Sorbonne Université, Université Paris-Saclay
Master Sciences de l'Océan, de l'Atmosphère et du Climat (SOAC)
Parcours Météorologie, Océanographie, Climat, Ingénierie pour les Observations Spatiales (MOCIS)

Institut polytechnique de Paris
Master 2 Water, Air, Pollution and Energy (WAPE)

Année 2025-2026

Course *Introduction to data assimilation* (*NUM2.2*, *ADOMO*)

From numerical modelling to data assimilation

Olivier Talagrand

6 January 2026

- What is assimilation ?
  - *Numerical weather prediction. Principles and performances*
  - *Definition of initial conditions*

- Bayesian Estimation

- One first step towards assimilation : 'Optimal Interpolation'

- The *Best Linear Unbiased Estimate* (*BLUE*)

The temporal dimension : Kalman Filter and Variational Assimilation

- *Numerical weather prediction. Principles and performances*

Results from

European Centre for Medium-range Weather Forecasts (ECMWF)

(Centre Européen pour les Prévisions Météorologiques à Moyen Terme, CEPMMT)

Located in Reading (UK), Bologna (Italy) and Bonn (Germany)

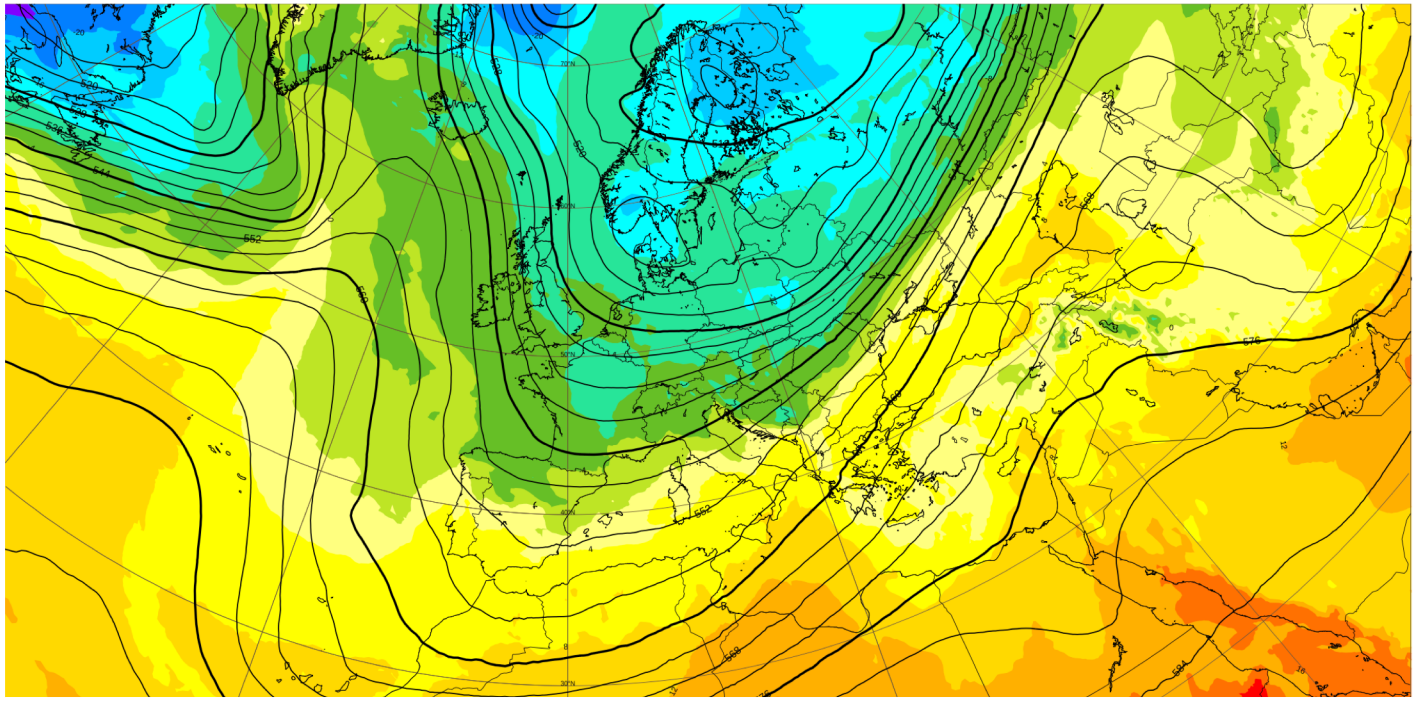
Base time: Wed 31 Dec 2025 00 UTC Valid time: Mon 05 Jan 2026 00 UTC (+120h) Area : Europe

