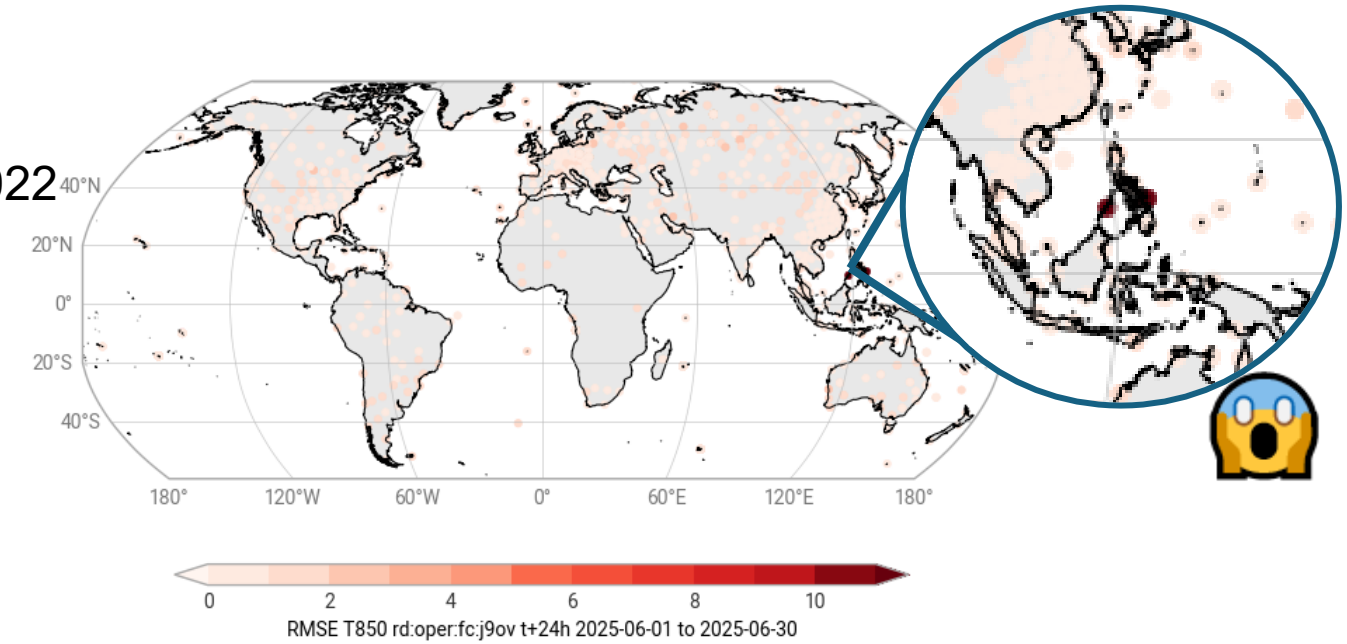


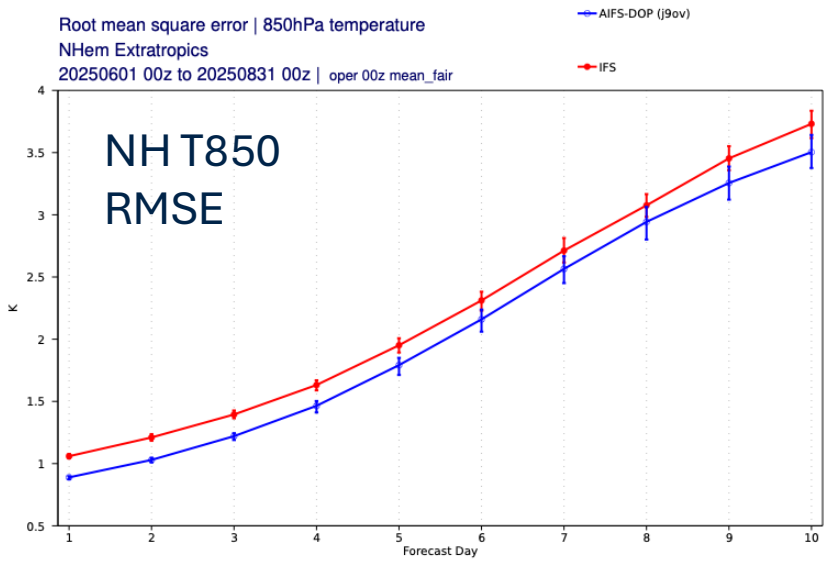


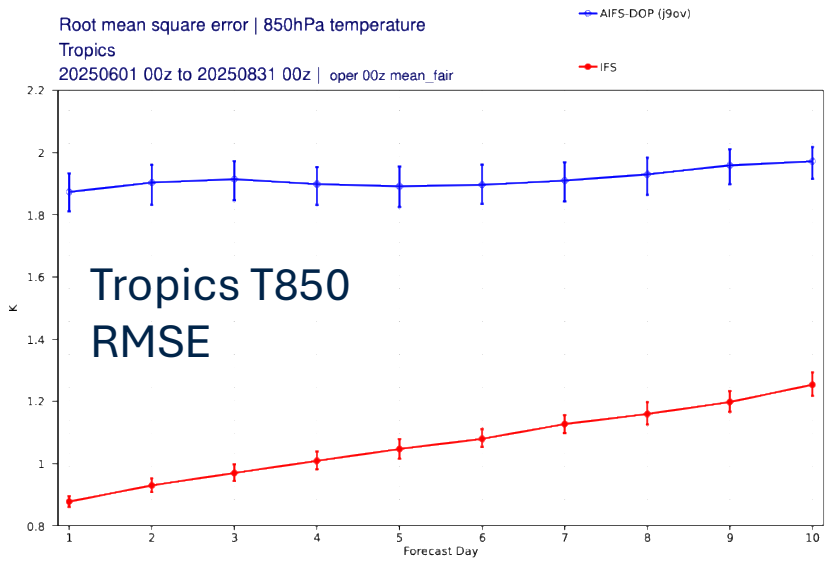


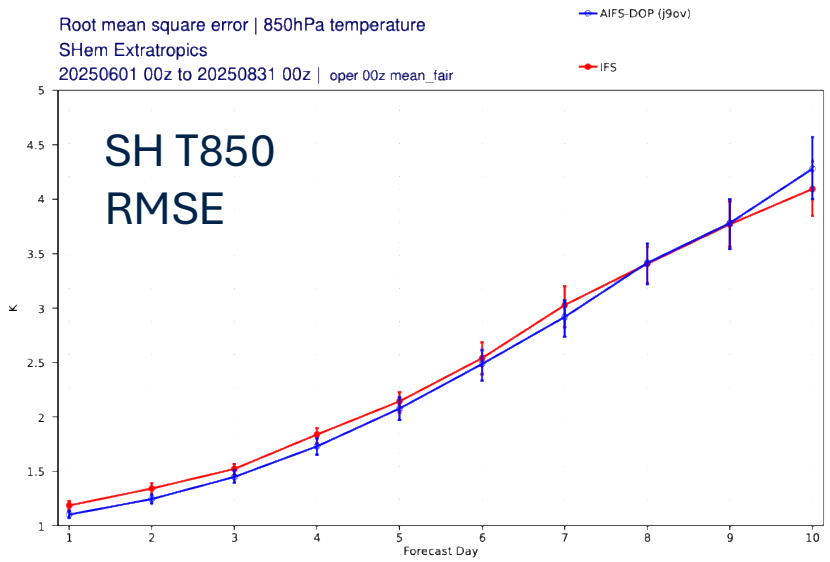
Moving to real time

- Paper trained until end of 2020 and validated 2022
- Retraining model up to end of 2024
- Transitioning from older sensors to newer ones
  - *e.g.* HIRS missing – pickup CrIS
- Initial results look promising



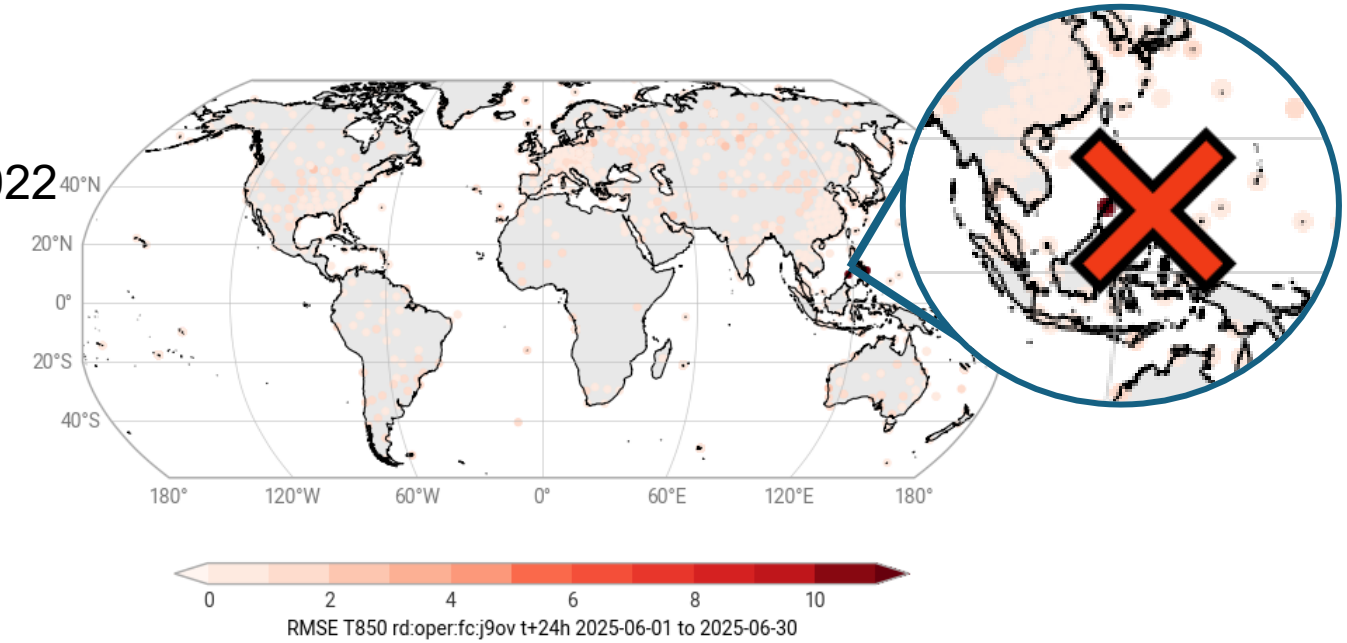


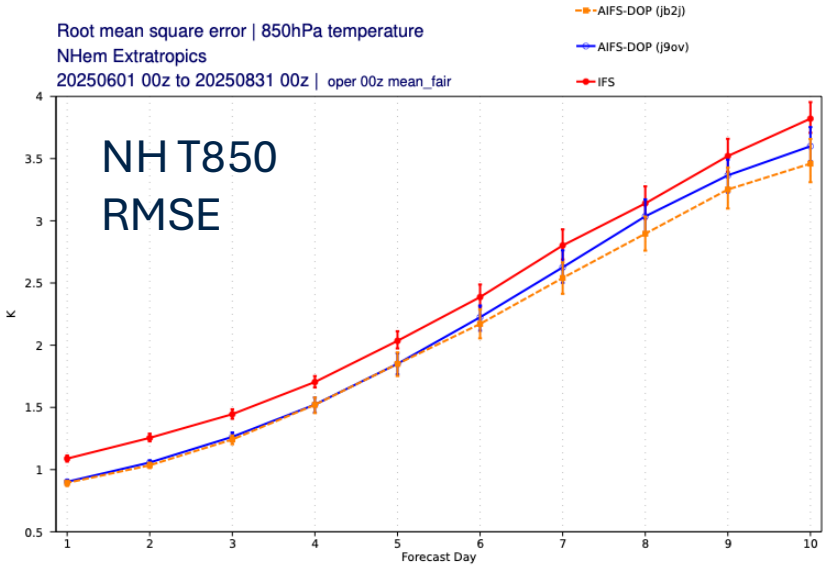


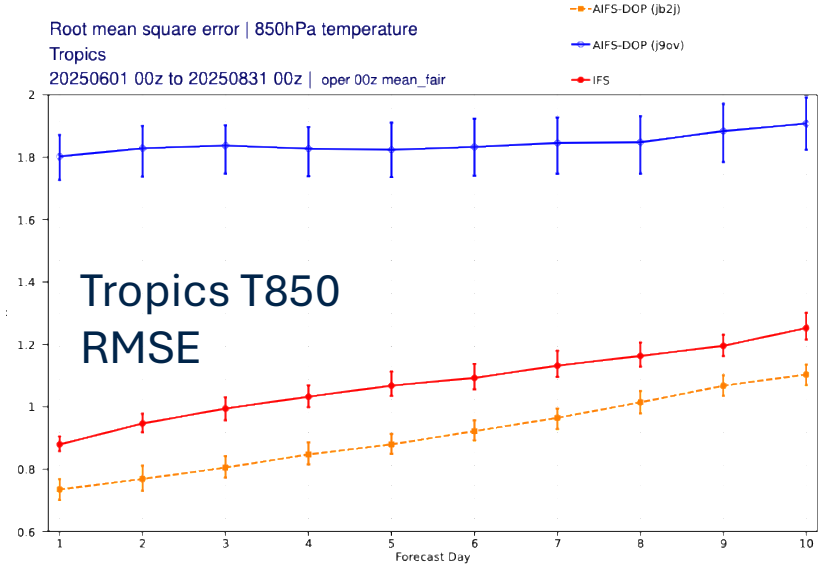


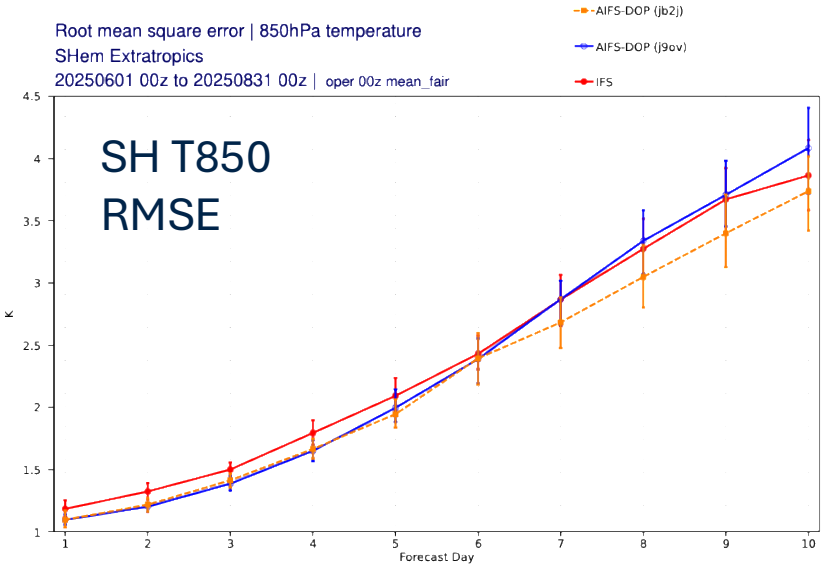
Moving to real time

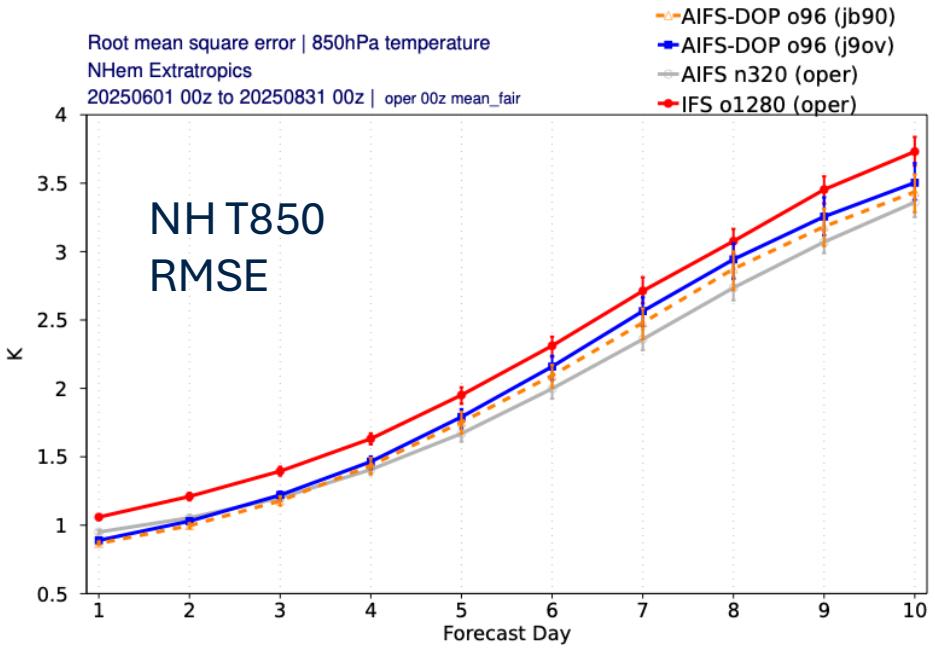
- Paper trained until end of 2020 and validated 2022
- Retraining model up to end of 2024
- Transitioning from older sensors to newer ones
  - *e.g.* HIRS missing – pickup CrIS
- Initial results look promising

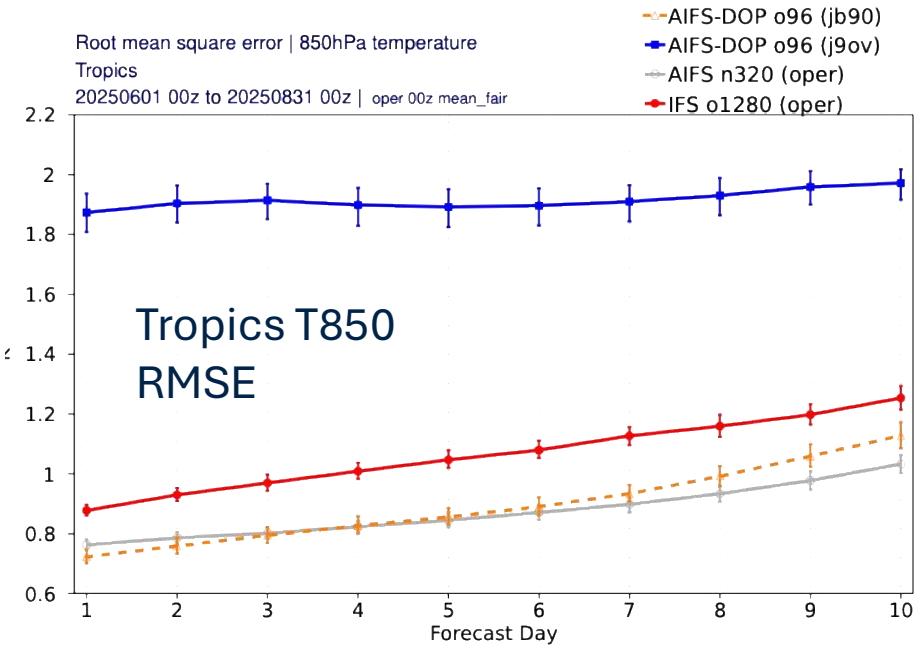


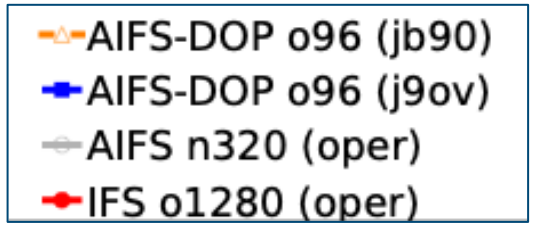


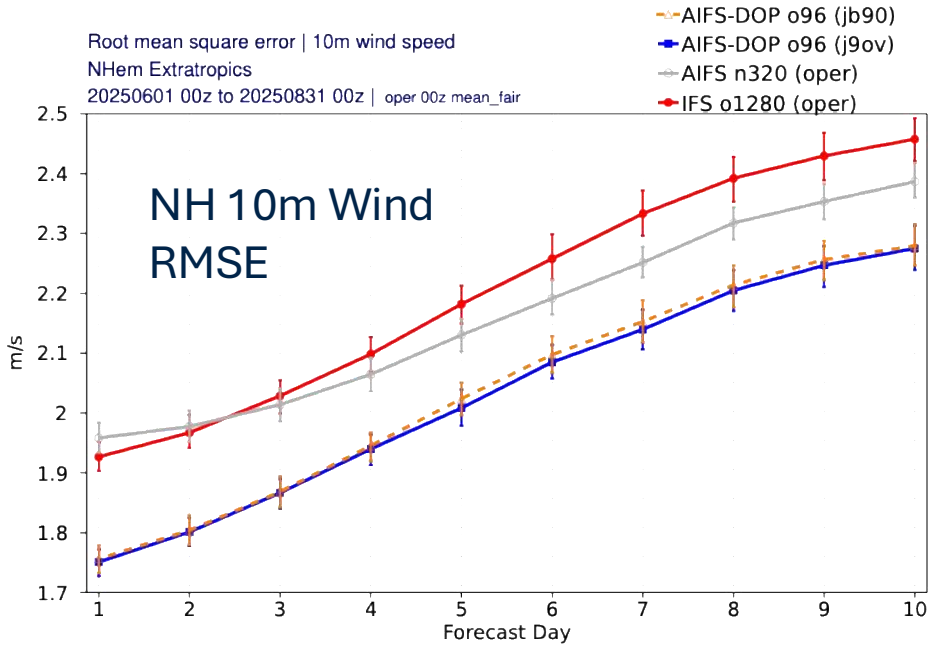


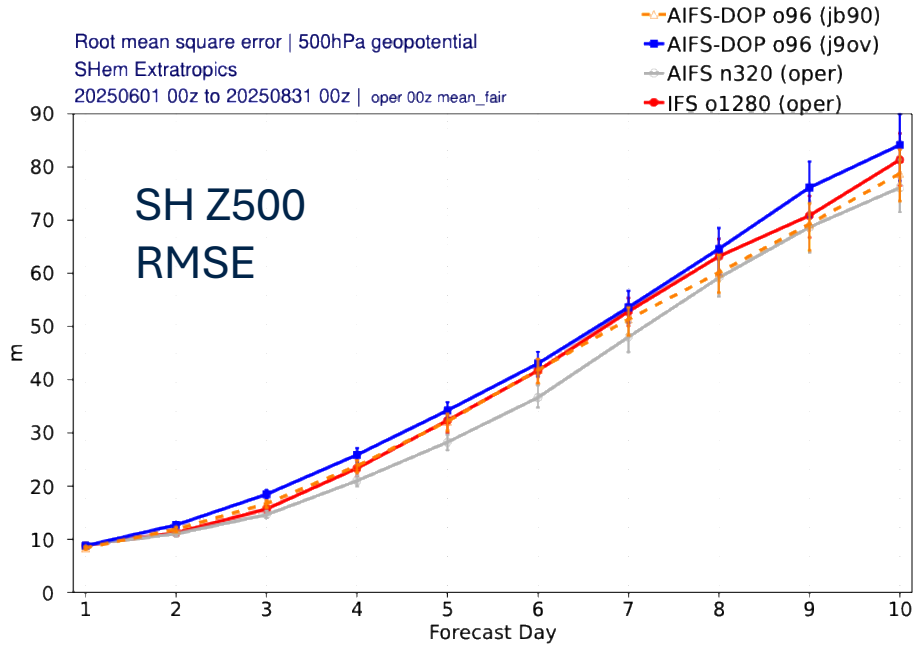



Root mean square error | 850hPa temperature
NHem Extratropics
20250601 00z to 20250831 00z | oper 00z mean_fair
AIFS-DOP o96 (jb90)
AIFS-DOP o96 (j9ov)
AIFS n320 (oper)
IFS o1280 (oper)
NH T850
RMSE
K
Forecast Day


Root mean square error | 850hPa temperature
Tropics
20250601 00z to 20250831 00z | oper 00z mean_fair
AIFS-DOP o96 (jb90)
AIFS-DOP o96 (j9ov)
AIFS n320 (oper)
IFS o1280 (oper)
Tropics T850
RMSE
Forecast Day

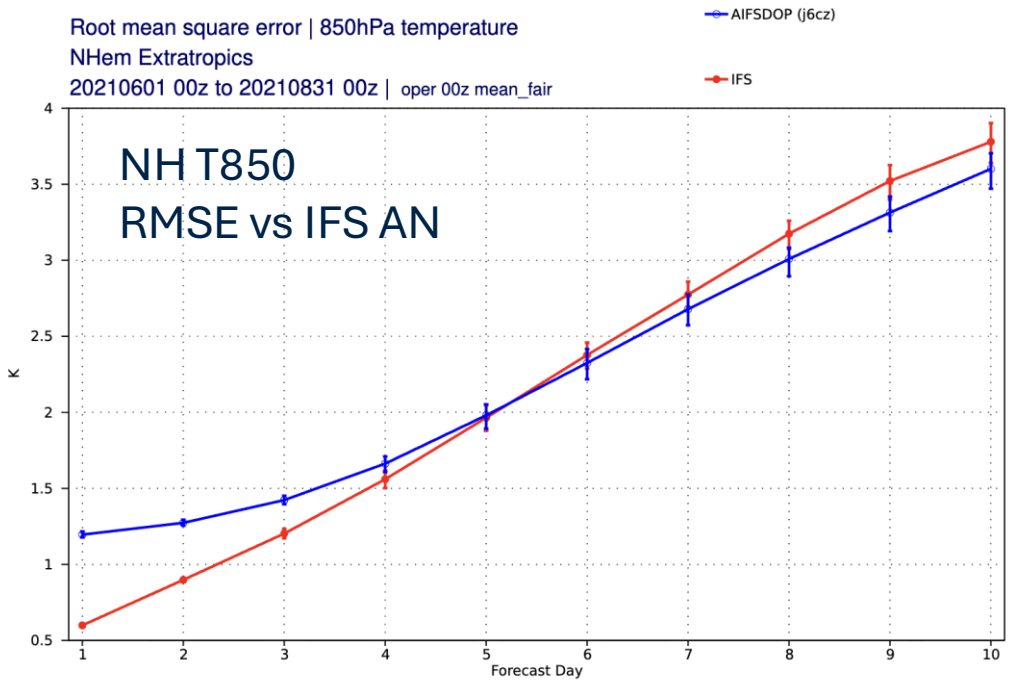

AIFS-DOP o96 (jb90)
AIFS-DOP o96 (j9ov)
AIFS n320 (oper)
IFS o1280 (oper)