ECMWF
HRES

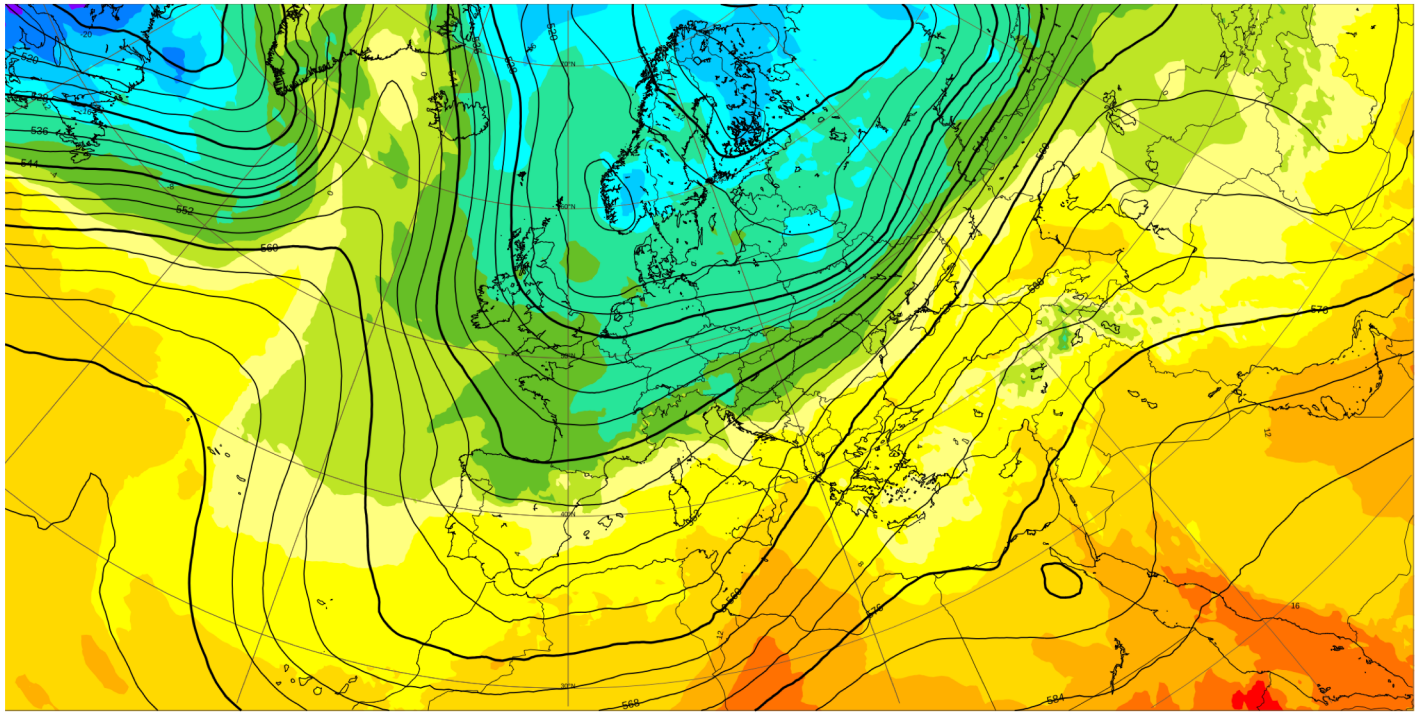
850 hPa temperature (C)