Root mean square error | 10m wind speed
NHem Extratropics
20250601 00z to 20250831 00z | oper 00z mean_fair
AIFS-DOP o96 (jb90)
AIFS-DOP o96 (j9ov)
AIFS n320 (oper)
IFS o1280 (oper)
NH 10m Wind
RMSE
m/s
Forecast Day


Root mean square error | 500hPa geopotential
SHem Extratropics
20250601 00z to 20250831 00z | oper 00z mean_fair
AIFS-DOP o96 (jb90)
AIFS-DOP o96 (j9ov)
AIFS n320 (oper)
IFS o1280 (oper)
SH Z500
RMSE
m
Forecast Day

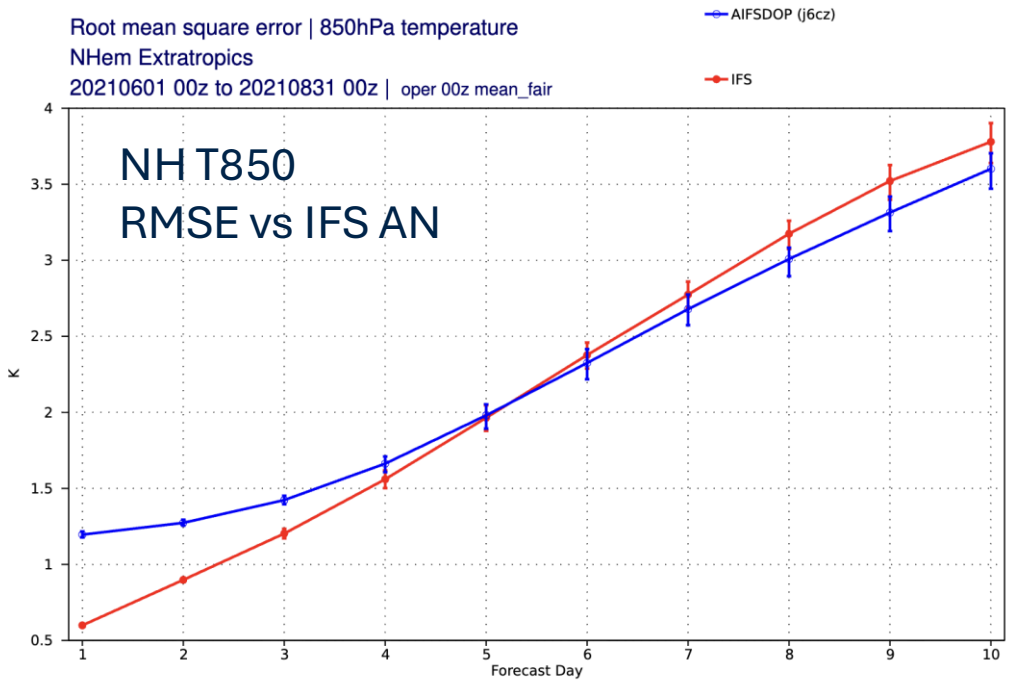
Lies, damned lies and Analysis verification…

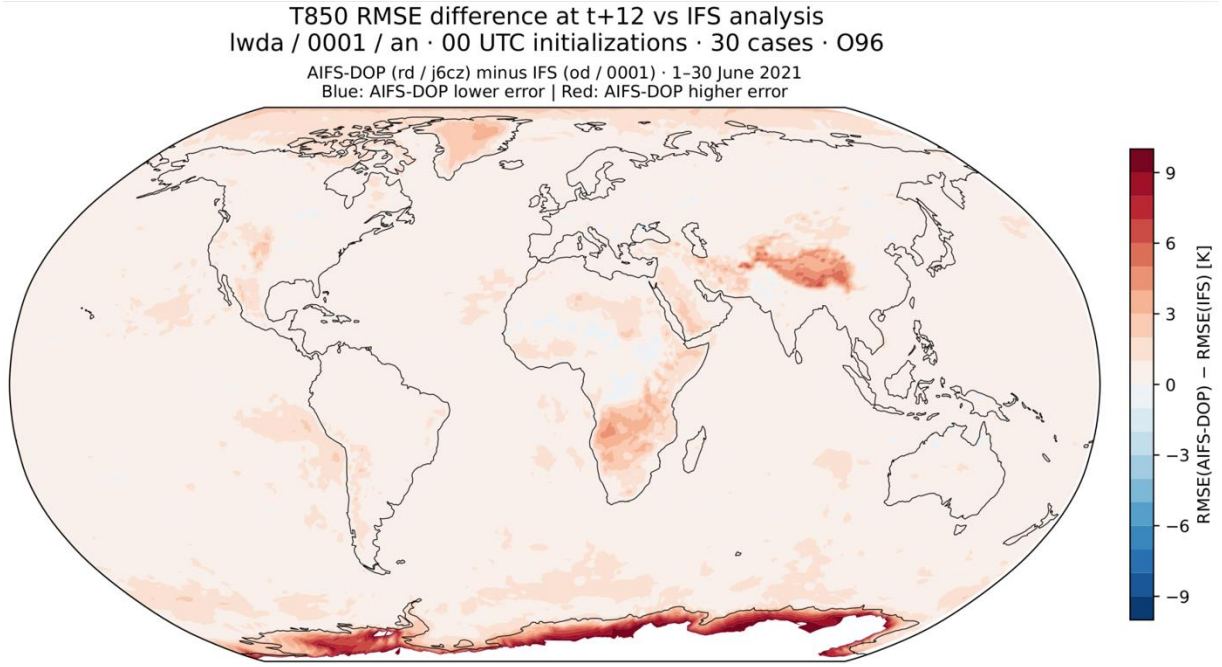
- Work started to understand analysis scores
  - Care needed when making comparisons!



Lies, damned lies and Analysis verification…

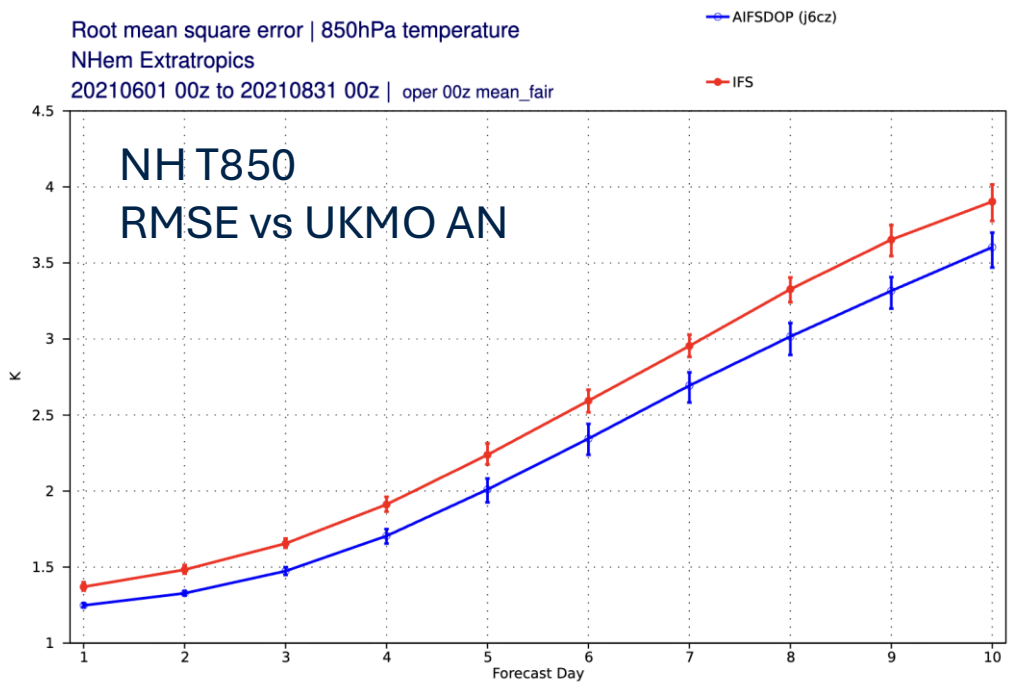
- Work started to understand analysis scores
  - Care needed when making comparisons!

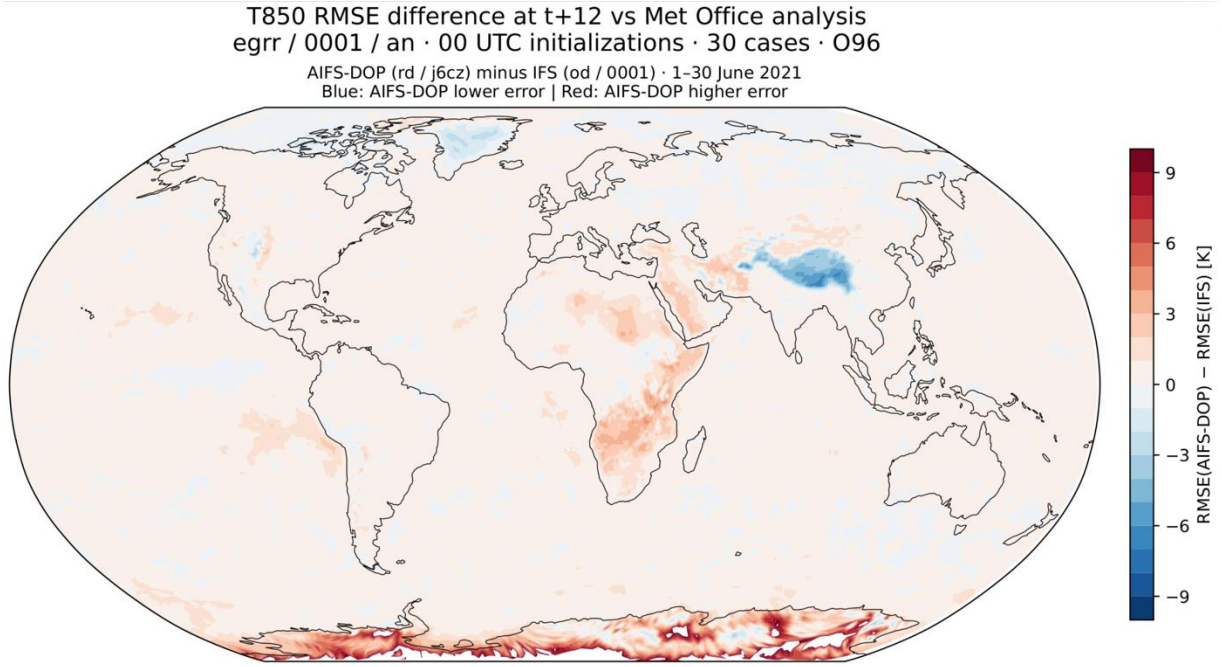




Lies, damned lies and Analysis verification…

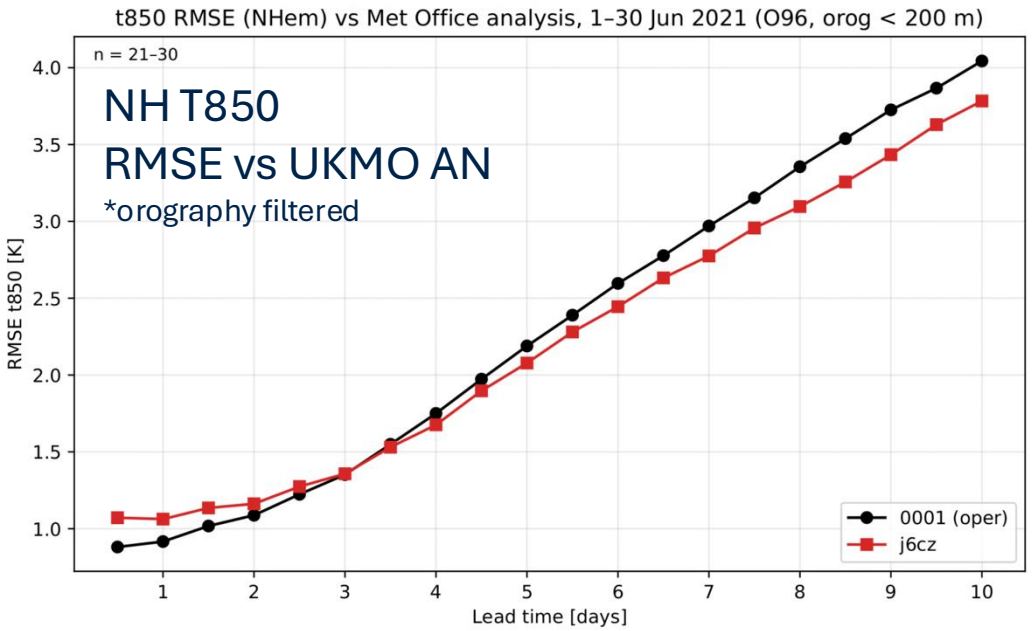
- Work started to understand analysis scores
  - Care needed when making comparisons!





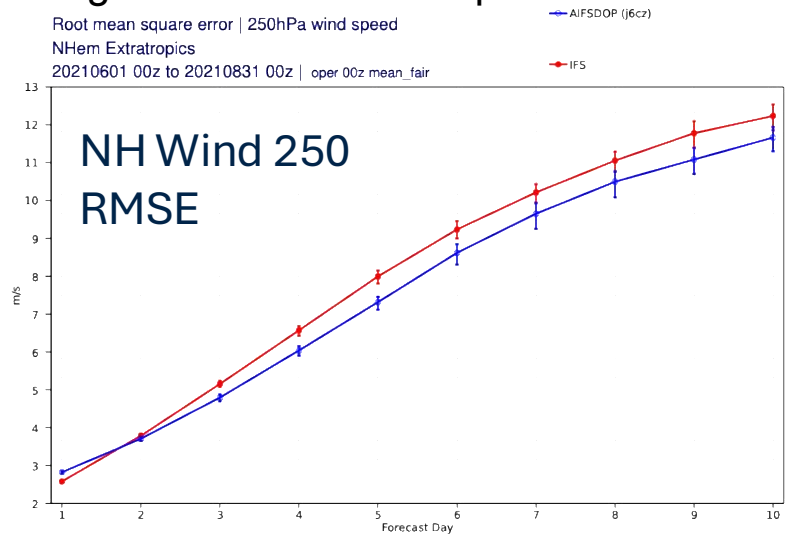
Lies, damned lies and Analysis verification…

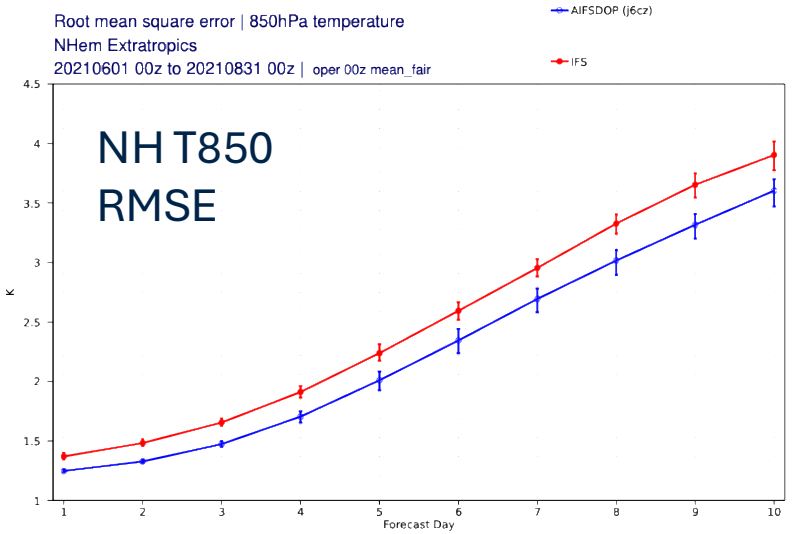
- Work started to understand analysis scores
  - Care needed when making comparisons!

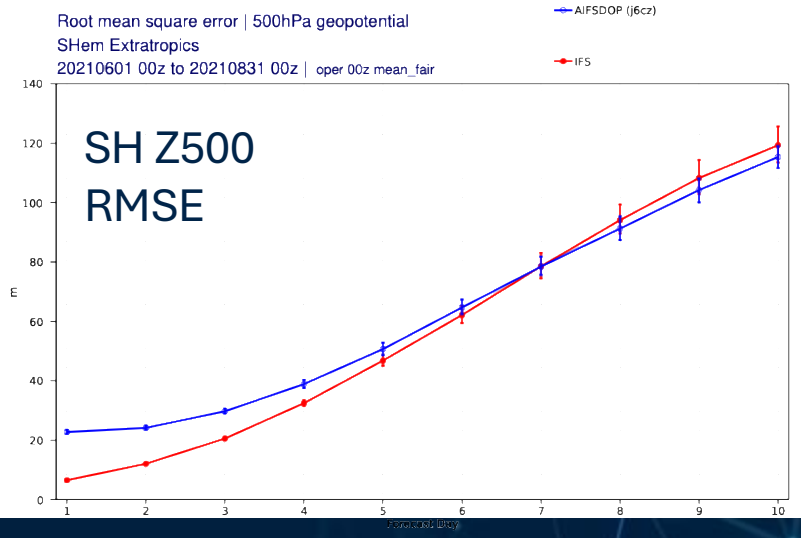


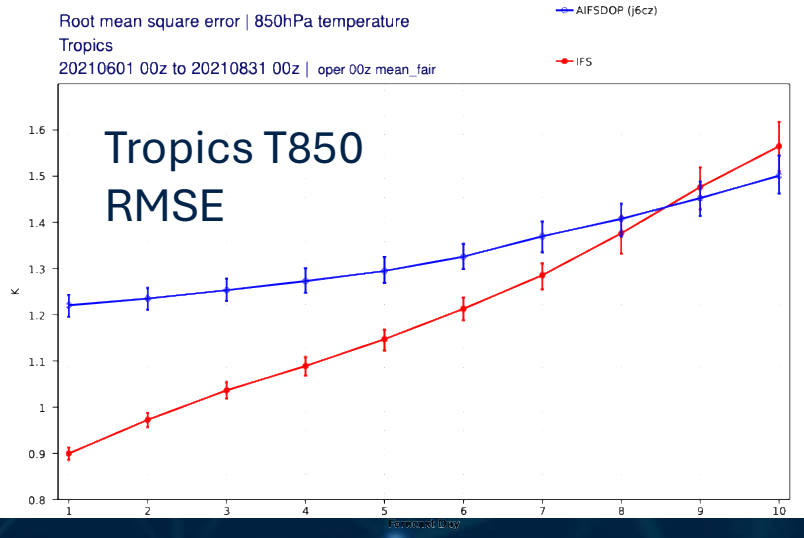
Comparisons to Analysis?

- Work started to understand analysis scores
  - Very helpful to diagnose issues and improve results!

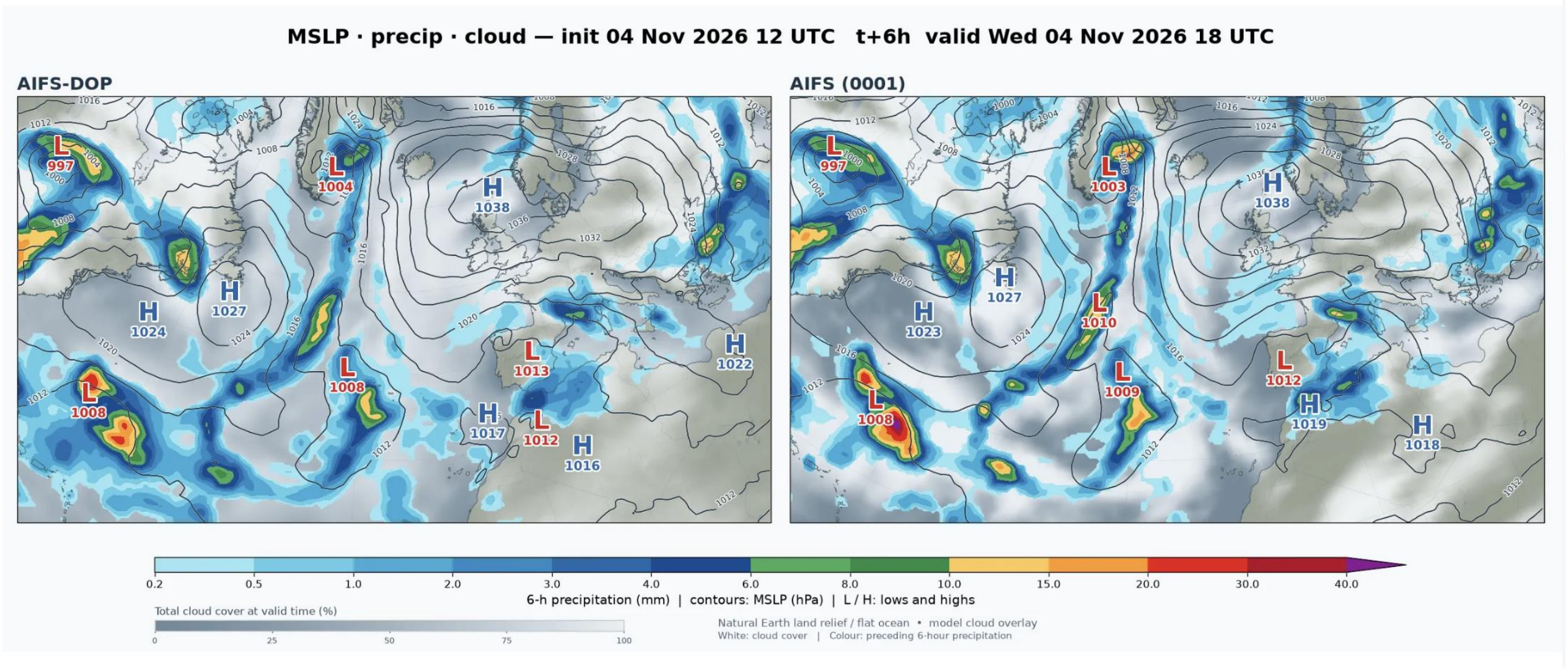




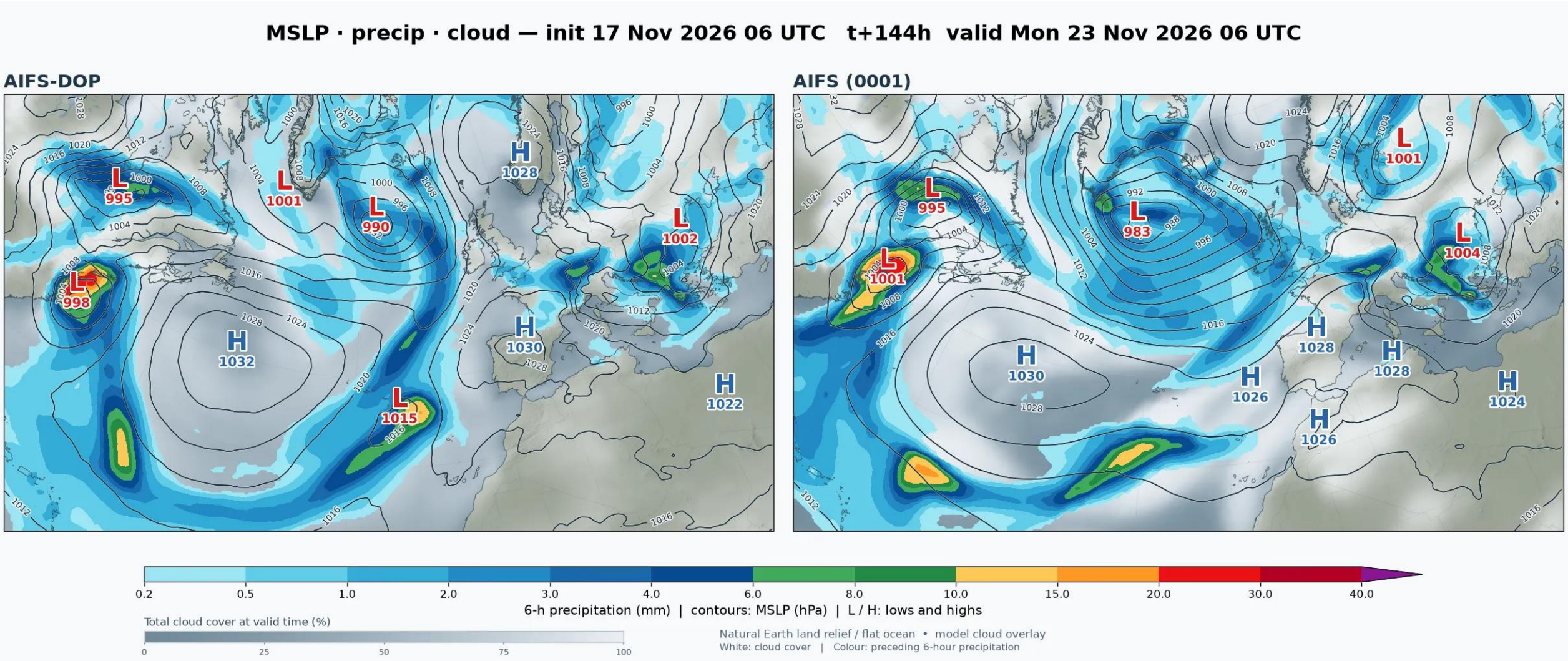




Real Time


MSLP · precip · cloud — init 04 Nov 2026 12 UTC t+6h valid Wed 04 Nov 2026 18 UTC
AIFS-DOP
AIFS (0001)
0.2
0.5
1.0
2.0
3.0
4.0
6.0
8.0
10.0
15.0
20.0
30.0
40.0
6-h precipitation (mm) | contours: MSLP (hPa) | L / H: lows and highs
Total cloud cover at valid time (%)
0
25
50
75
100
Natural Earth land relief / flat ocean • model cloud overlay
White: cloud cover | Colour: preceding 6-hour precipitation

Real Time


MSLP · precip · cloud — init 17 Nov 2026 06 UTC t+144h valid Mon 23 Nov 2026 06 UTC
AIFS-DOP
AIFS (0001)
0.2 0.5 1.0 2.0 3.0 4.0 6.0 8.0 10.0 15.0 20.0 30.0 40.0
6-h precipitation (mm) | contours: MSLP (hPa) | L / H: lows and highs
Total cloud cover at valid time (%)
0 25 50 75 100
Natural Earth land relief / flat ocean • model cloud overlay
White: cloud cover | Colour: preceding 6-hour precipitation

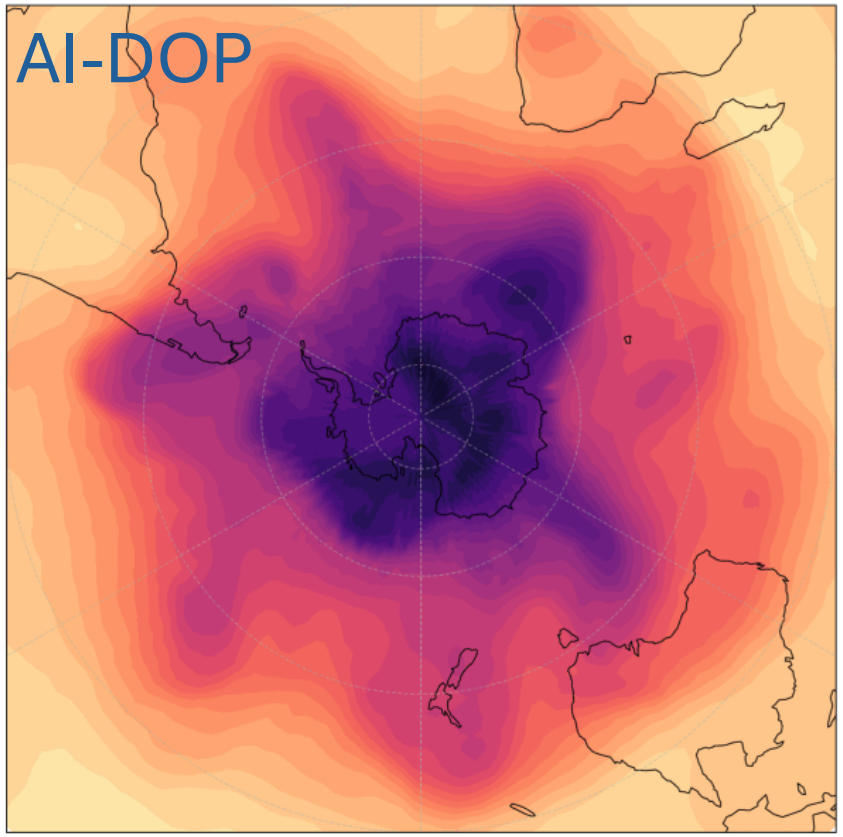
Prospects for using DOP for reanalysis



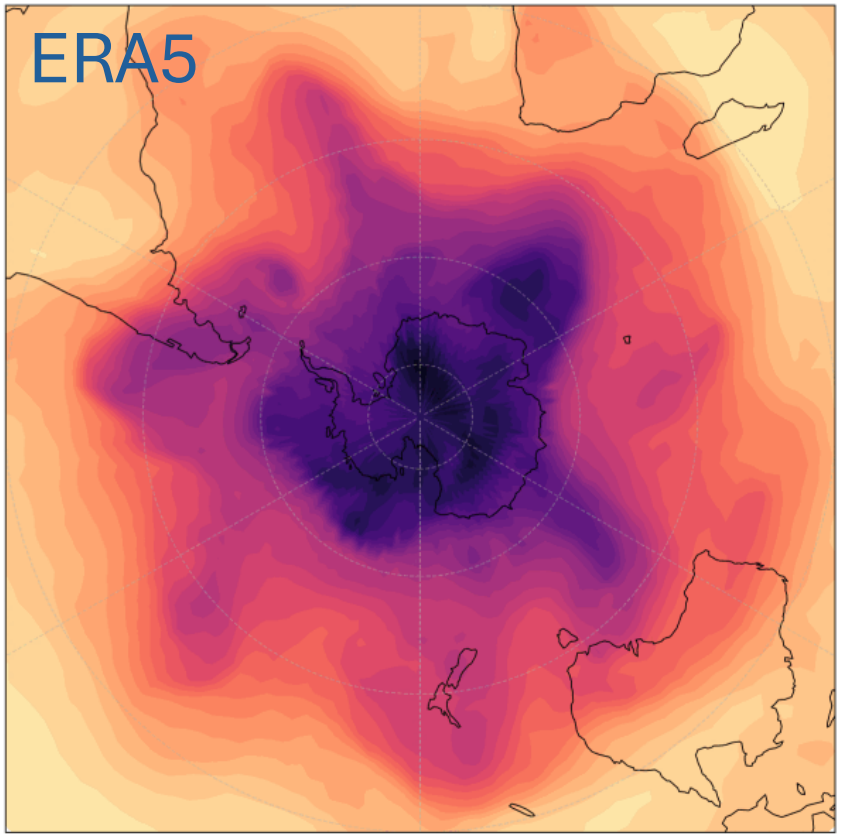
500hPa Temperature

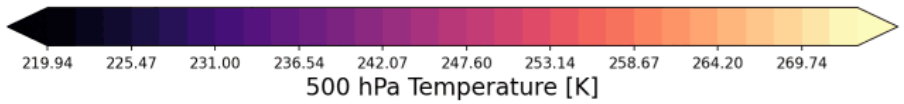
SH Extratropical T500 Evolution
Valid Time: 2017-09-03 00:00 UTC

AIFS-DOP



ERA5





* Sequence of analyses

Could Direct Observation Prediction be used to generate a reanalysis?

- **Can DOP produce dense, physically coherent fields from sparse observations?**
  - Data assimilation systems use a physical model and covariance statistics to produce dense, physically coherent fields from sparse observations
  - Can the representations of the physical Earth System learned during training perform a similar role?
- Overlapping training and test period, is this a problem?
- Prototype presented here: state estimates from t+6h DOP forecasts, every 6 hours at O96 resolution
- **Very fast** to produce: inference is cheap + embarrassingly parallel

Takes < 1 day to produce a 40-year reanalysis!

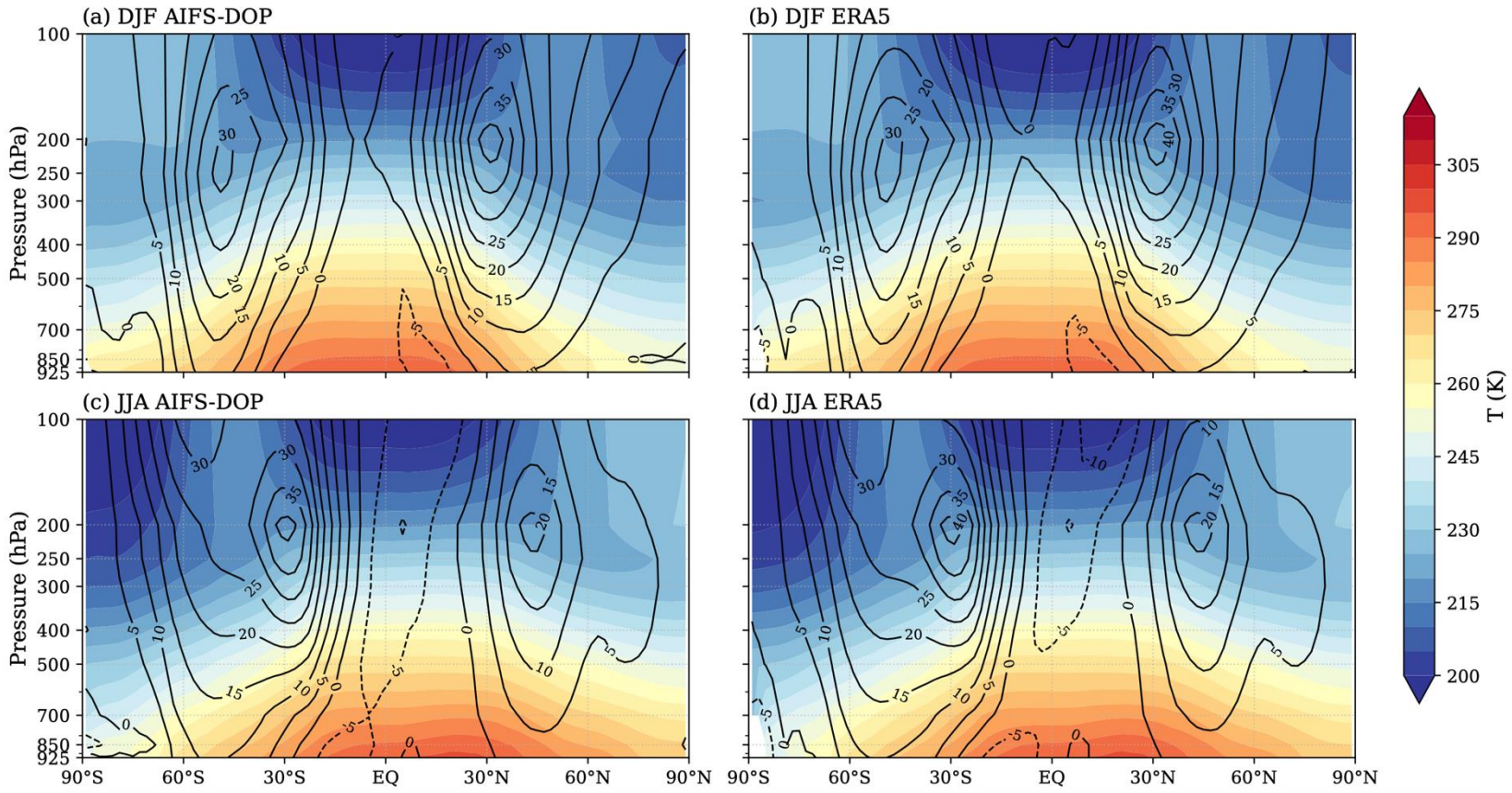
Zonal mean state

DOP

ERA5

DJF

JJA

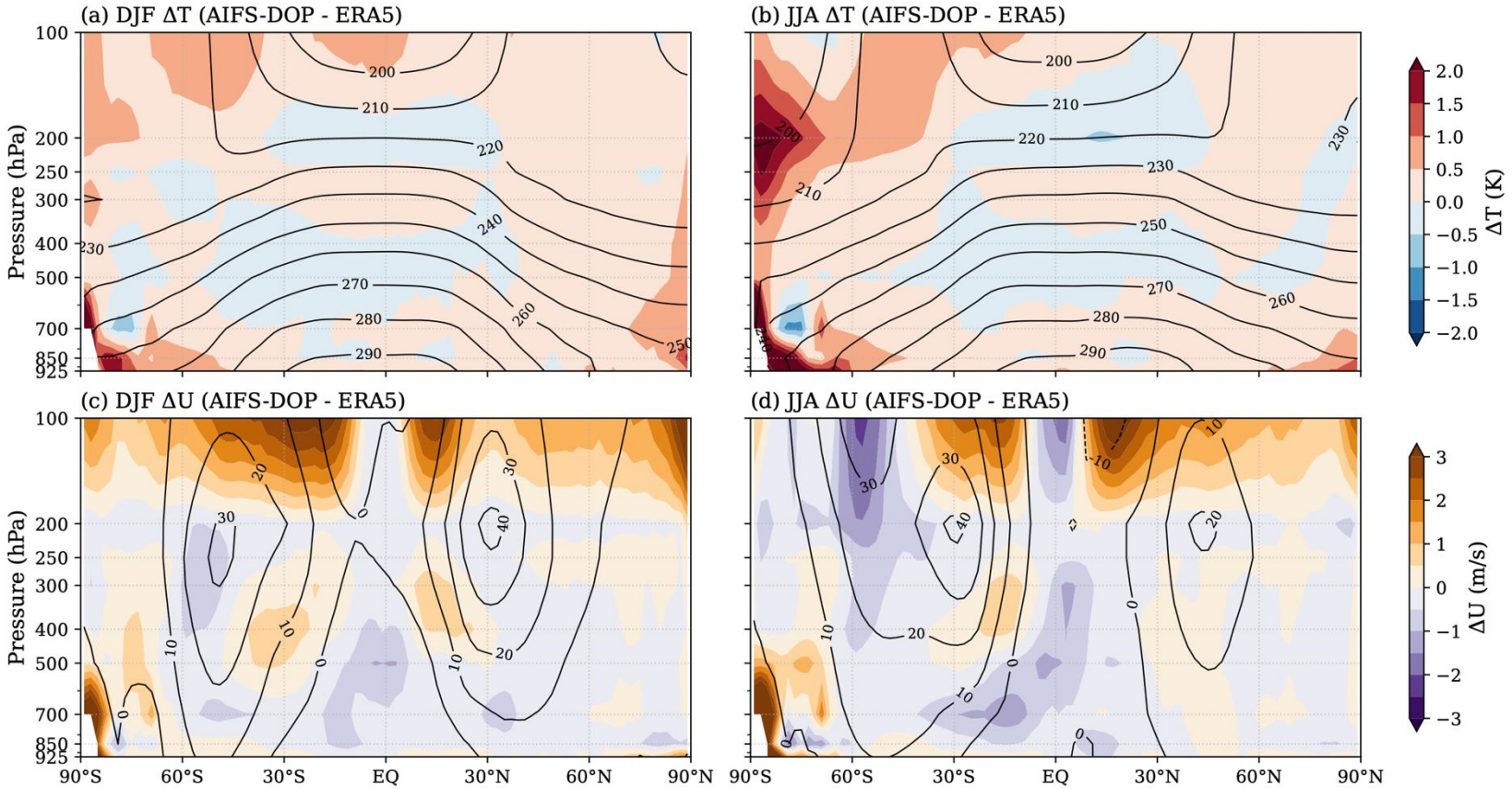


Zonal mean u-wind contours

Mean difference v ERA5

Temp diff

Wind diff

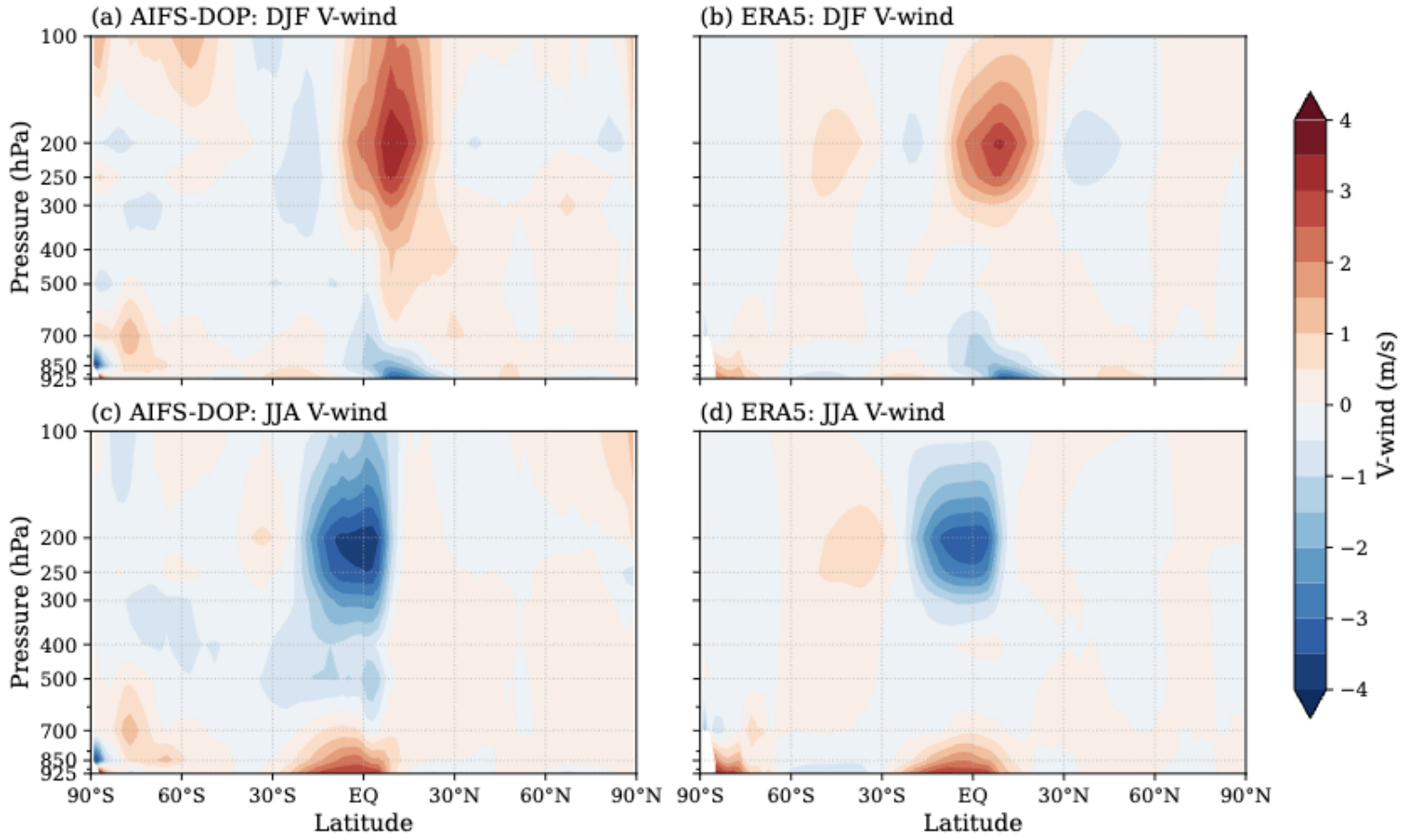


DJF

JJA

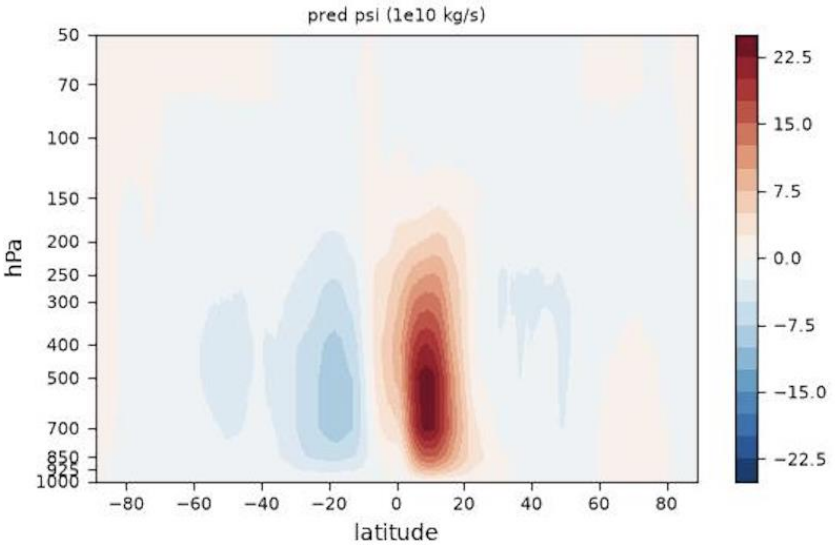
Zonal mean u-wind contours

Issues with the meridional wind

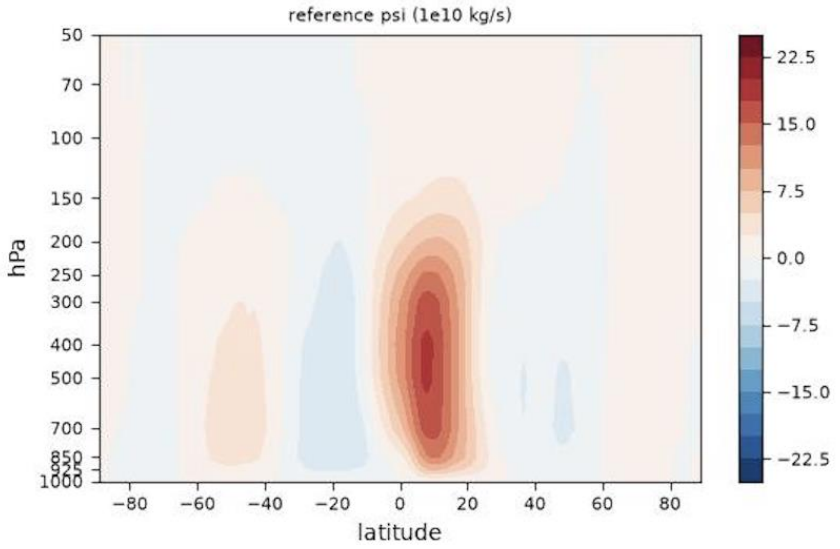


Issues with the meridional wind

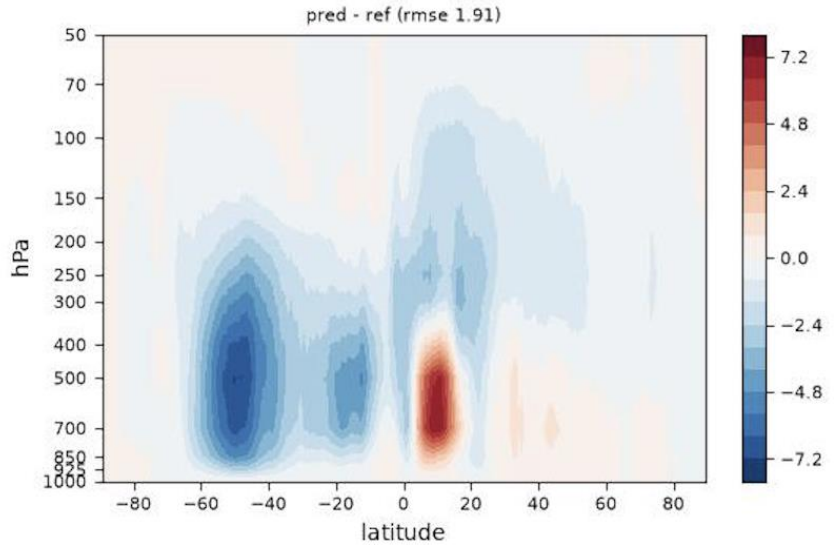
DOP



IFS

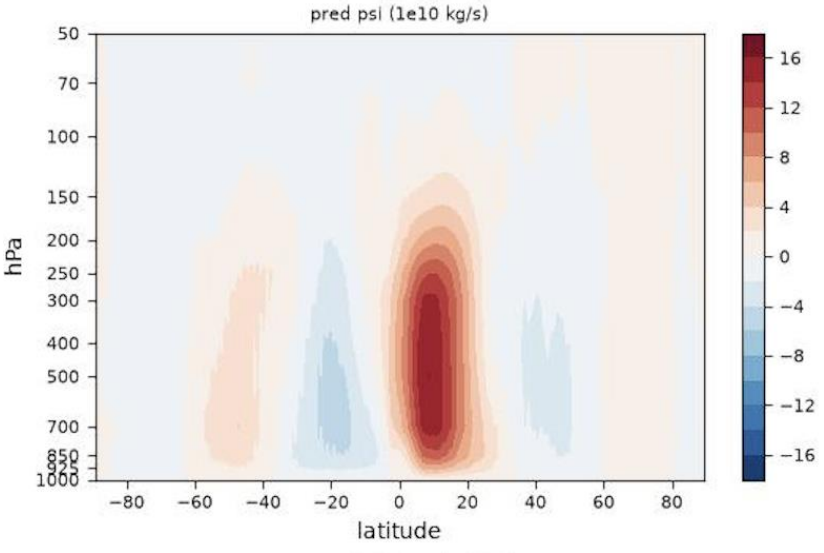


Error = 1.9

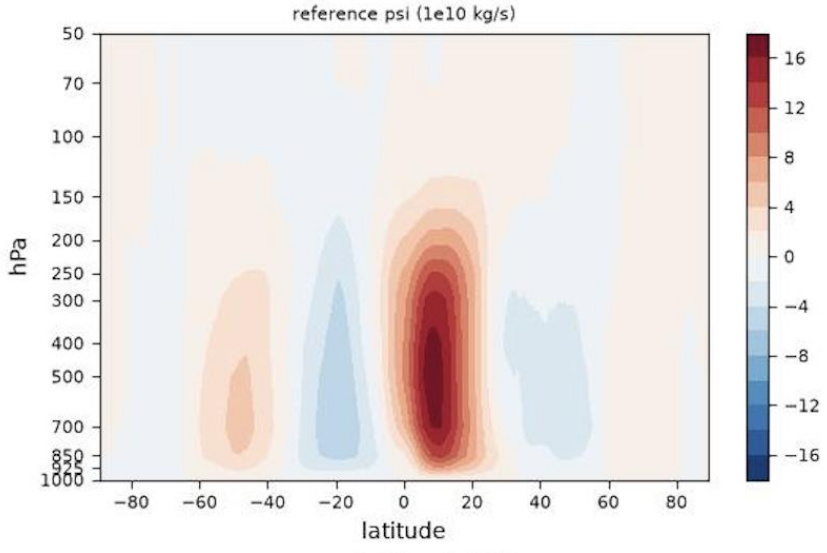


Issues with the meridional wind

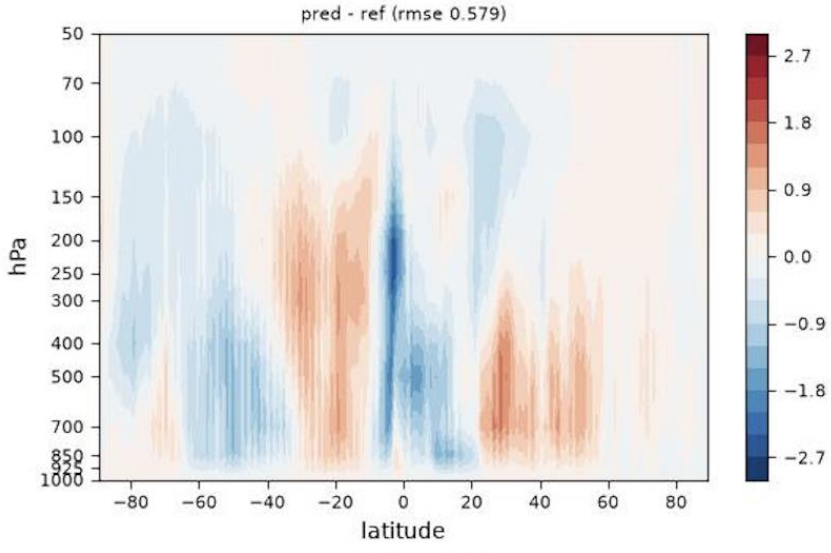
DOP



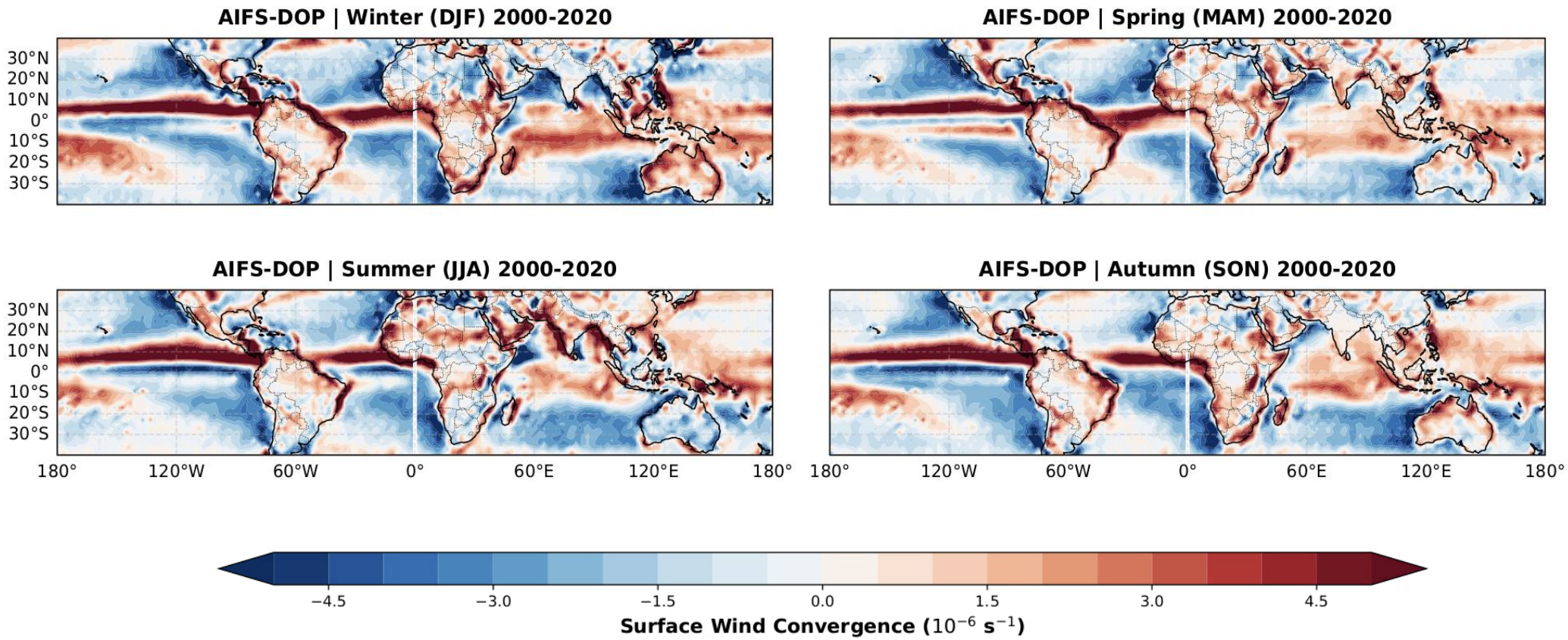
IFS



Error = 0.57

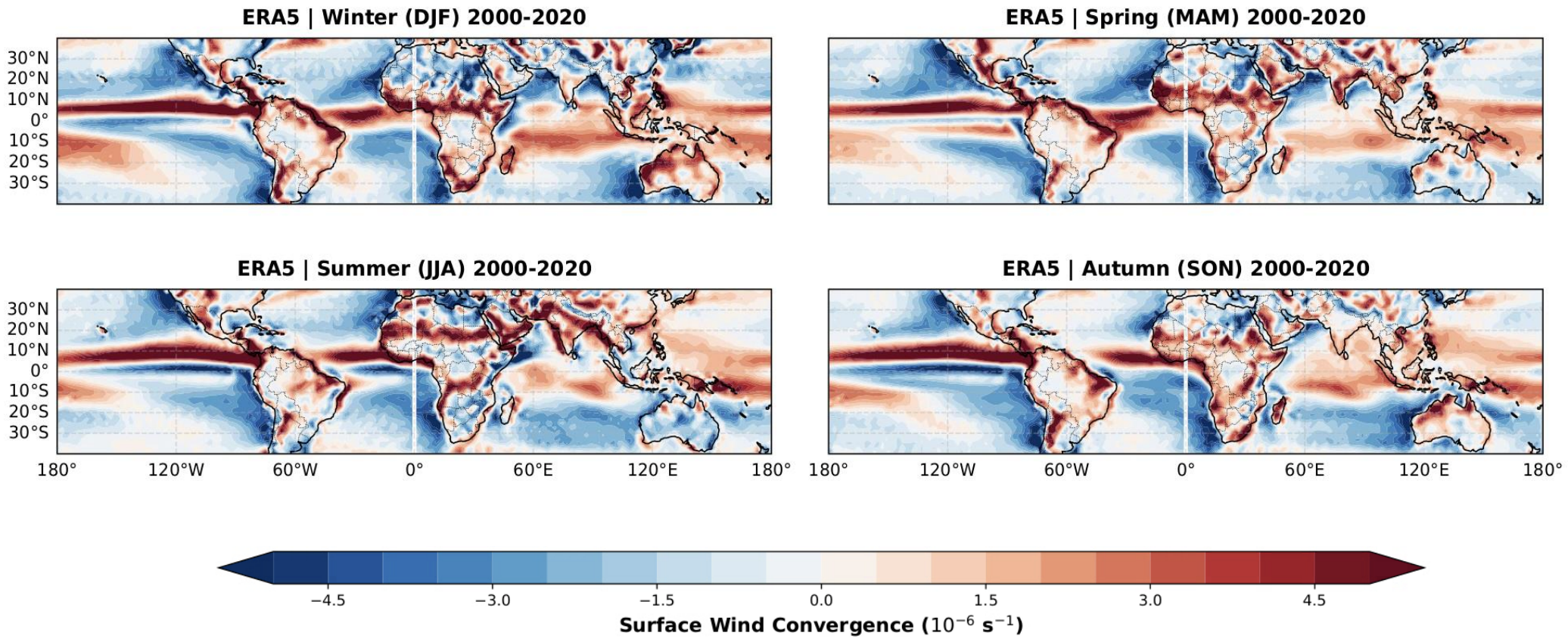


ITCZ: DOP



* No scatterometer data used in DOP yet

ITCZ: ERA5



* No scatterometer data used in DOP yet

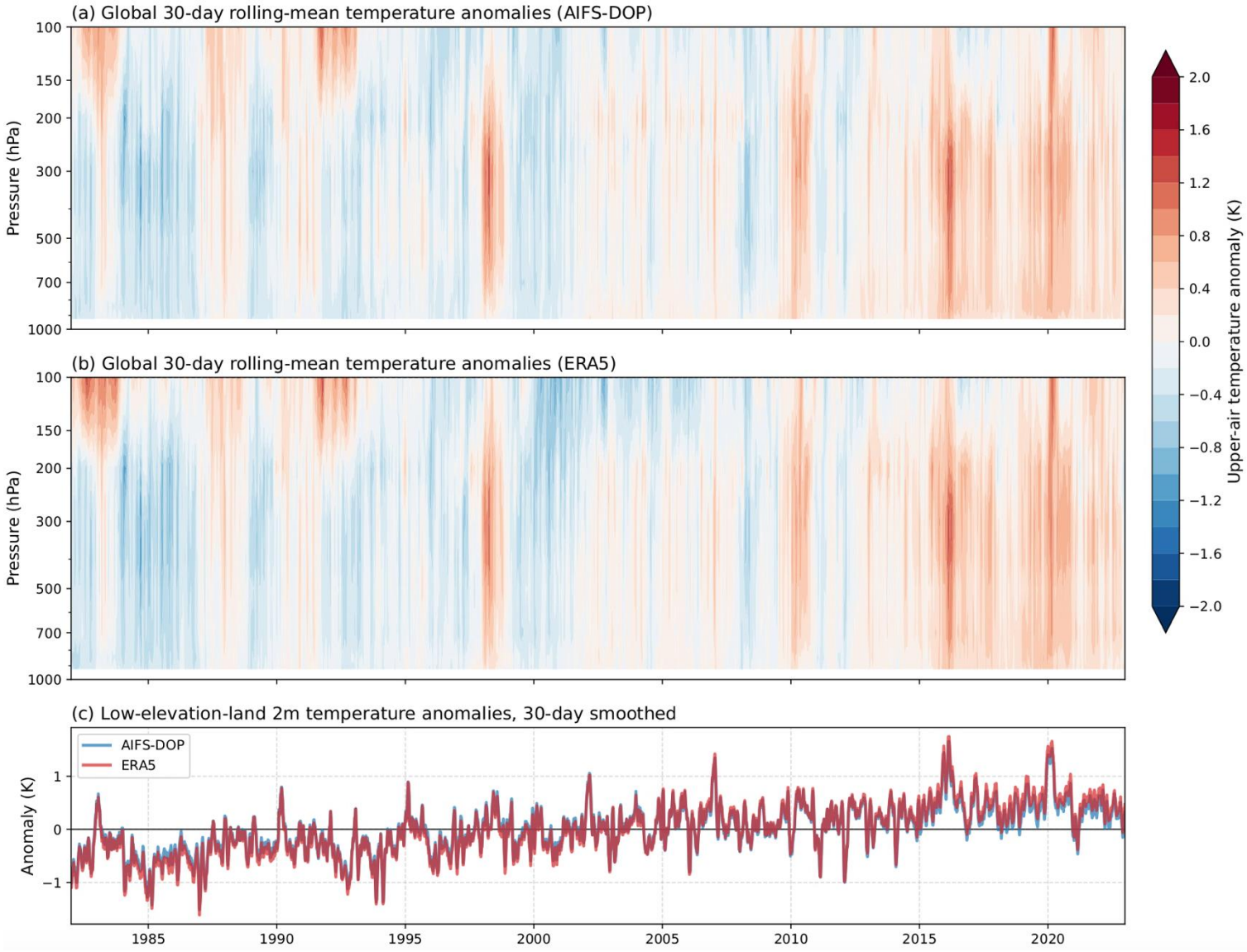
Temperature anomalies

DOP

Upper air

ERA5

2m temp
low-elevation land



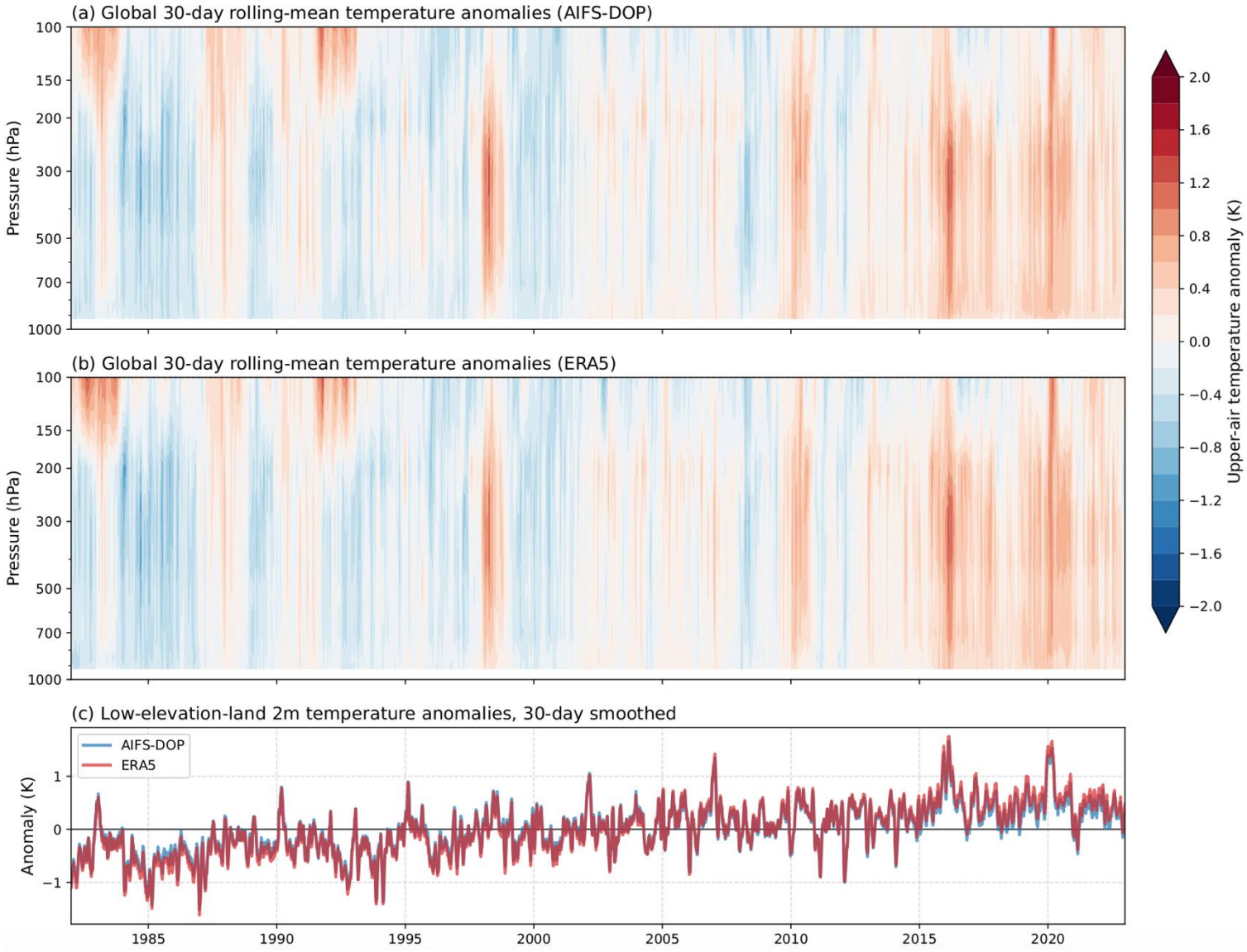
Temperature anomalies

DOP

Upper air

ERA5.1

2m temp
low-elevation land

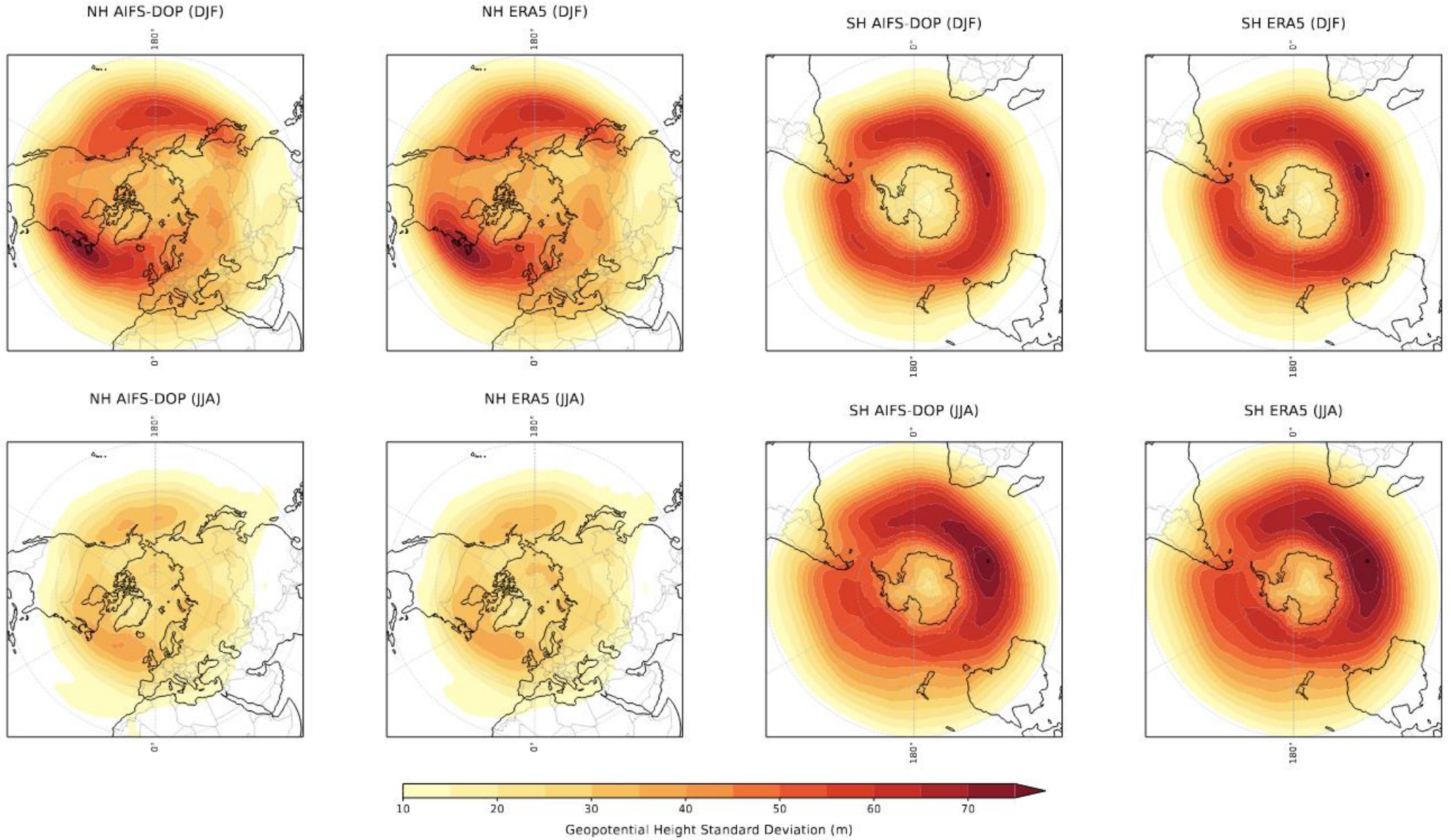


Variability - storm tracks

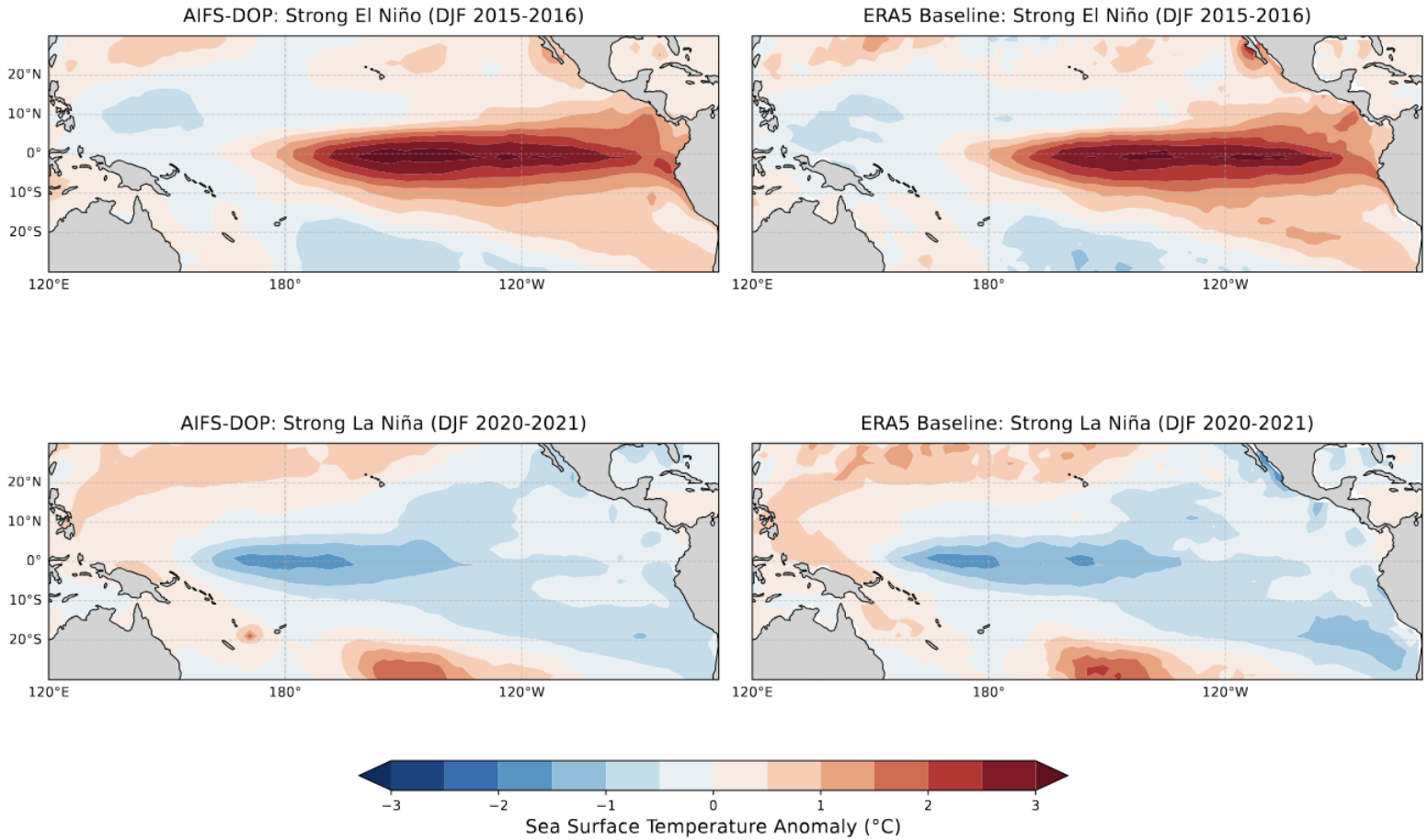
Standard deviation of band-pass filtered z500

DJF

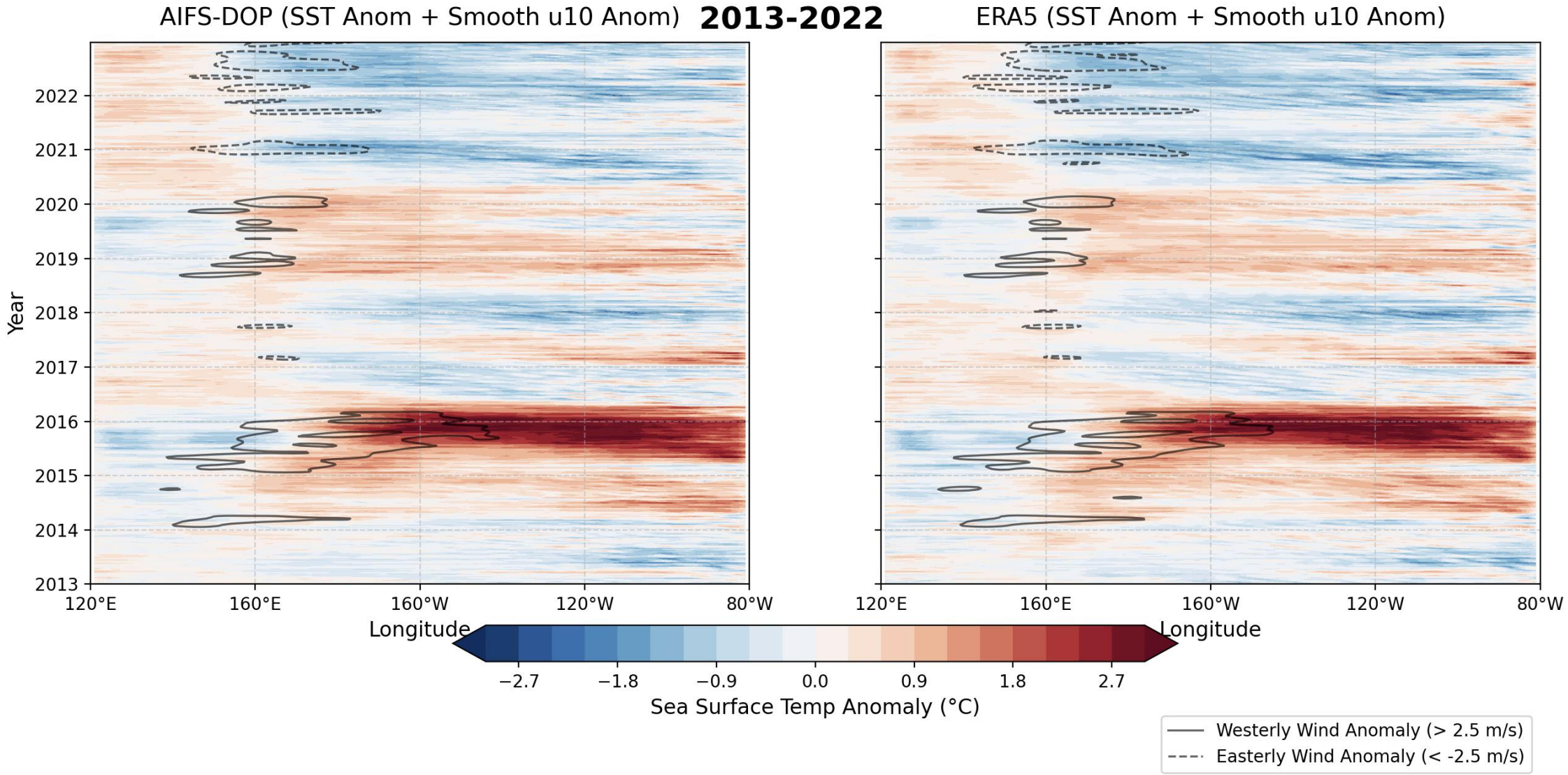
JJA



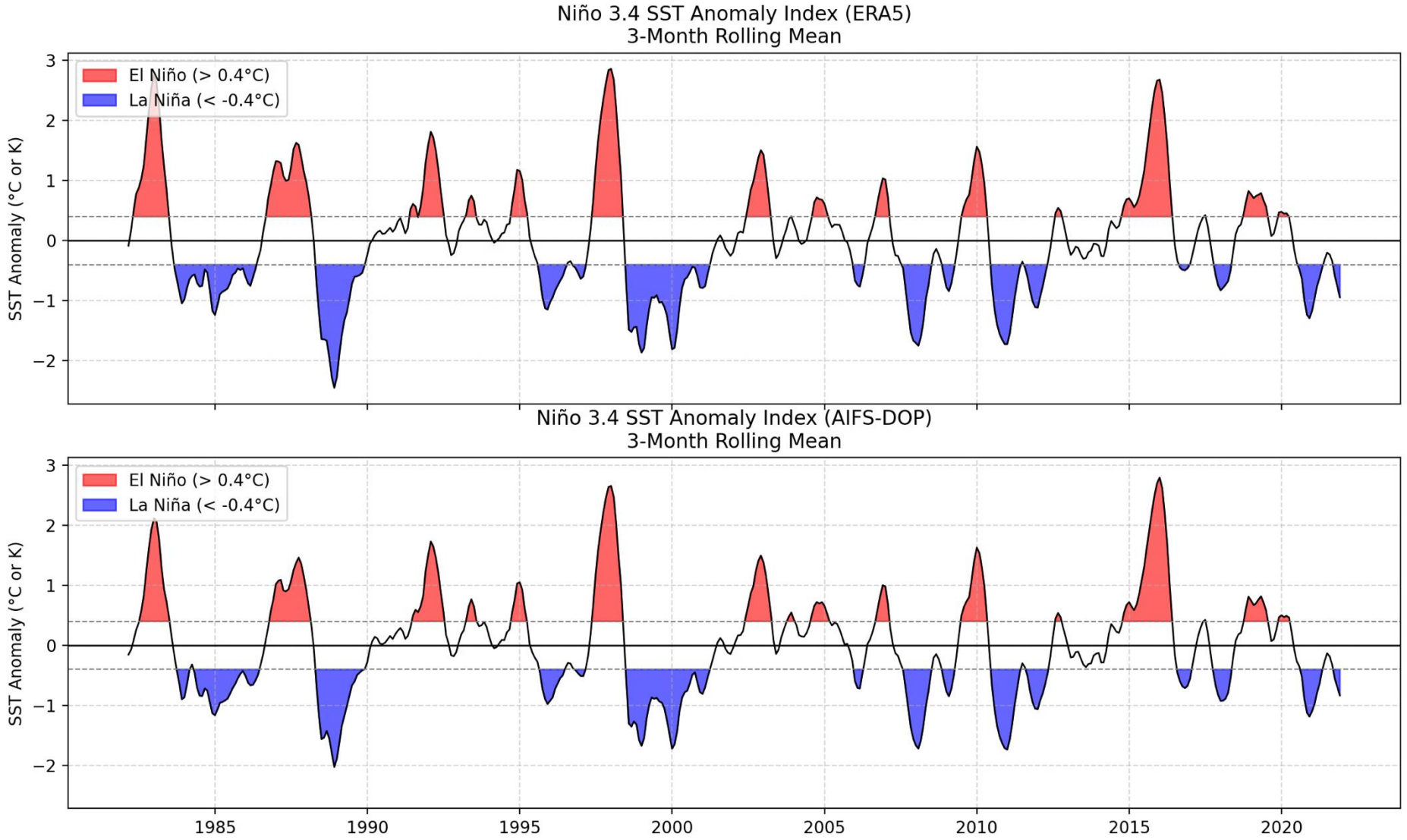
ENSO anomalies



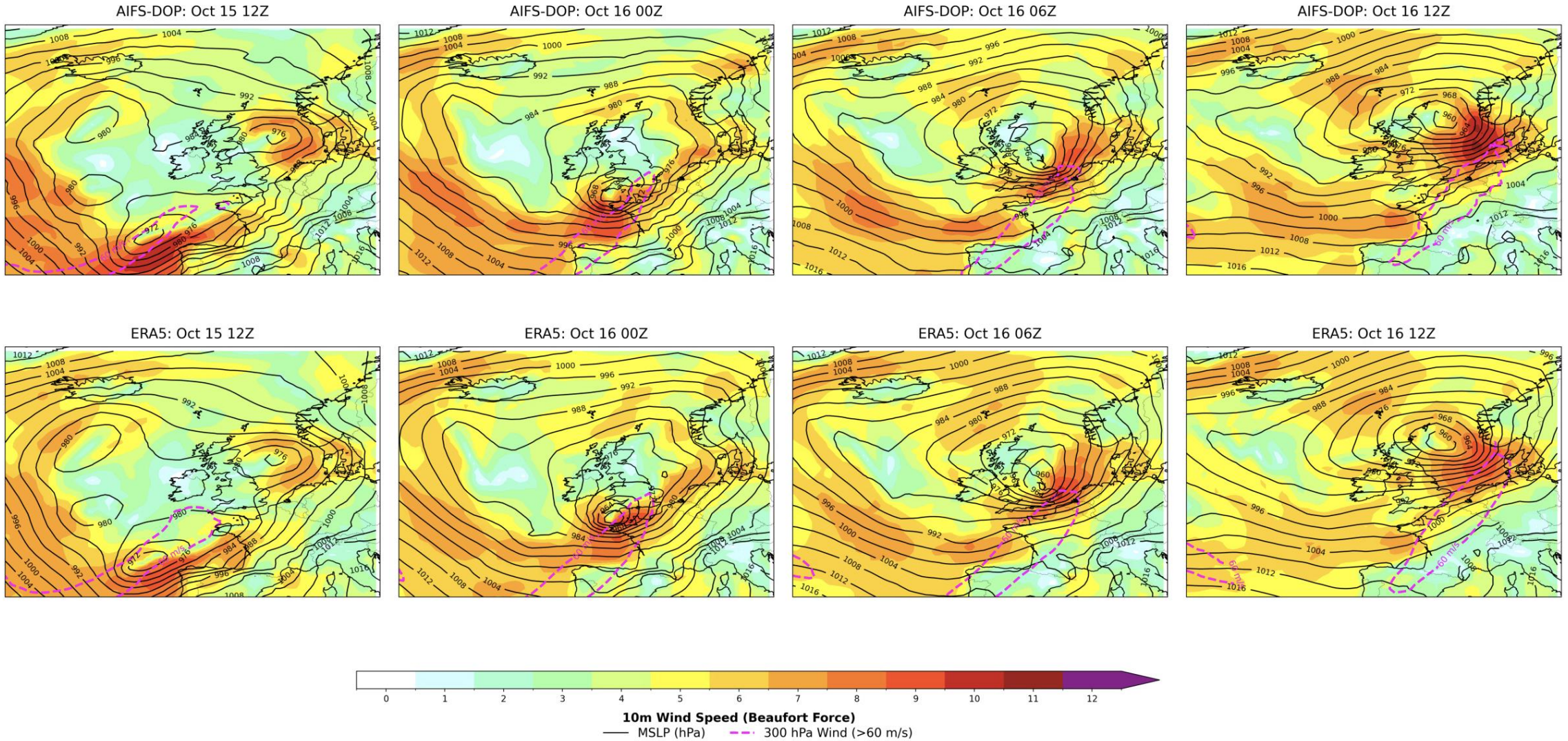
Nino 3.4 SST anomalies



Nino 3.4 SST anomalies



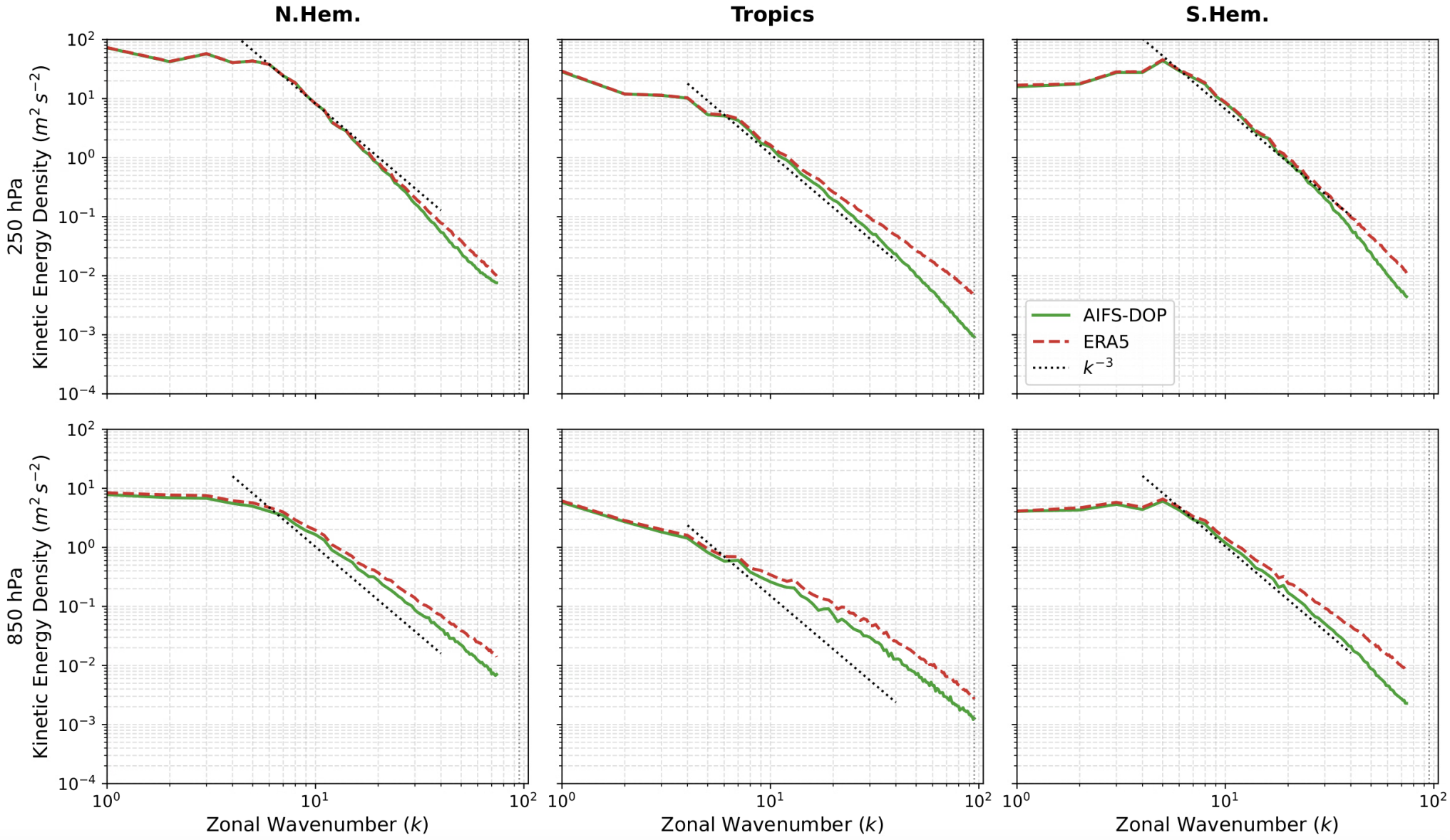
Case study: 1987 extreme wind storm



Kinetic Energy Spectra

KE 250hPa

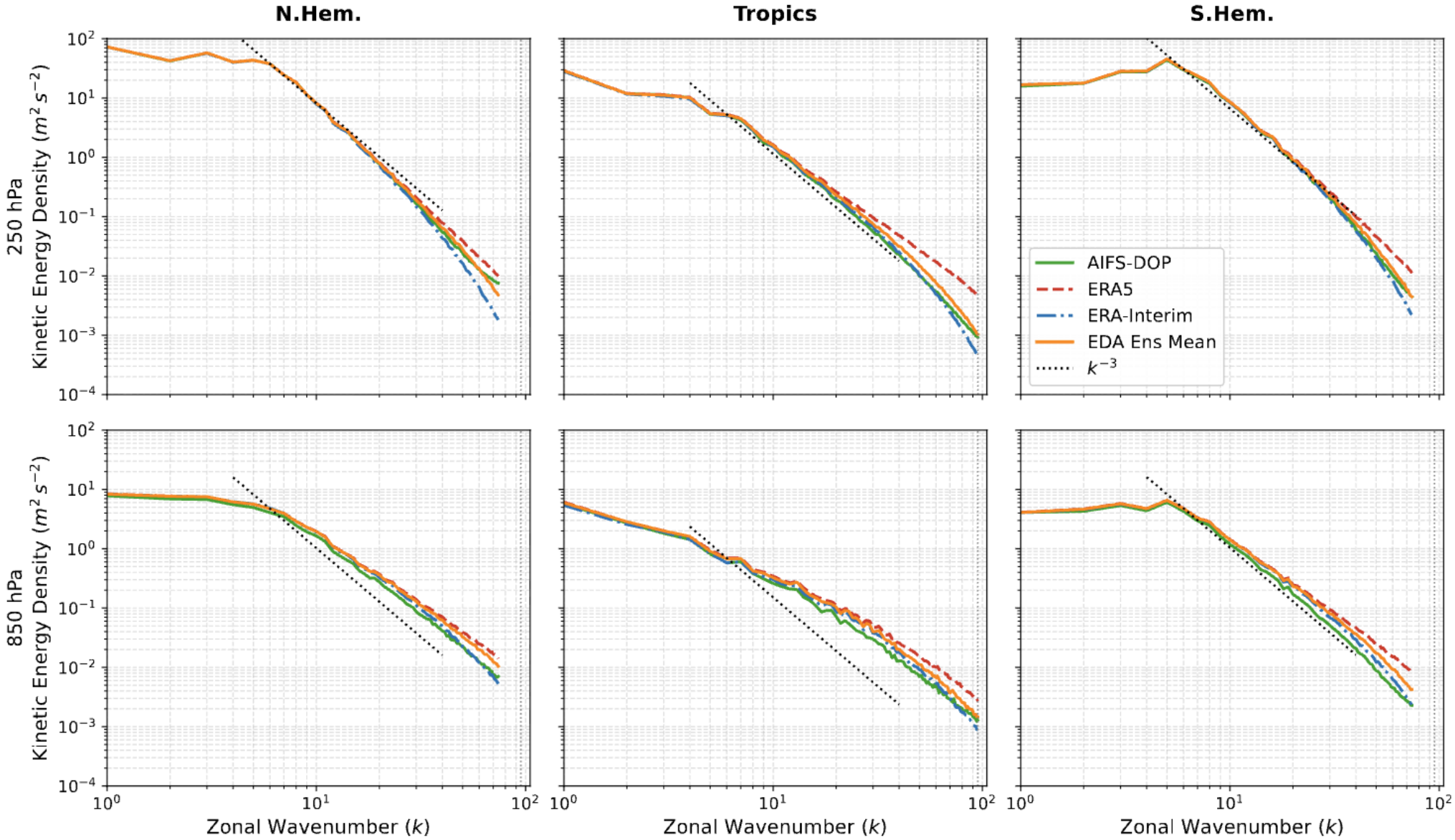
KE 500hPa



Kinetic Energy Spectra

KE 250hPa

KE 500hPa

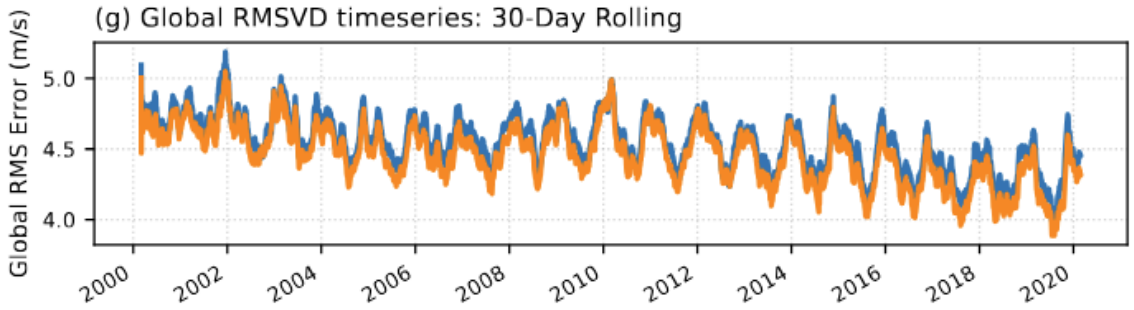


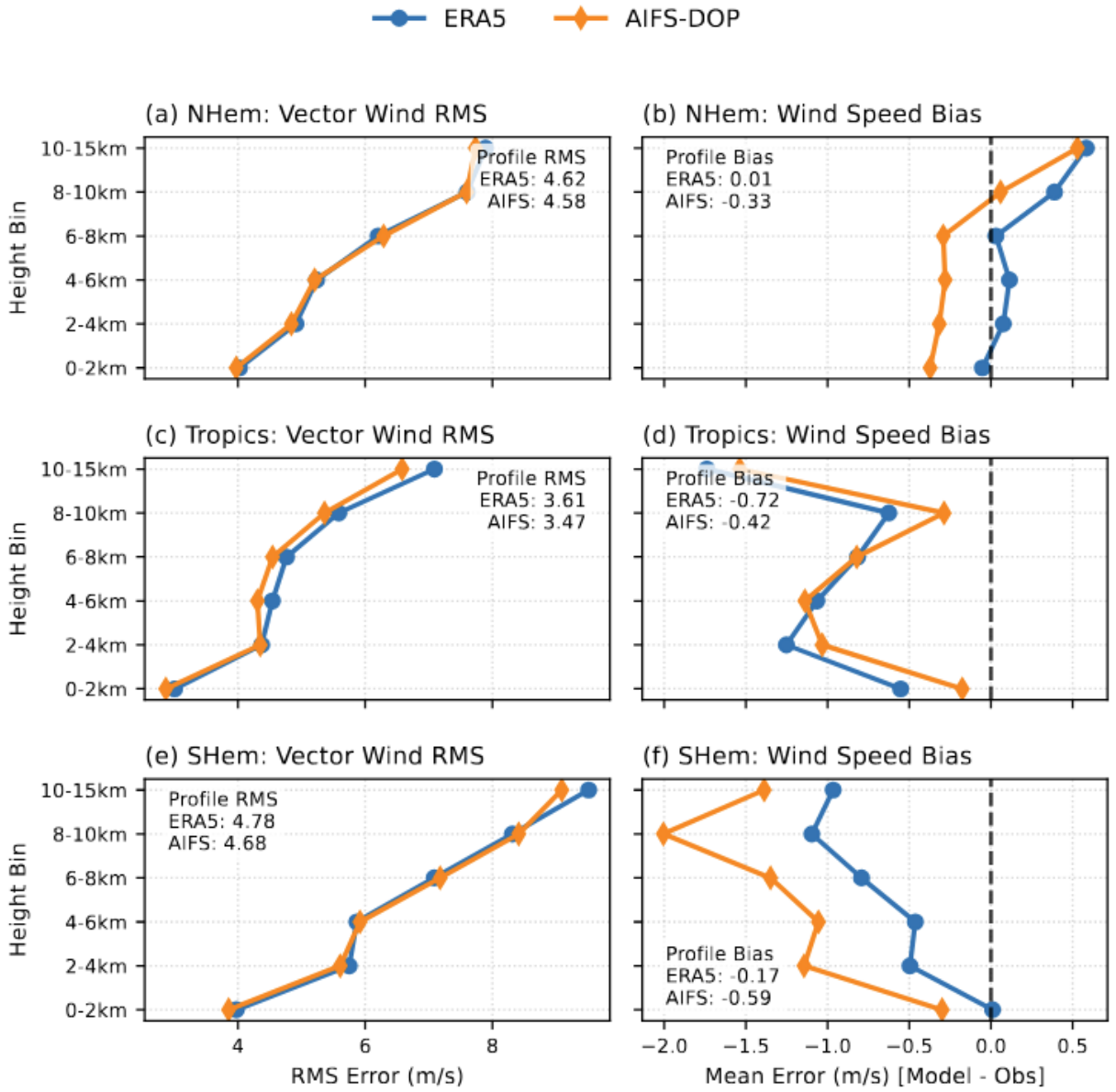
More smoothing in tropics

Evaluation against held-out observations: MISR AMVs

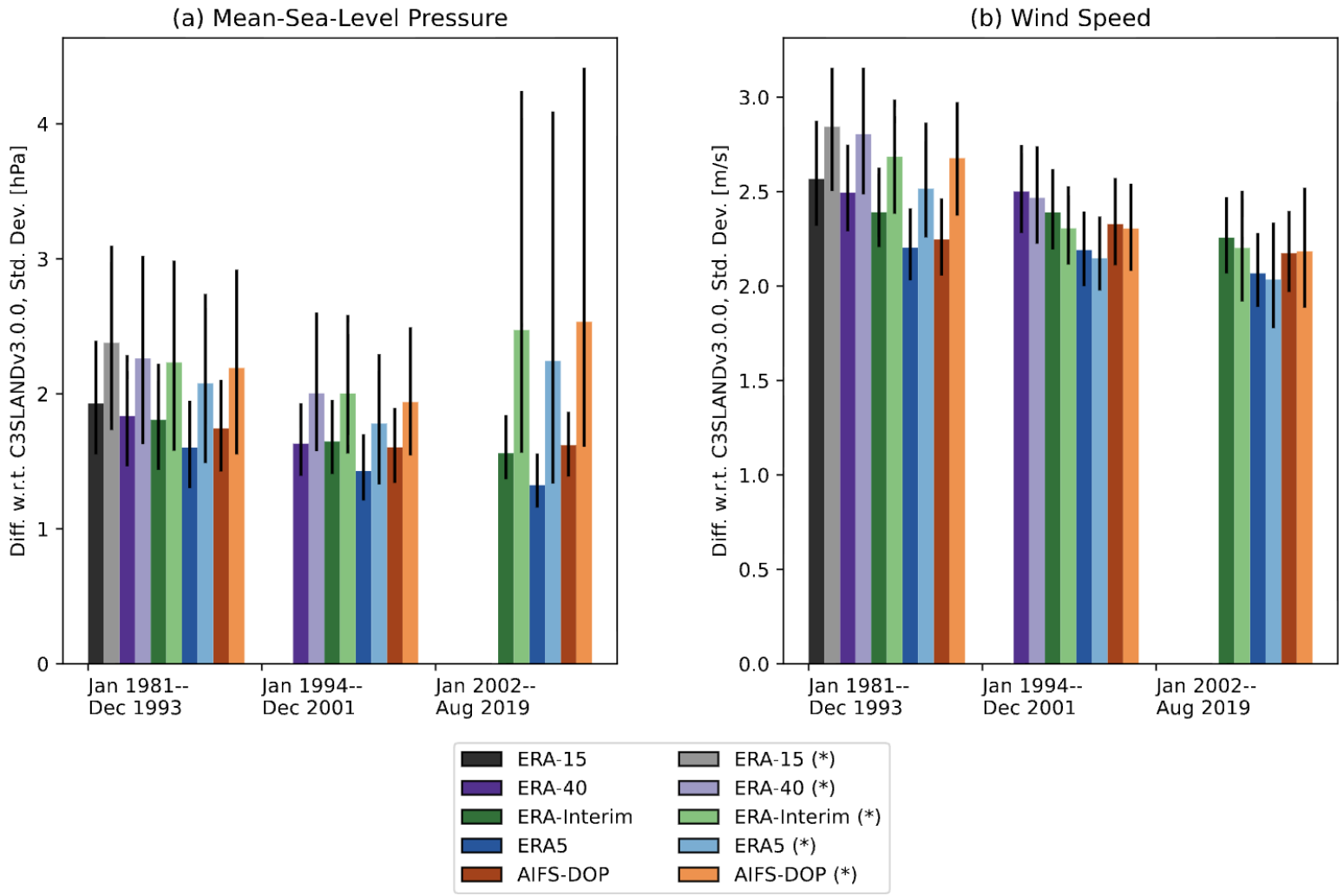
Has the model learned to generalize spatially or is it just memorizing the training dataset?

MISR AMVs are independent observations that aren't used by either ERA5 or DOP





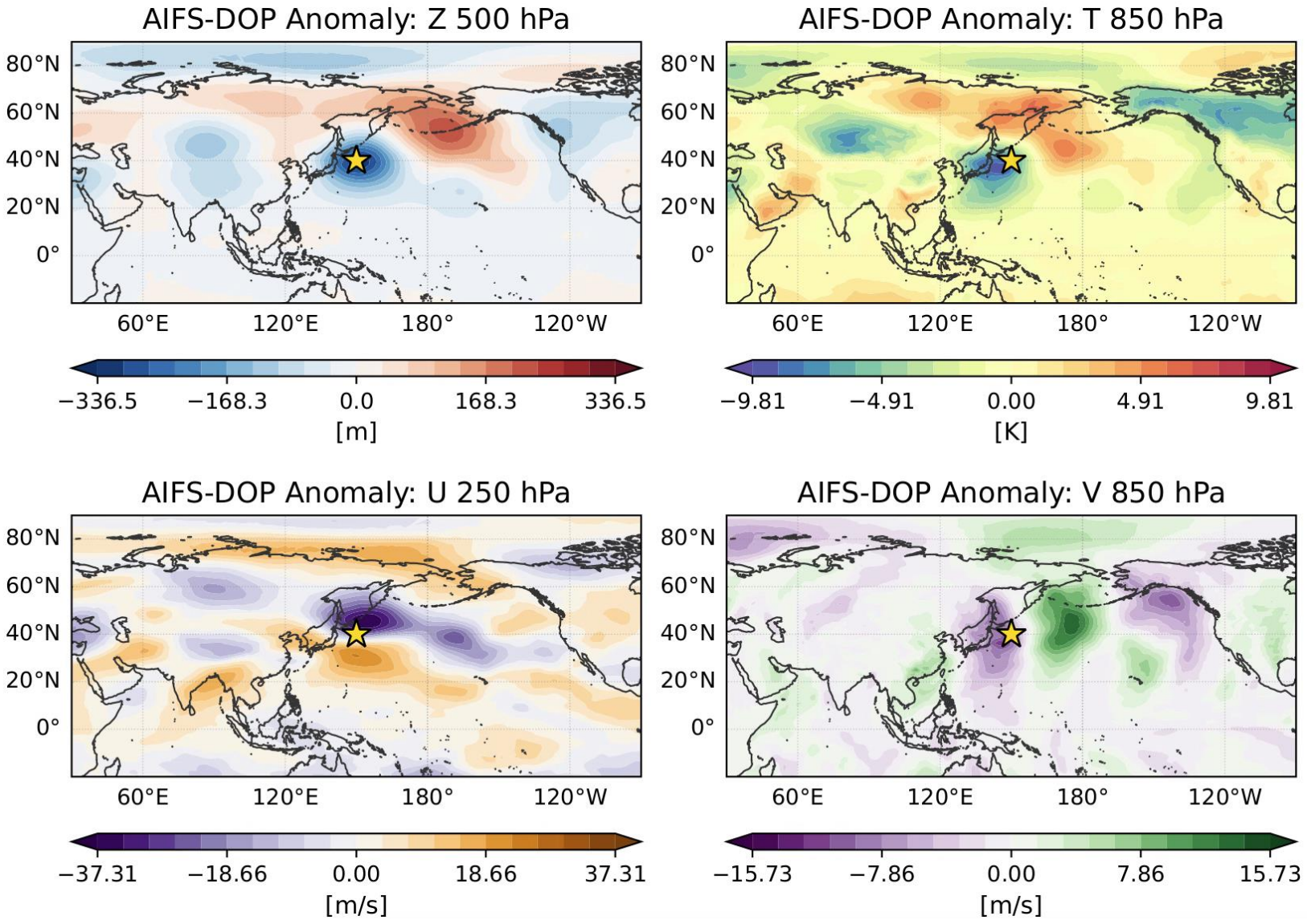
Evaluation against held-out observations: surface stations



Paul Poli

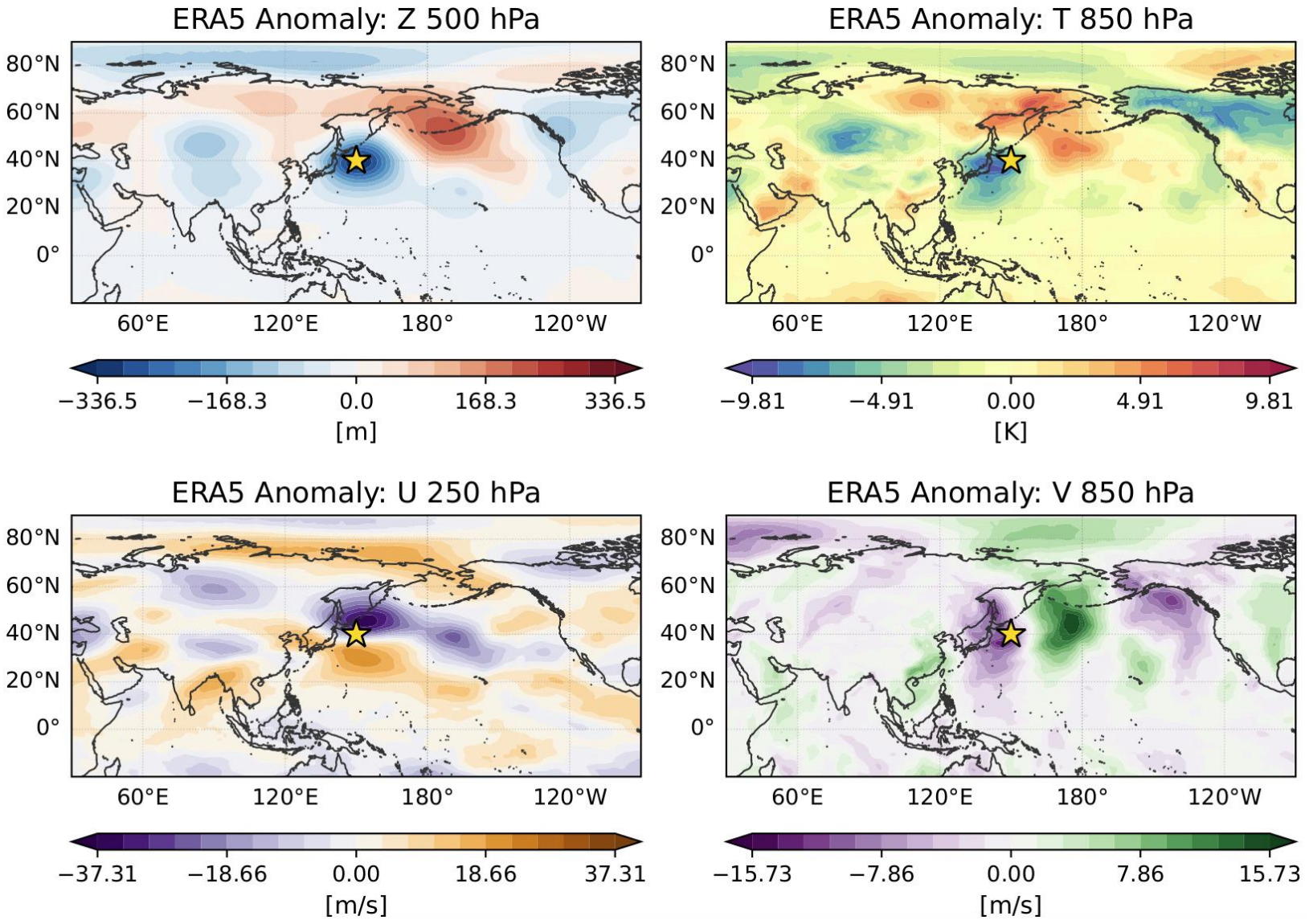
Physical consistency of fields:
Multi-variate regression patterns

AIFS-DOP



Physical consistency of fields:
Multi-variate regression patterns

ERA5



Summary of Direct Observation Prediction reanalysis prototype

- Data assimilation combines observations with a model prior to provide a spatially complete state estimate
- DOP learns a model of the Earth System during training which performs a similar role at inference time;
  - allowing gaps between sparse observations to be filled in a coherent way
- 40-year DOP reanalysis can be completed in **O(1-day)** opening-up many new possibilities

- **Encouraging results from this first prototype, but plenty of work still to do**

  to better understand, evaluate and improve the performance of data-driven models for reanalysis

Summary

- Data assimilatio
- DOP learns a m
  - allowing gaps b
- 40-year DOP re
- **Encouraging re**

  to better under

ototype

ate estimate

ence time;

reanalysis

Global reanalysis from observations alone with machine learning

Peter Lean[1,*], Ewan Pinnington[1], Patrick Laloyaux[1], Mihai Alexe[2], Eulalie Boucher[1], Simon Lang[1], Tomas Kral[1], Paul Poli[2], Hans Hersbach[1], Niels Bormann[1], Matthew Chantry[1], Anthony McNally[1]

[1]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
[2]ECMWF, Bonn, Germany
[*]Correspondence: peter.lean@ecmwf.int

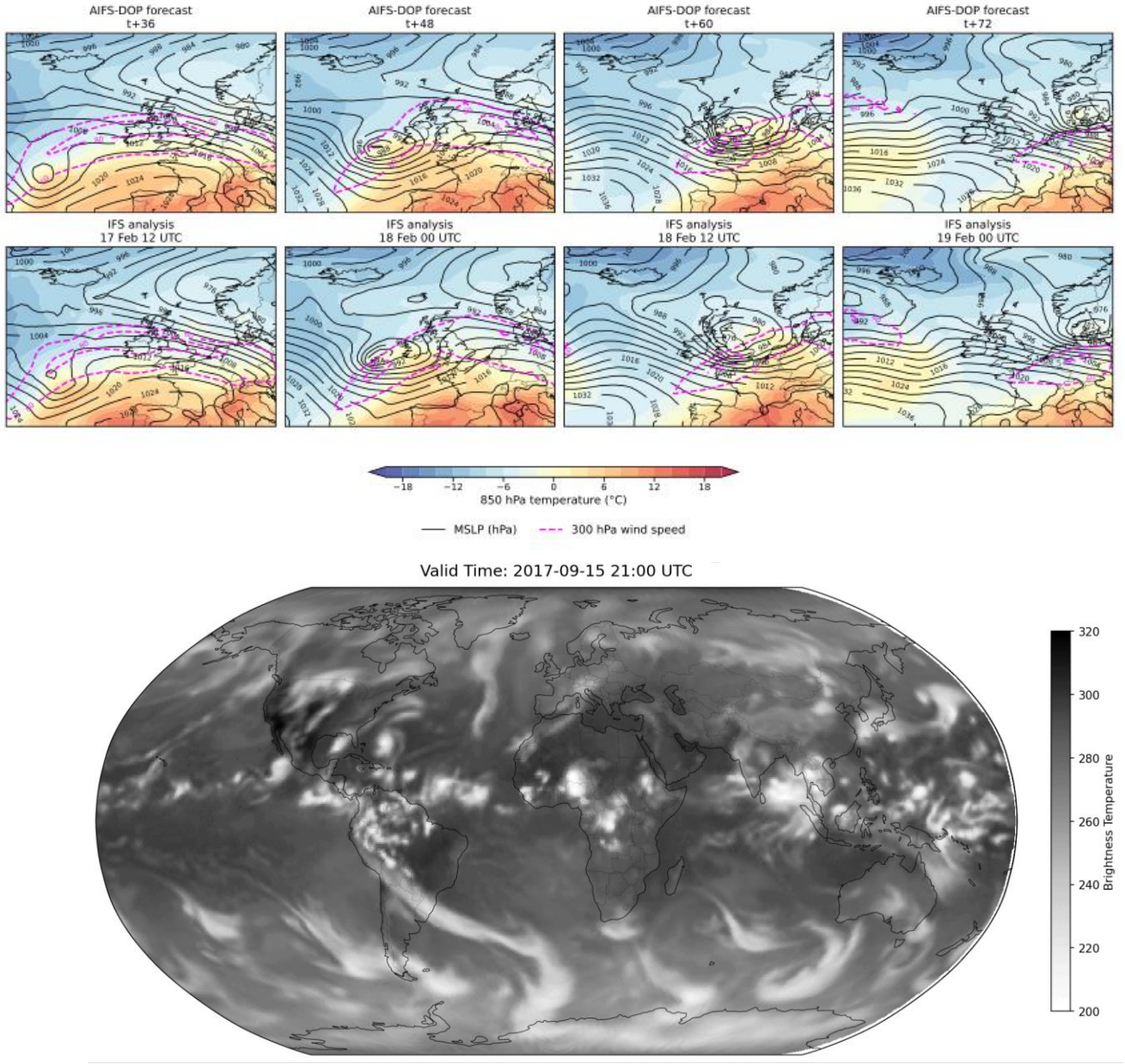
[physics.ao-ph] 8 Jul 2026

Abstract

Earth system reanalysis datasets are foundational for weather and climate research and provide the gridded training data used by most machine learning weather prediction systems. Here we show results from a prototype system that suggest that machine learning models trained only on Earth system observations can potentially be used to generate multi-decade global reanalyses without using physics-based numerical models. The resulting gridded fields capture large-scale atmospheric structure and variability across multiple timescales, while exhibiting signs of physical coherence in several key dynamical diagnostics. Evaluations of the prototype against held-out independent atmospheric observations indicate that the root mean square vector error of upper-level winds is close to that of ERA5 when compared at a consistent resolution, and that the standard deviation of the error at the surface is between that of 4th- and 5th-generation ECMWF reanalyses (ERA-Interim and ERA5). Furthermore, while traditional reanalysis production is computationally expensive, typically taking several years to produce, the reanalysis presented here was generated during the course of a single working day. These results suggest that observation-trained machine learning models offer a promising new approach for reanalysis production from observations alone.

Summary

- Training models directly from observations
- Dataset preparation and cleaning becomes extremely important!
- Extending training datasets back in time using harmonized EUMETSAT & NOAA records
- Extended observation dataset has allowed production of prototype “reanalysis”
- Pushing on to *real time* system and bridging gap with previous data records
- Lots more development needed!



Can a **data-driven** model handle **data-sparse** settings?

- Pre-satellite era

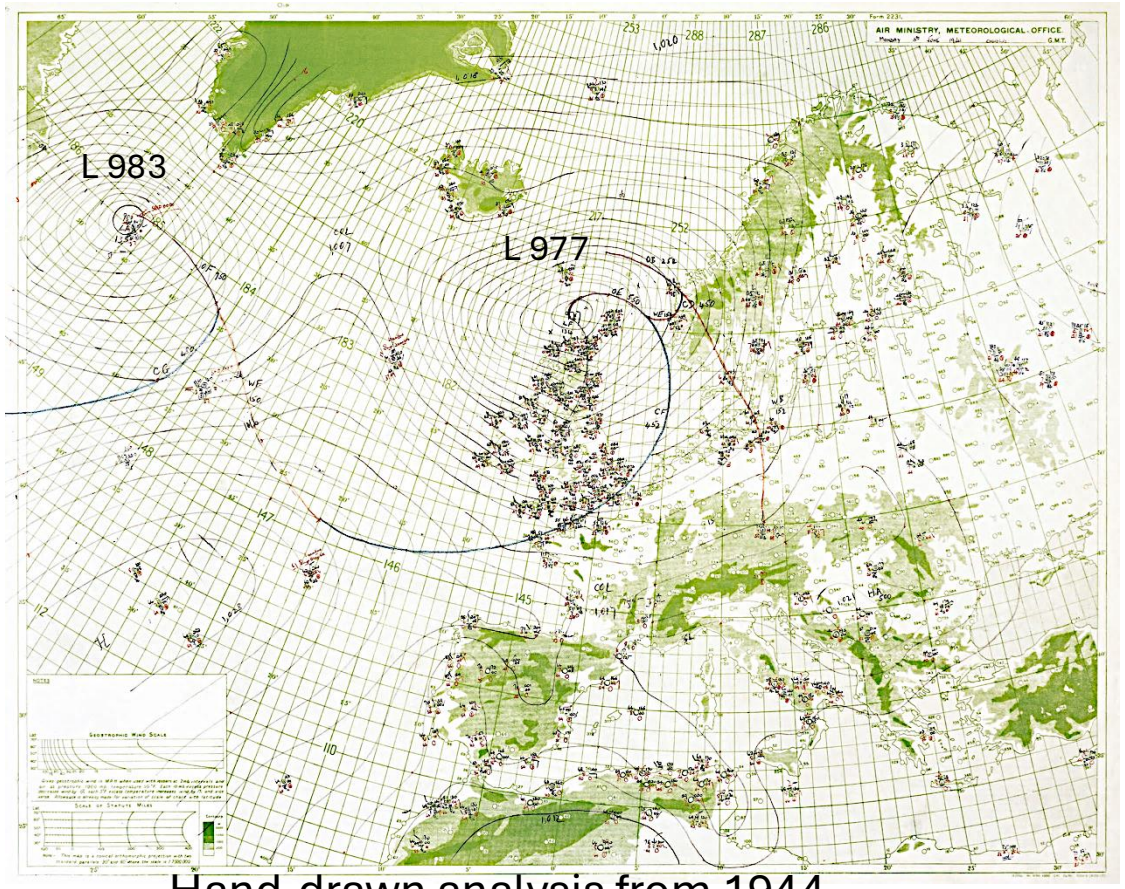
- Ocean

D-Day case study

Paul Poli, Patrick Laloyaux

Fine-tune model to only use conventional observations

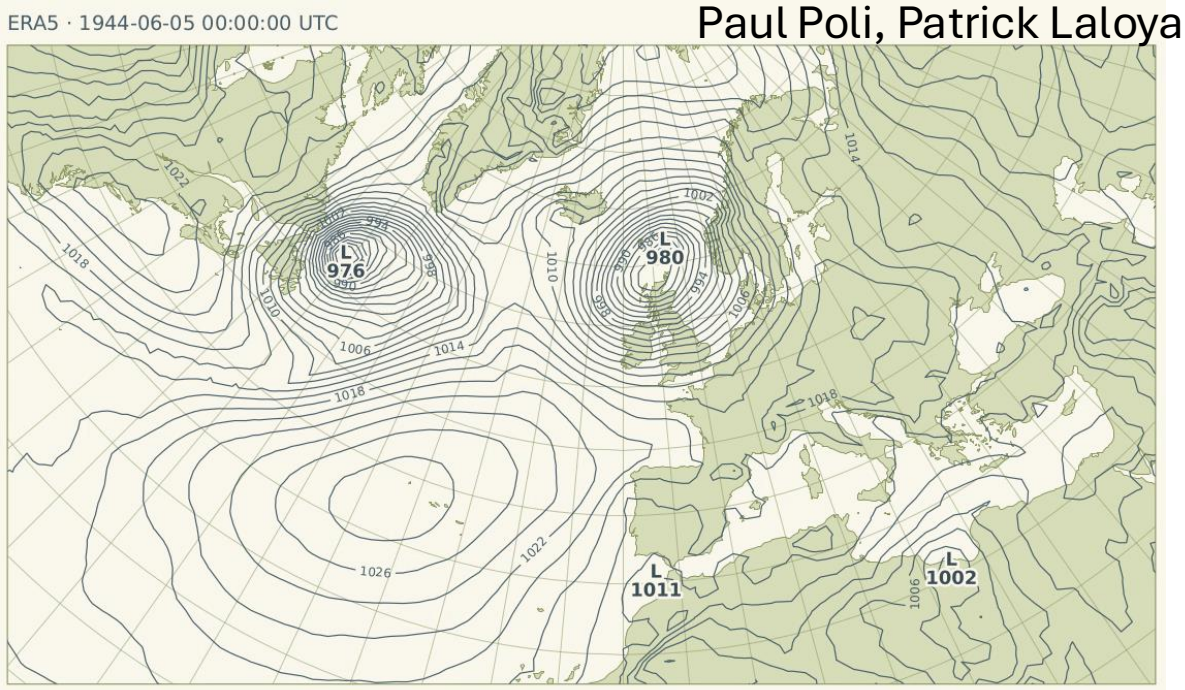
June 5th, 1944 00 UTC



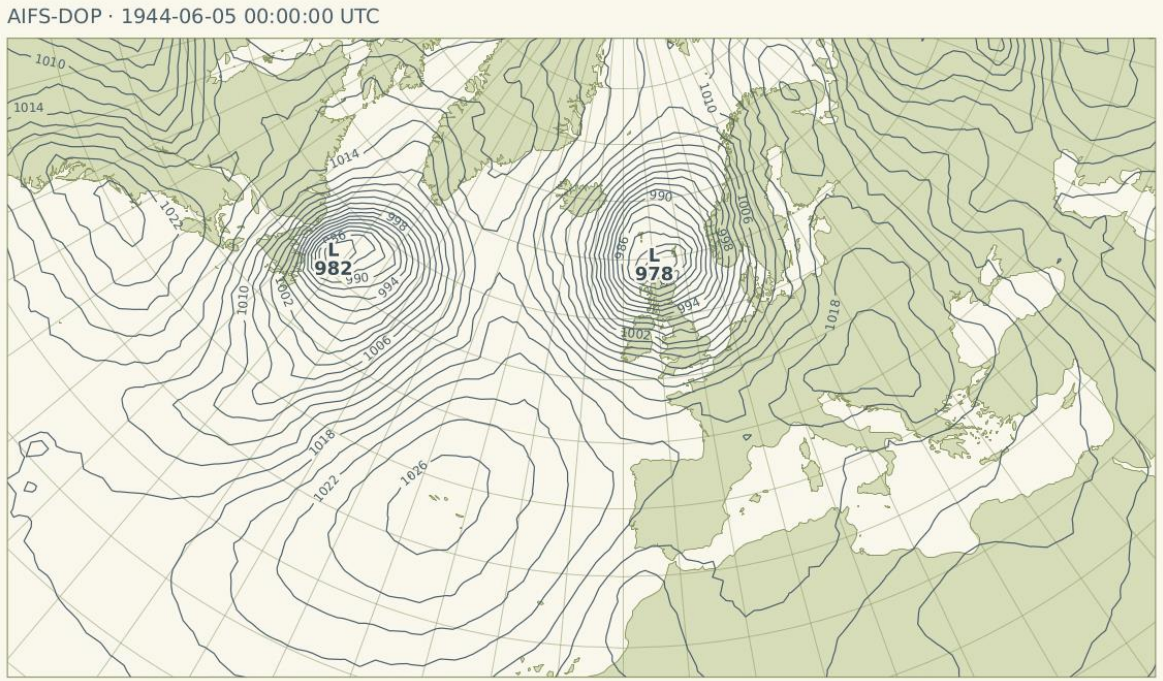
Hand-drawn analysis from 1944

Information provided by the National Meteorological Library and Archive – Met Office, UK
© Crown Copyright 1944.

ERA5



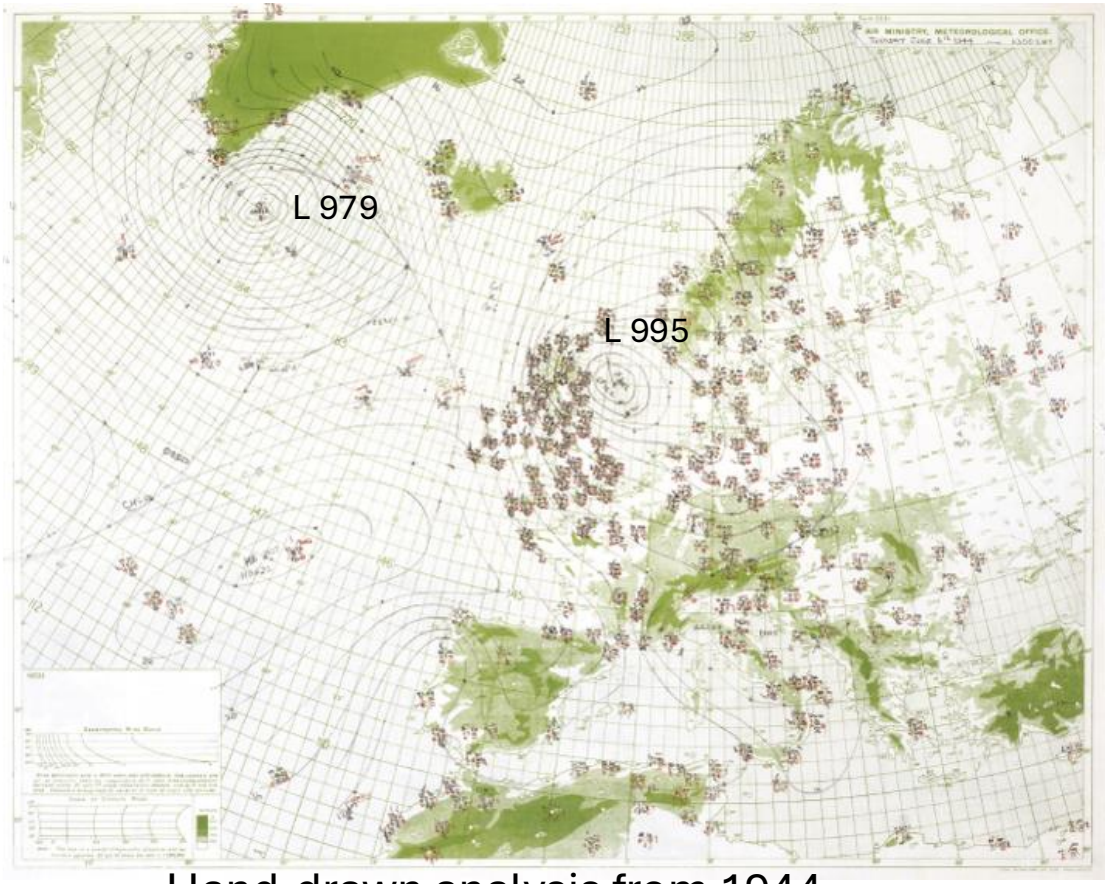
AIFS-DOP



D-Day case study

Paul Poli, Patrick Laloyaux

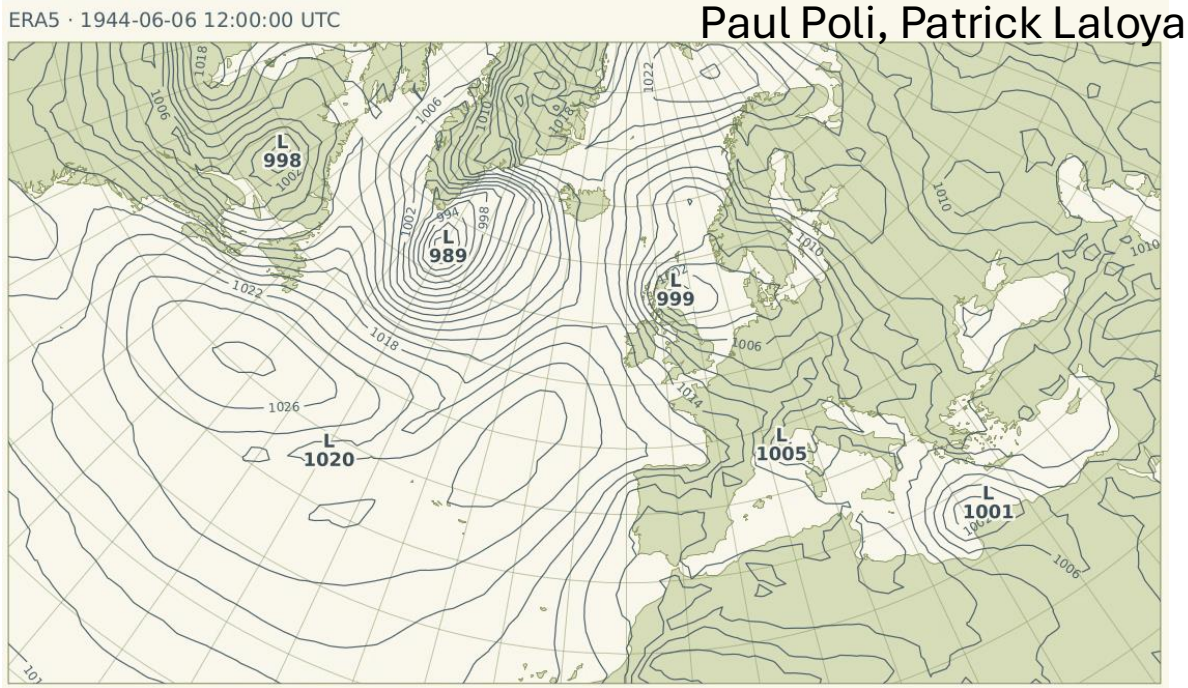
June 6th, 1944 12 UTC

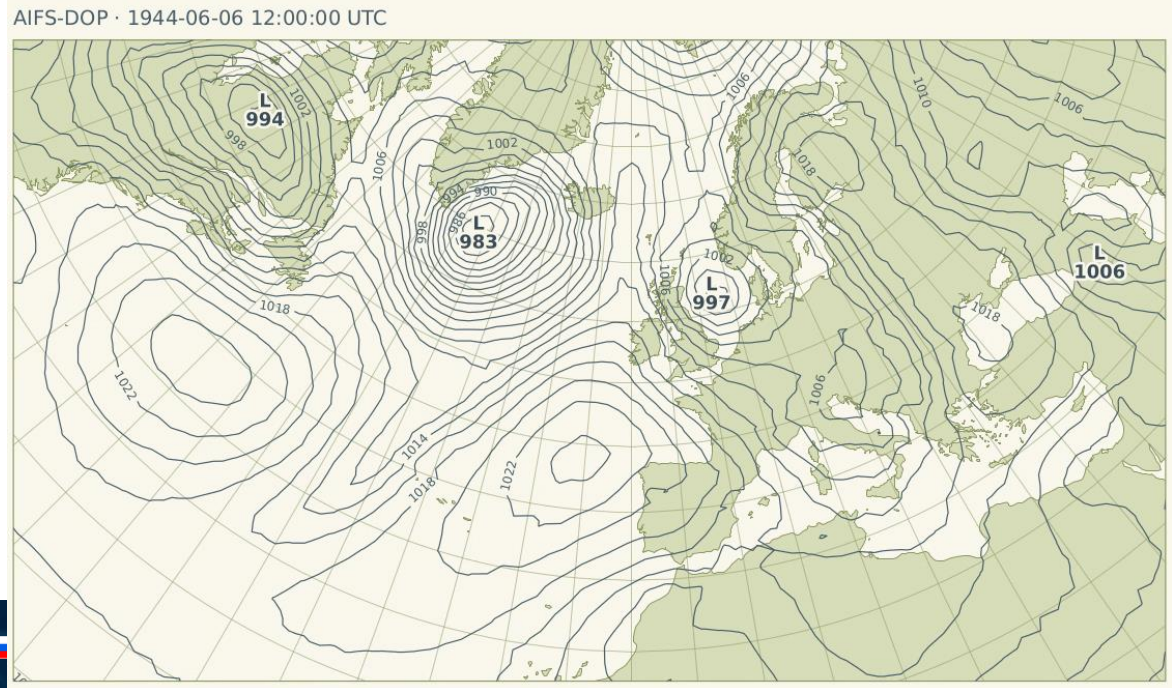


Hand-drawn analysis from 1944

Information provided by the National Meteorological Library and Archive – Met Office, UK
© Crown Copyright 1944.

ERA5





AIFS-DOP

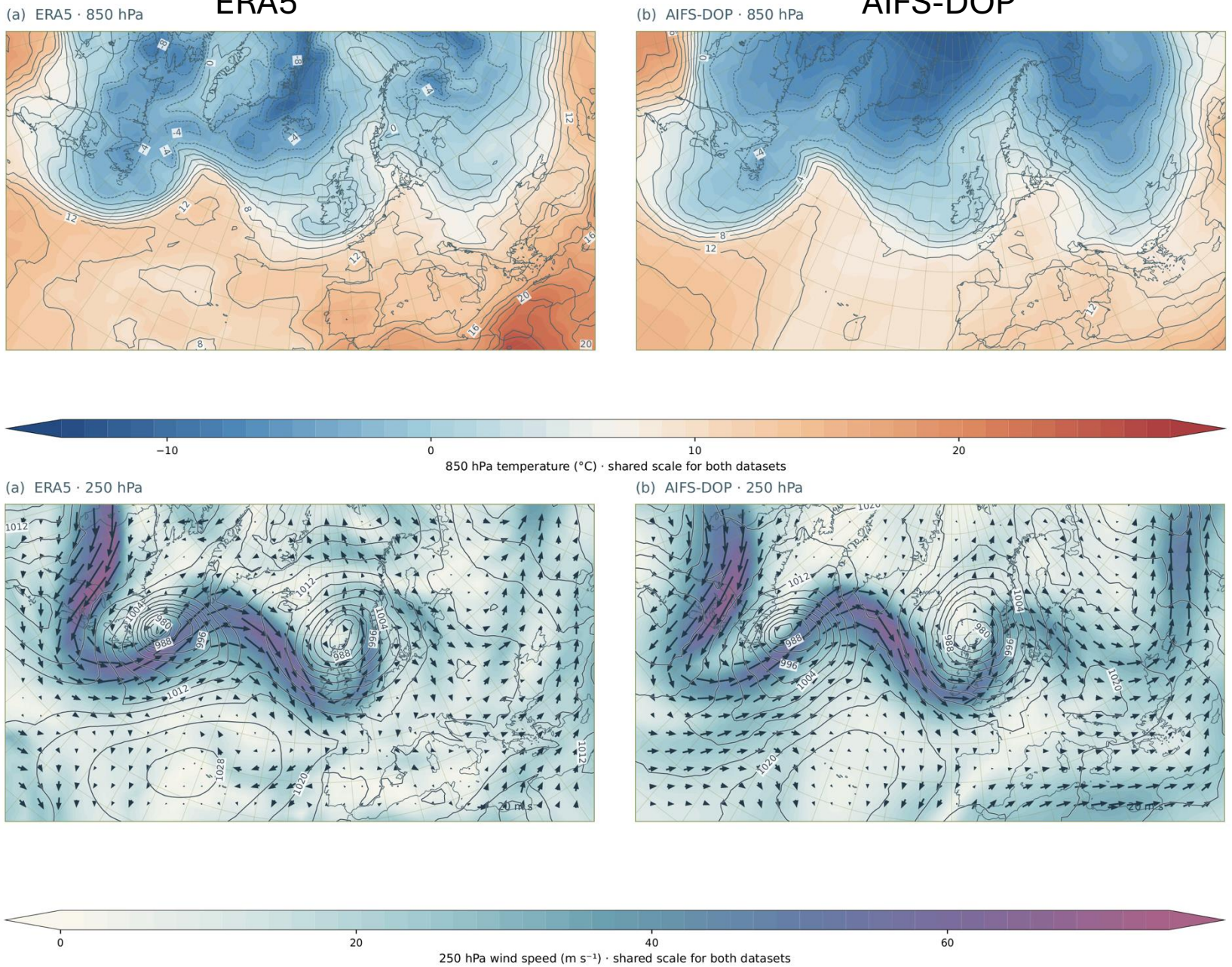
Case study: D-Day

June 5th, 1944 00 UTC

- No upper air temperature observations were input to AIFS-DOP
- No wind observations above 500hPa

ERA5

AIFS-DOP



850hPa temp

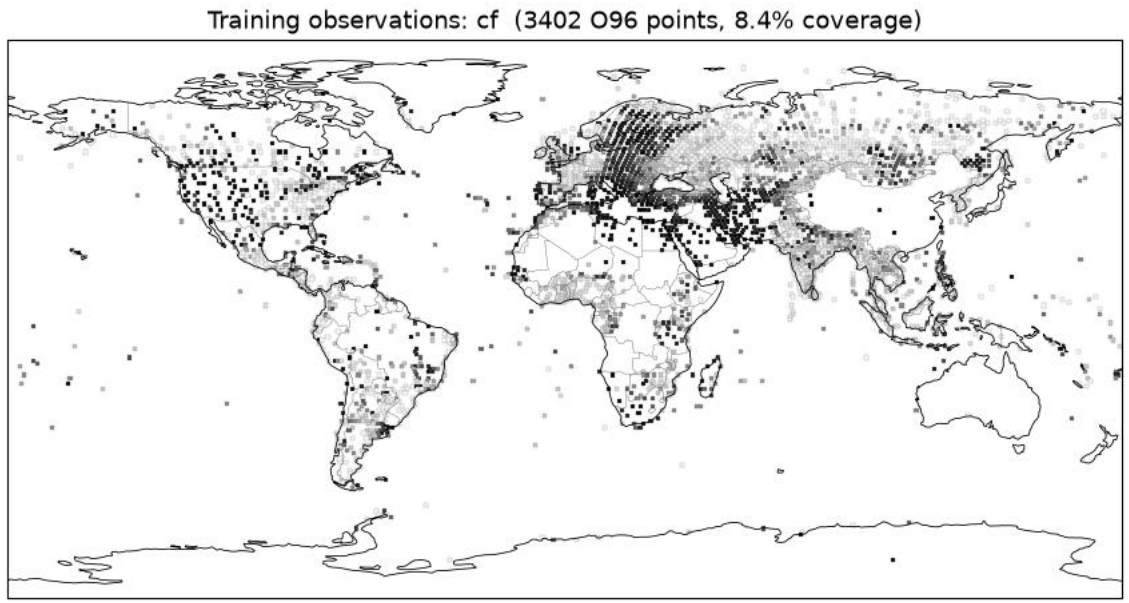
250hPa wind

Precipitation and Cloud from Sparse Observations

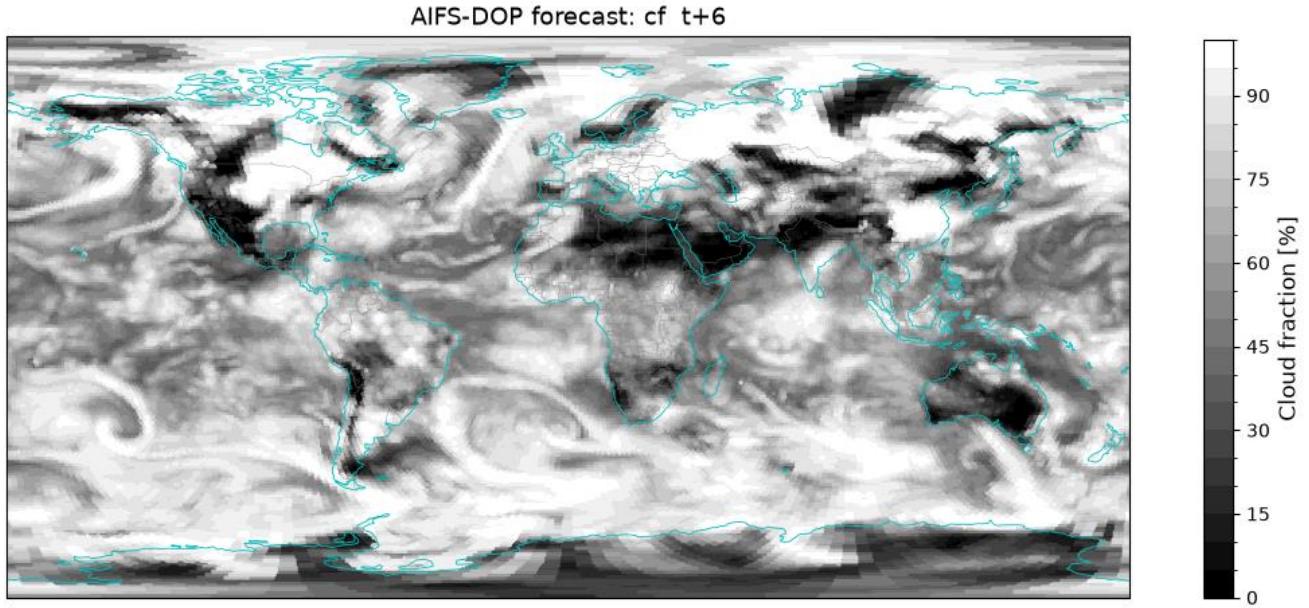
Clouds and precipitation in AIFS-DOP
Where is the information coming from?

- Cloud: targeting **SYNOP reports**, input information coming from all-sky IR (geostationary (training only), IASI/CrIS) and cloud-sensitive MW sounder channels

Example CF training observations



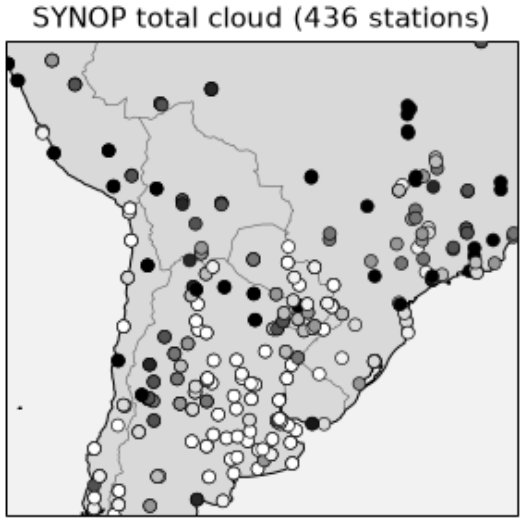
Example CF forecast



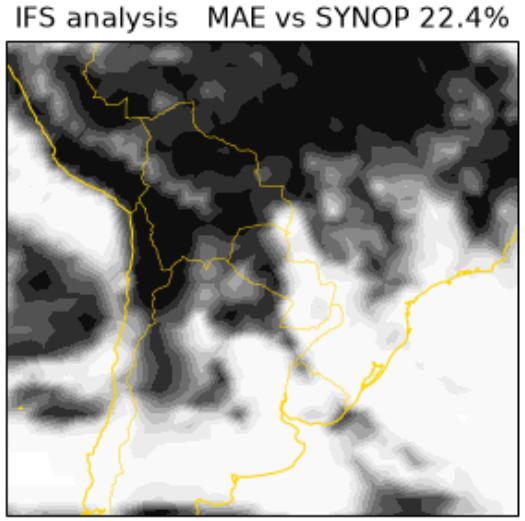
MW denial experiments

- Denial experiments give an indication on which observations are being used to produce cloud fraction forecasts
- Experiment: remove AMSU-A-like channels (23.8-89 GHz) channels at inference time, *from an already trained model*
- Results in degraded CF forecasts at all lead times (shown t+6h)
  - Changes the spatial cloud pattern
  - Increases the MAE against SYNOP observations

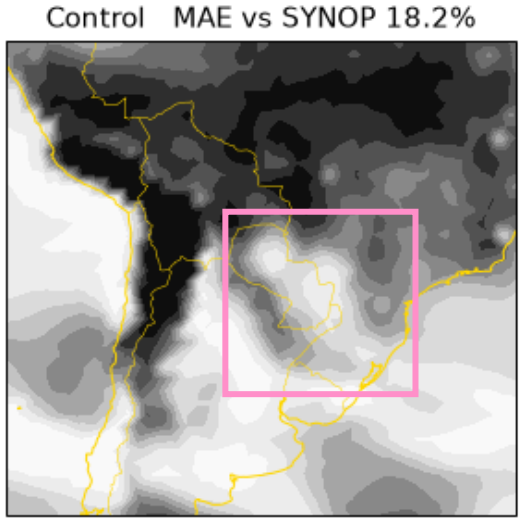
SYNOP CF



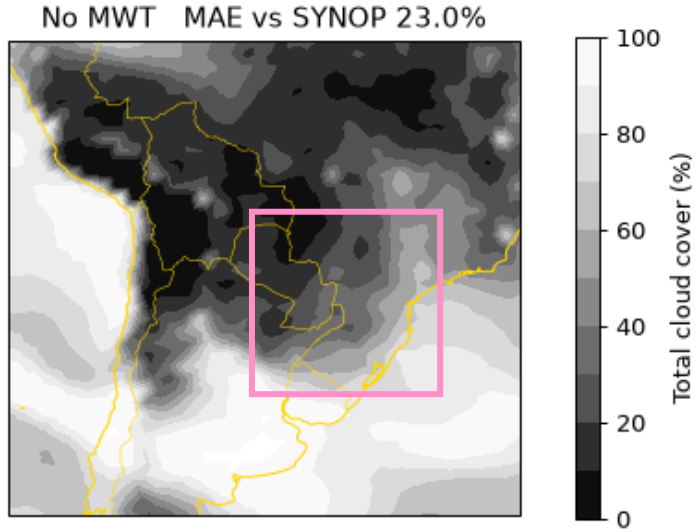
IFS ANALYSIS



CONTROL AIFS-DOP

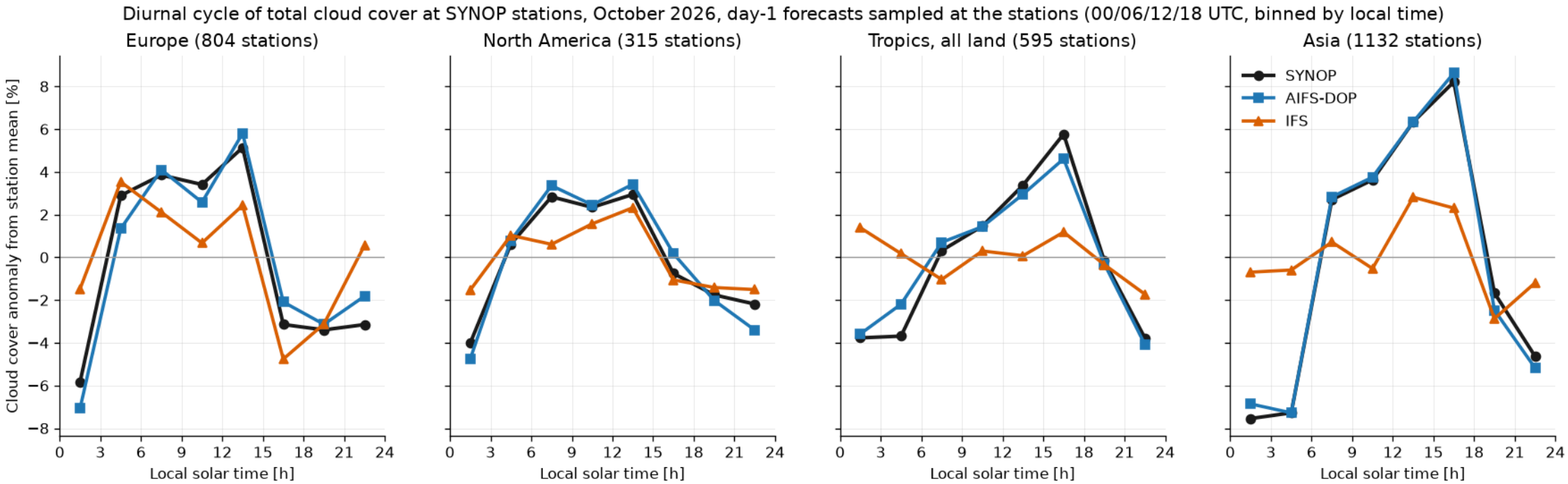


NO MWT AIFS-DOP



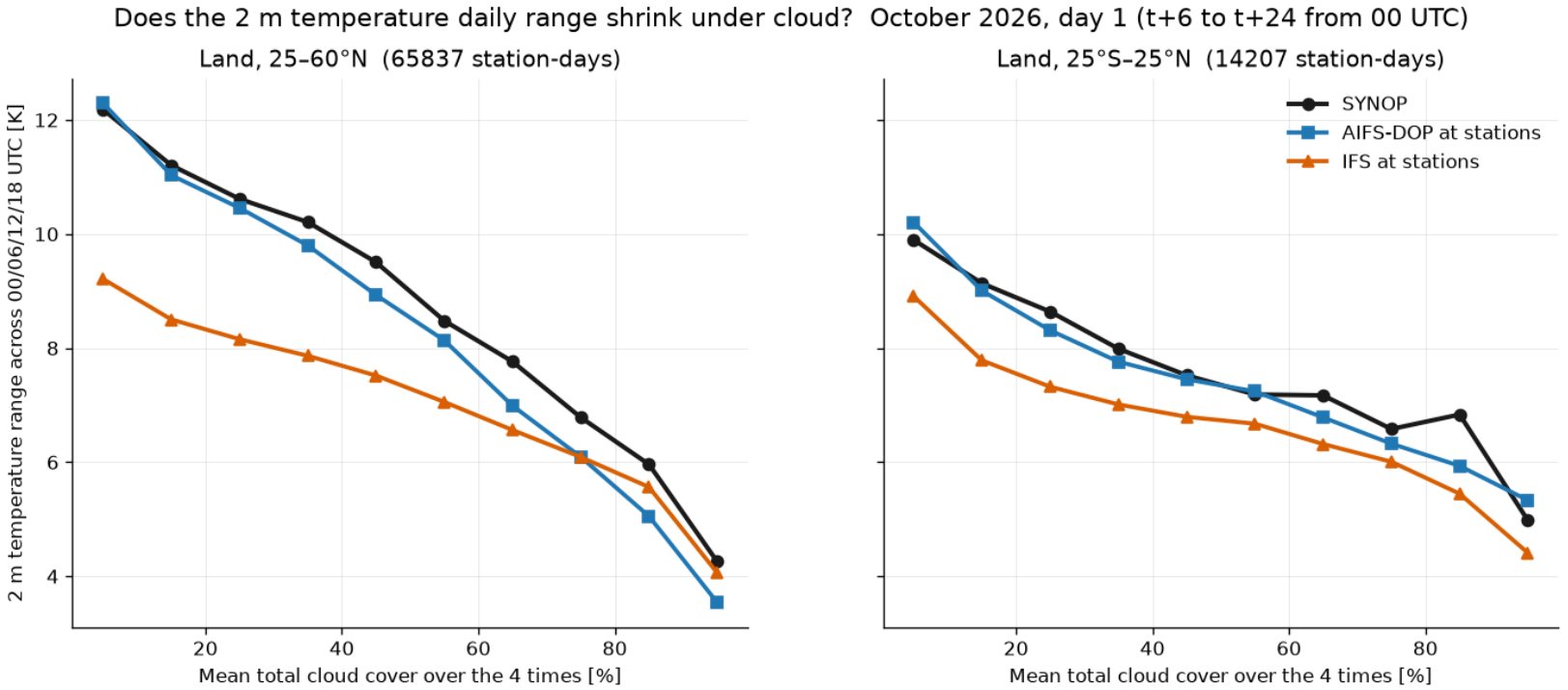
Diurnal cycle of total cloud cover

- Training directly on SYNOP cloud reports helps AIFS-DOP capture the observed diurnal cycle
- Follows the observed daily variation in cloud cover more closely
- Clearest over tropical land and Asia where AIFS-DOP captures the strong afternoon cloud peak



Relationship between 2m temperature and cloud cover

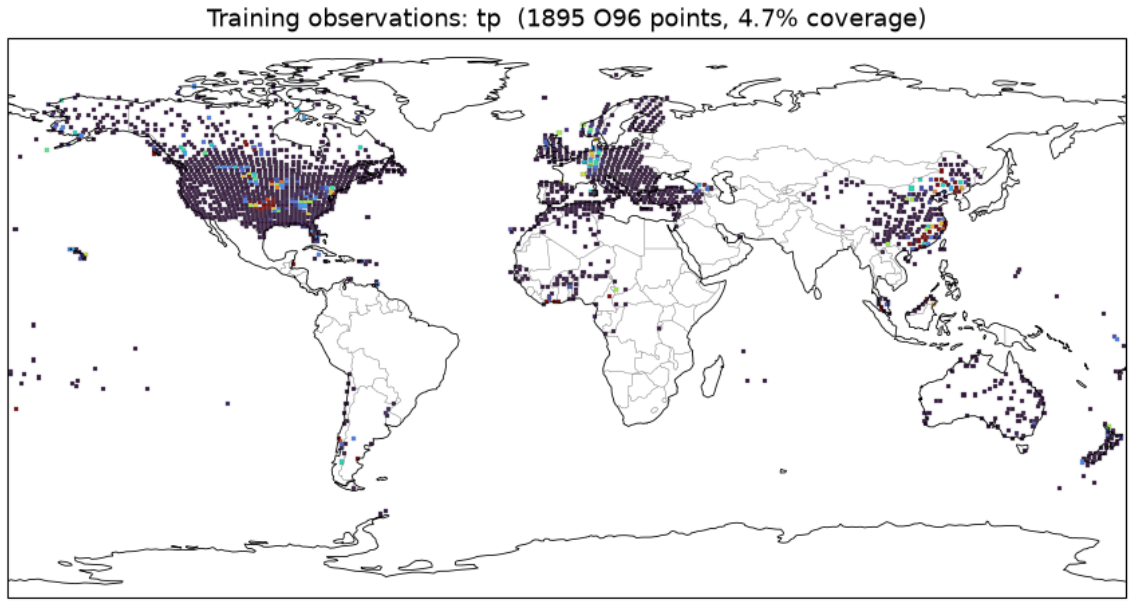
- AIFS-DOP captures the observed reduction in the 2m temperature range as total cloud cover increases
- AIFS-DOP follows this relationship closely
- This provides evidence that training on observations help AIFS-DOP learn a physically meaningful relationship between clouds and temperature



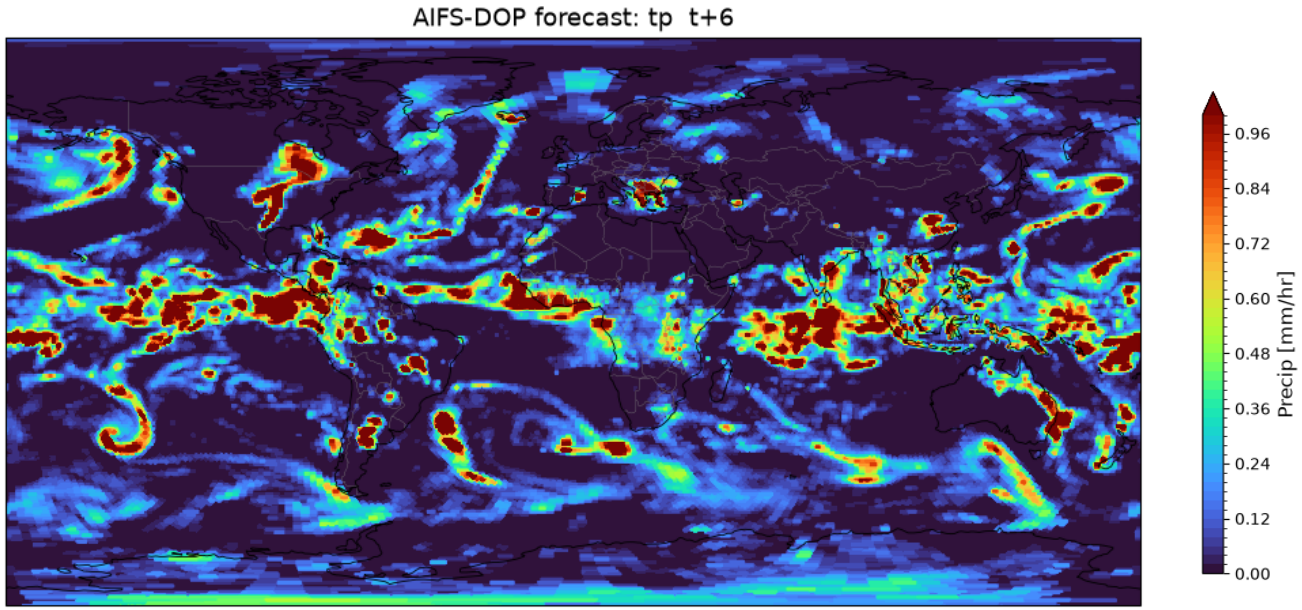
Clouds and precipitation in AIFS-DOP
Where is the information coming from?

- Total precipitation: GHCN SYNOP reports (rain gauges)

Example TP training observations

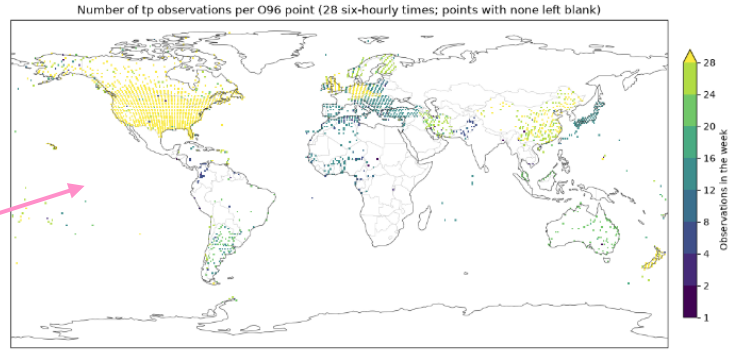


Example TP forecast

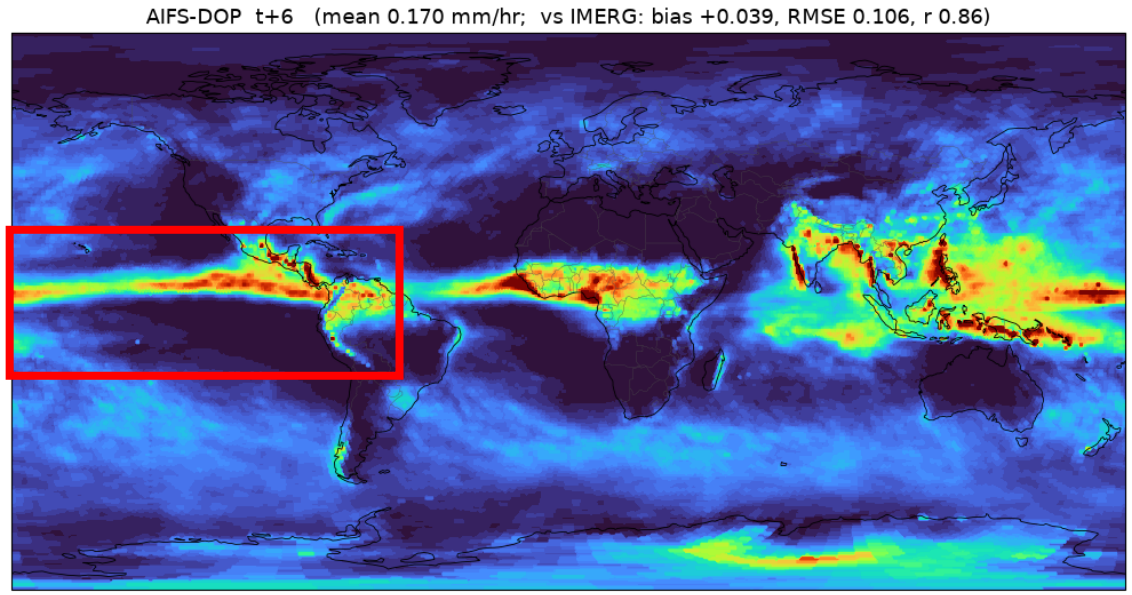


Comparison with IMERG

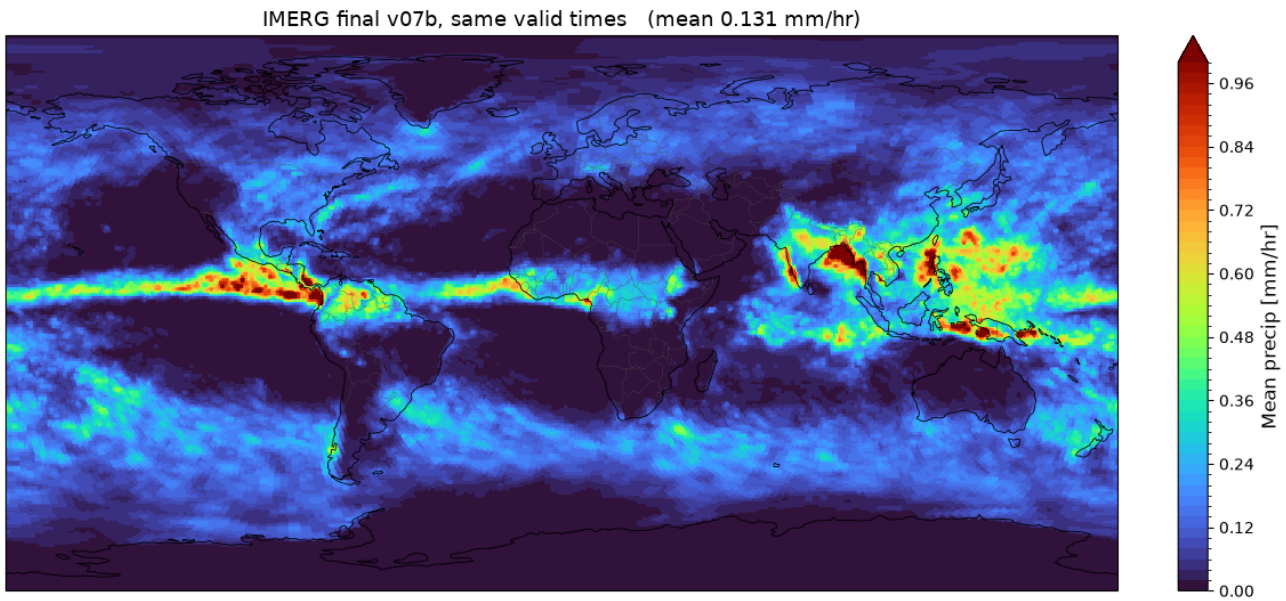
- IMERG here is used only for validation (not used in training)
- **Average rainfall June-July 2025**
- Large-scale rainfall patterns are present; Asian monsoon around India, tropical rain belt…
- All of this learnt only from land stations!
- Wet bias compared to IMERG



Average tp in AIFS-DOP t+6h forecasts



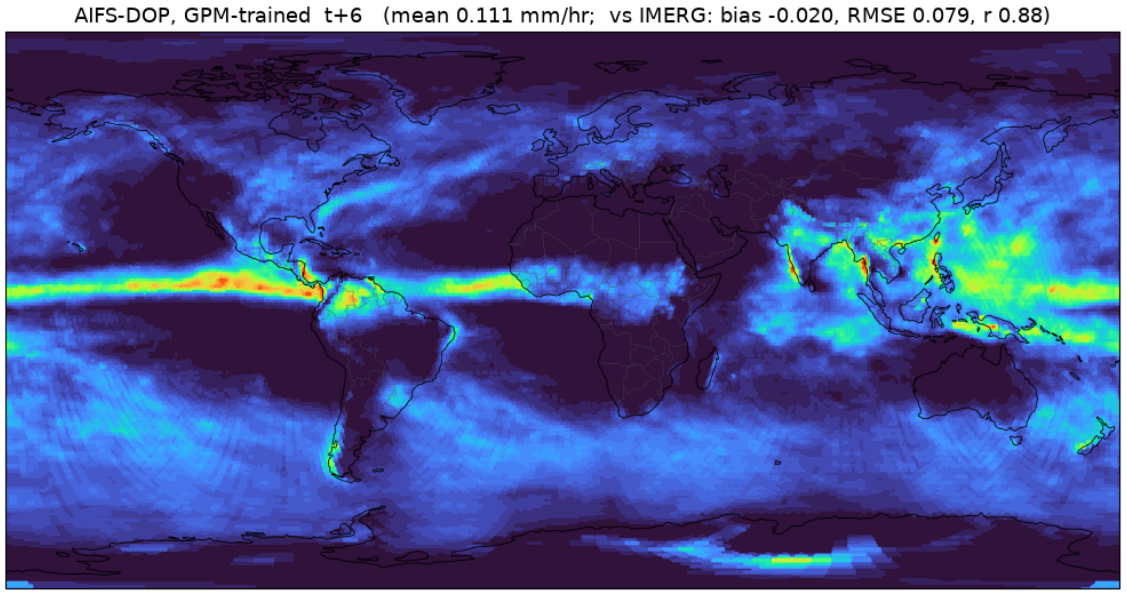
Average tp in IMERG at same valid times



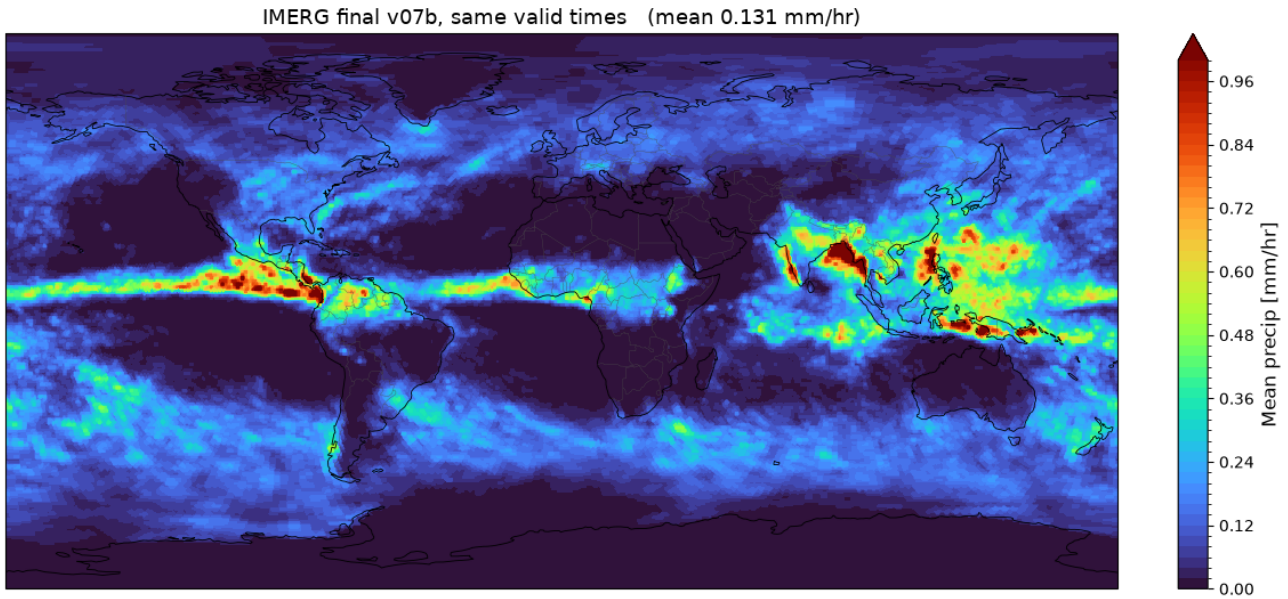
Comparison with IMERG

- Another possibility is to train towards GPM radar data
- Slightly closer agreement with IMERG: lower error and higher correlation than the gauge-trained run.
- This time overall dry bias compared to IMERG
- Maybe using a combination of both observation types would be beneficial

Average tp in AIFS-DOP t+6h forecasts



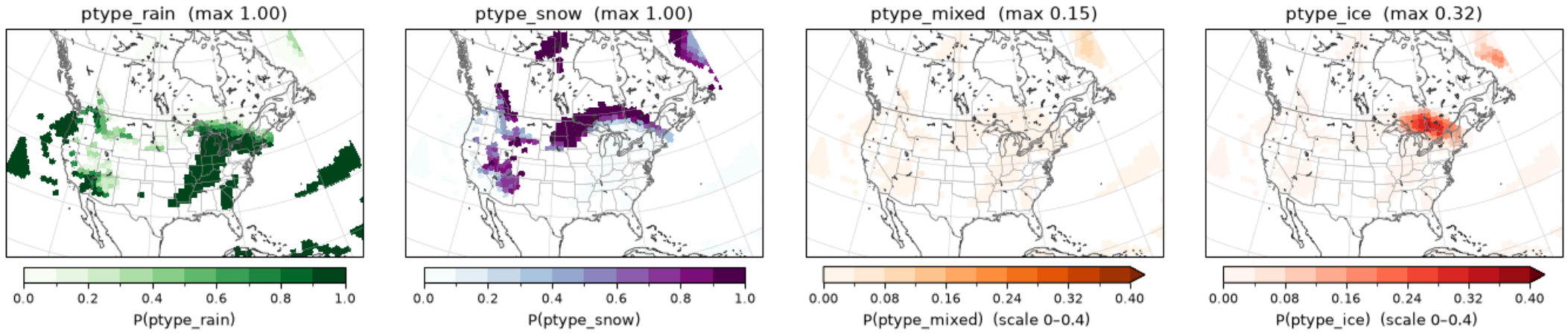
Average tp in IMERG at same valid times



Experimental: precipitation type

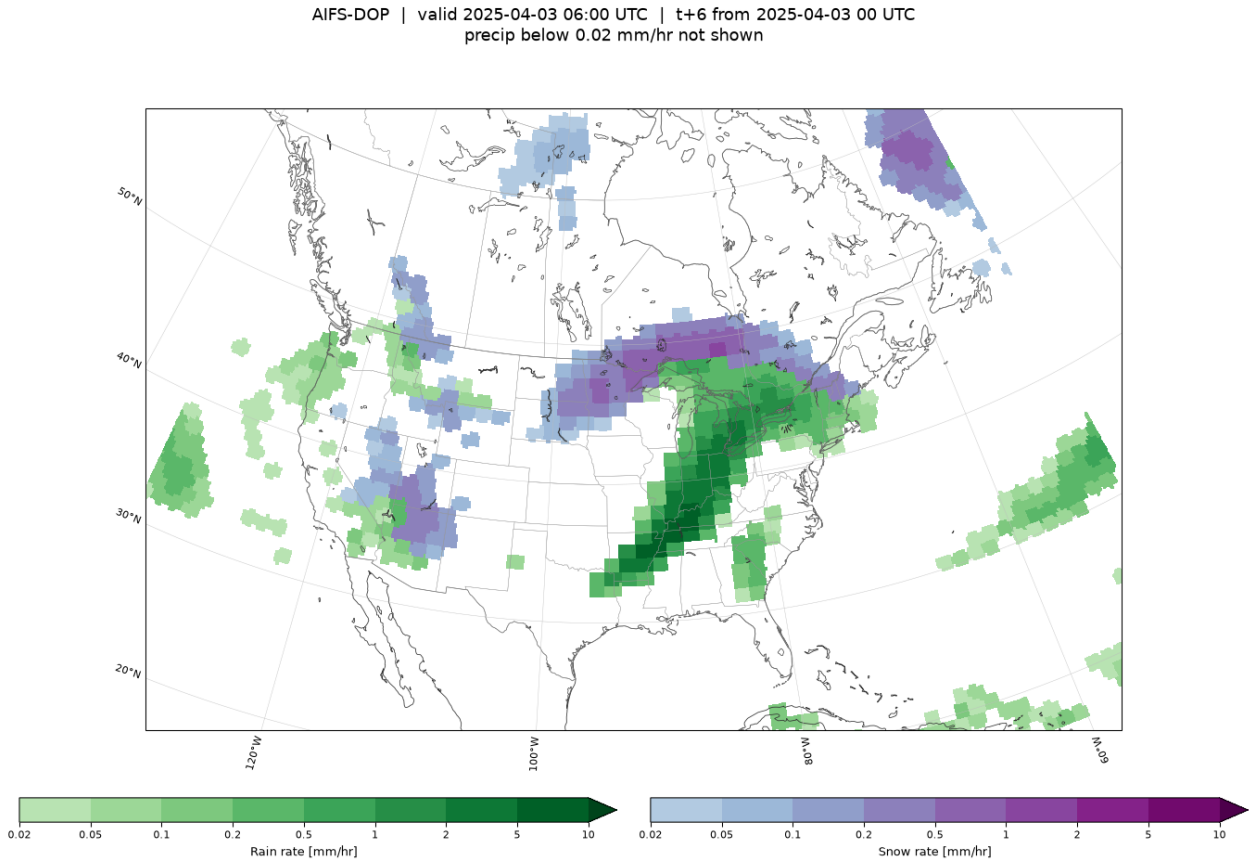
- Manual SYNOP reports in the present weather category (“snow showers”, “heavy rain”, “ice pellets” etc.
- Assigned these weather codes into one of four categories: ptype_rain, ptype_snow, ptype_mixed, ptype_ice
- Here, AIFS-DOP predicts “if there is precipitation, which type will it be” ?

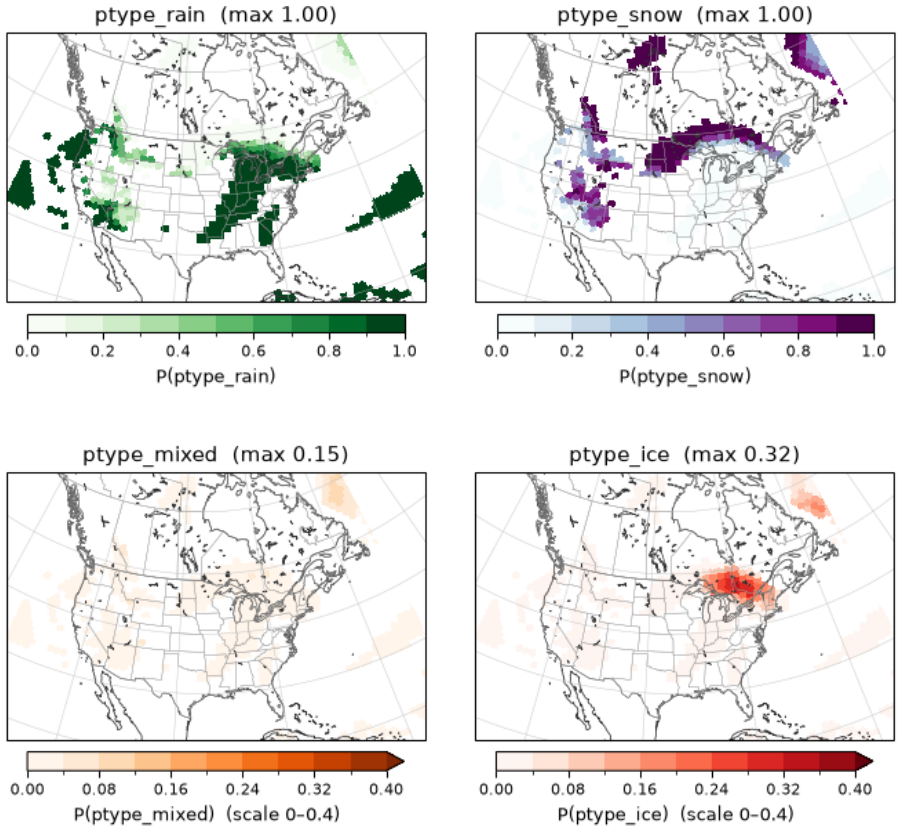
Example of the predicted probabilities for each category (North America, 3 April 2025)



Experimental: precipitation type

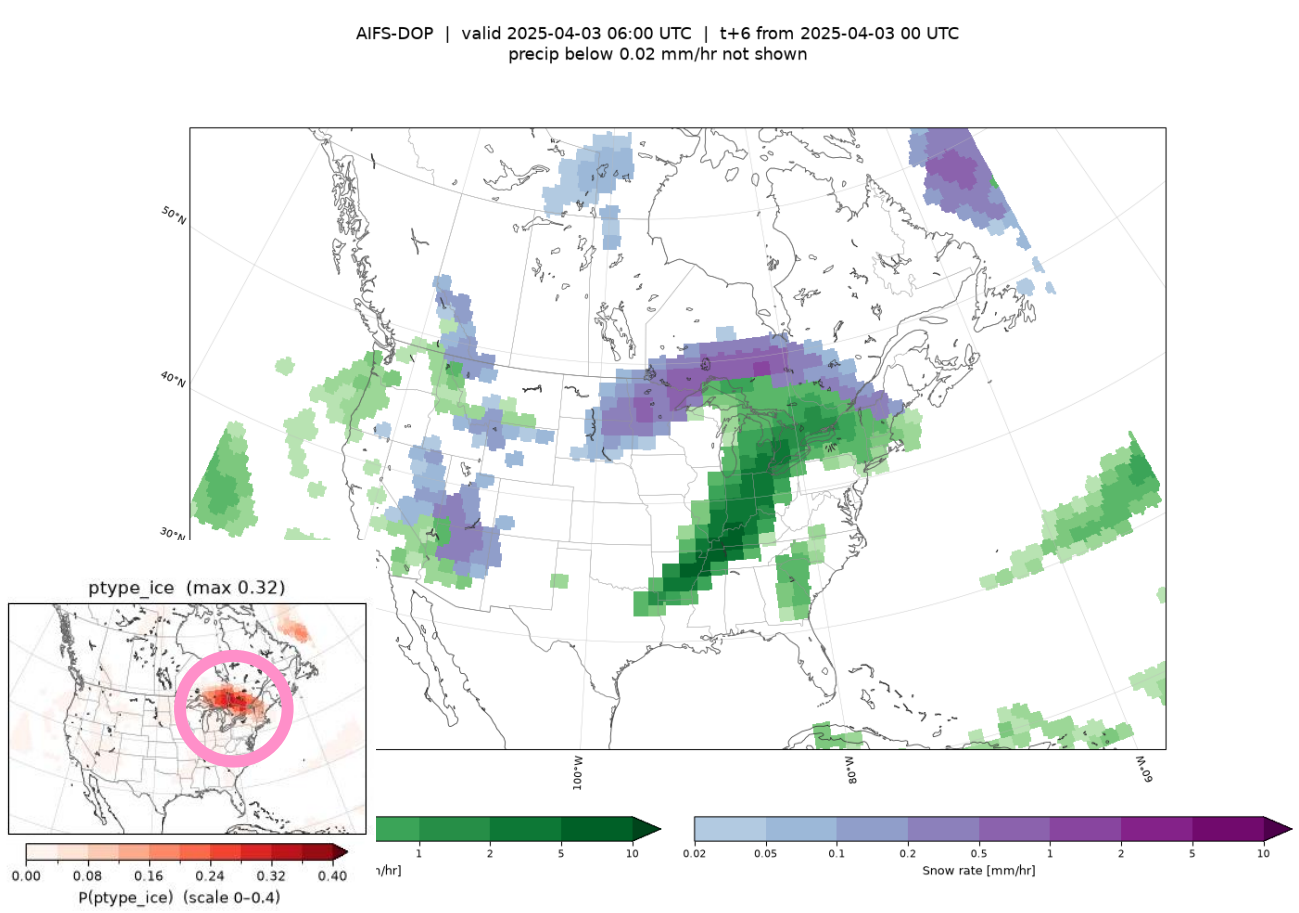
- Associated with the predicted GPM precipitation rate
- Colour shading adjusted based on the ptype category with the highest value for this grid point

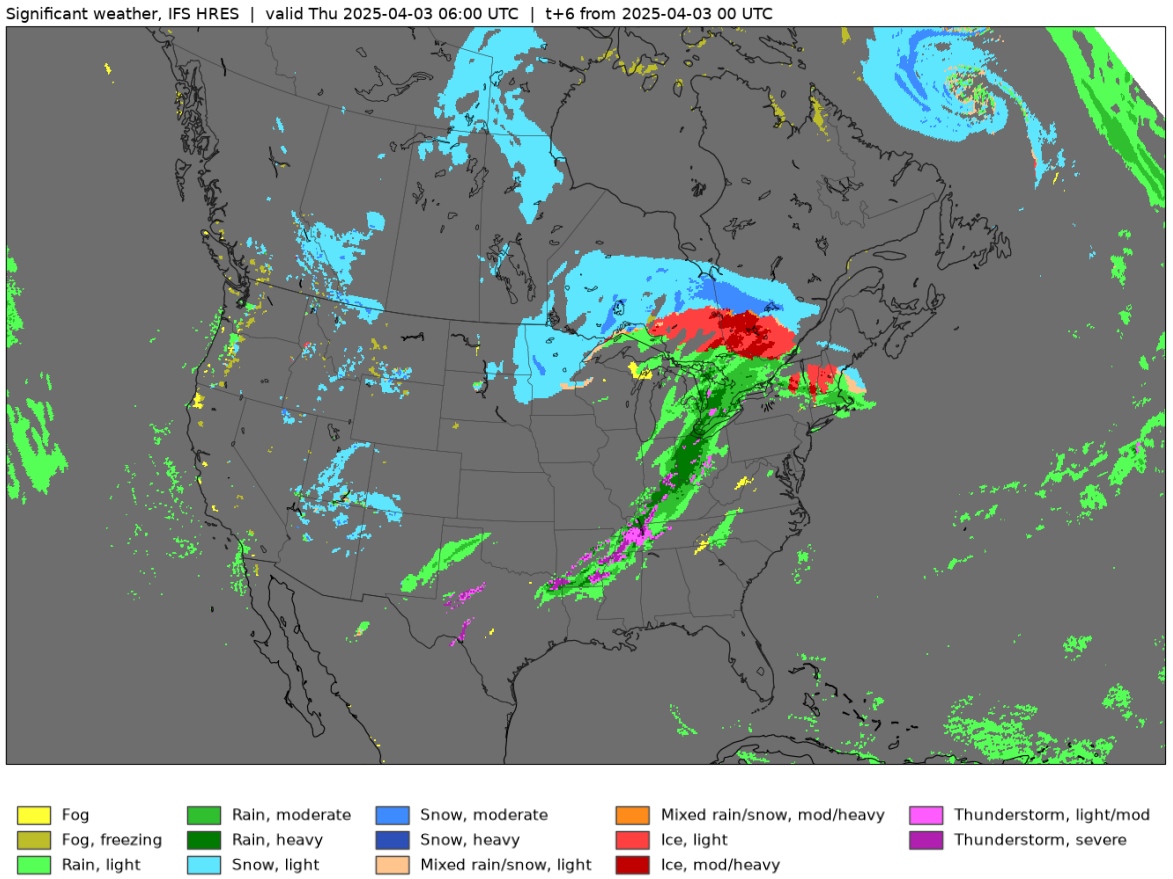




Experimental: precipitation type

- Associated with the predicted GPM precipitation rate
- Colour shading adjusted based on the ptype category with the highest value for this grid point





Thanks!

Ewan Pinnington

ECMWF

ewan.pinnington@ecmwf.int



www.ecmwf.int