56 52 48 44 40 36 32 28 24 20 16 12 8 4 0 -4 -8 -12 -16 -20 -24 -28 -32 -36 -40 -44 -48 -52 -60 -70 -80

500 hPa geopotential (dm)

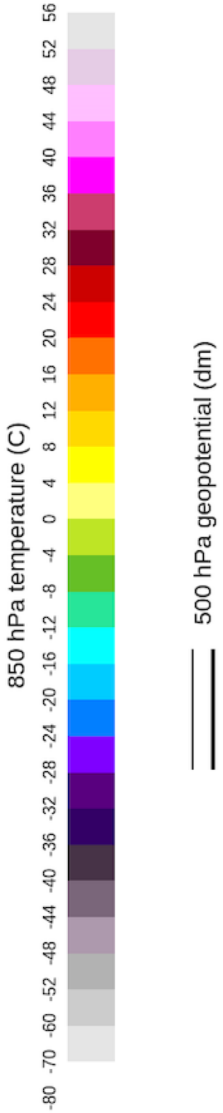


Base time: Mon 05 Jan 2026 00 UTC Valid time: Mon 05 Jan 2026 00 UTC (+0h) Area : Europe

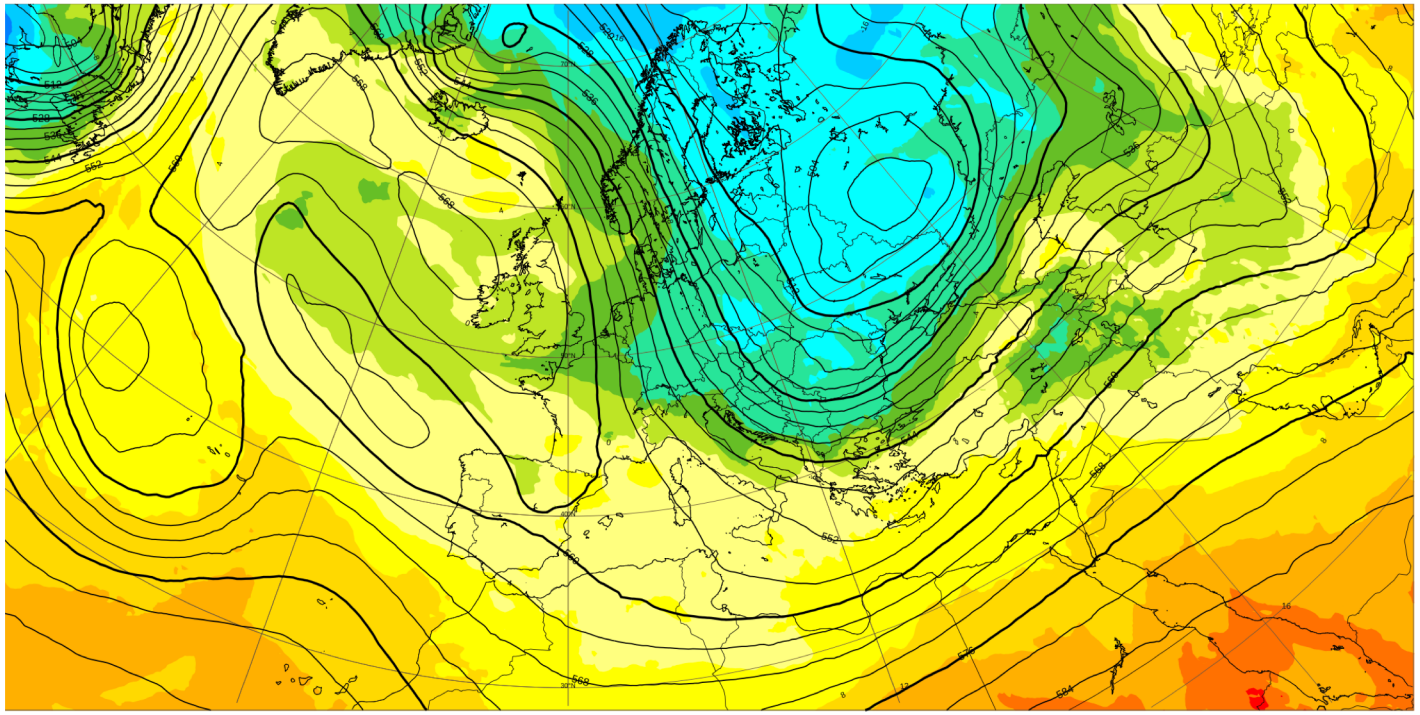


Base time: Wed 31 Dec 2025 00 UTC Valid time: Wed 31 Dec 2025 00 UTC (+0h) Area : Europe

ECMWF

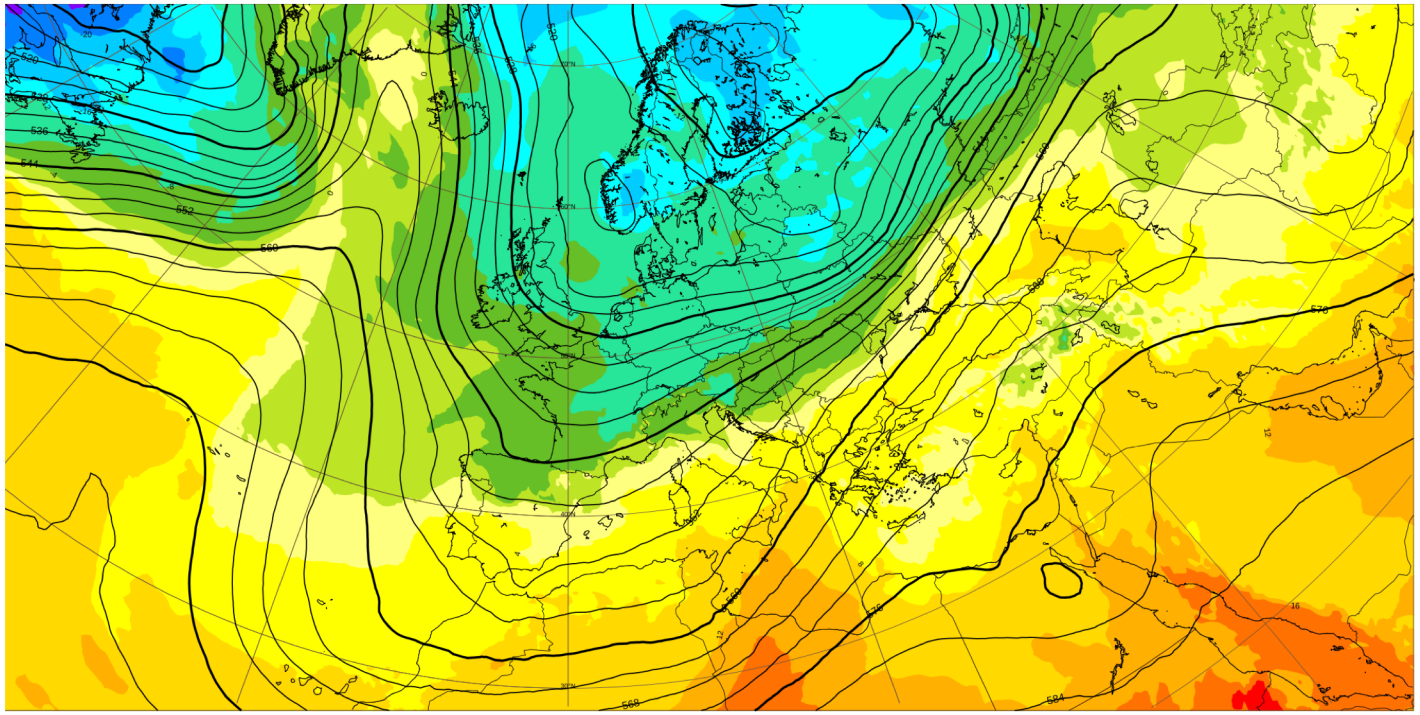


500 hPa geopotential (dm)



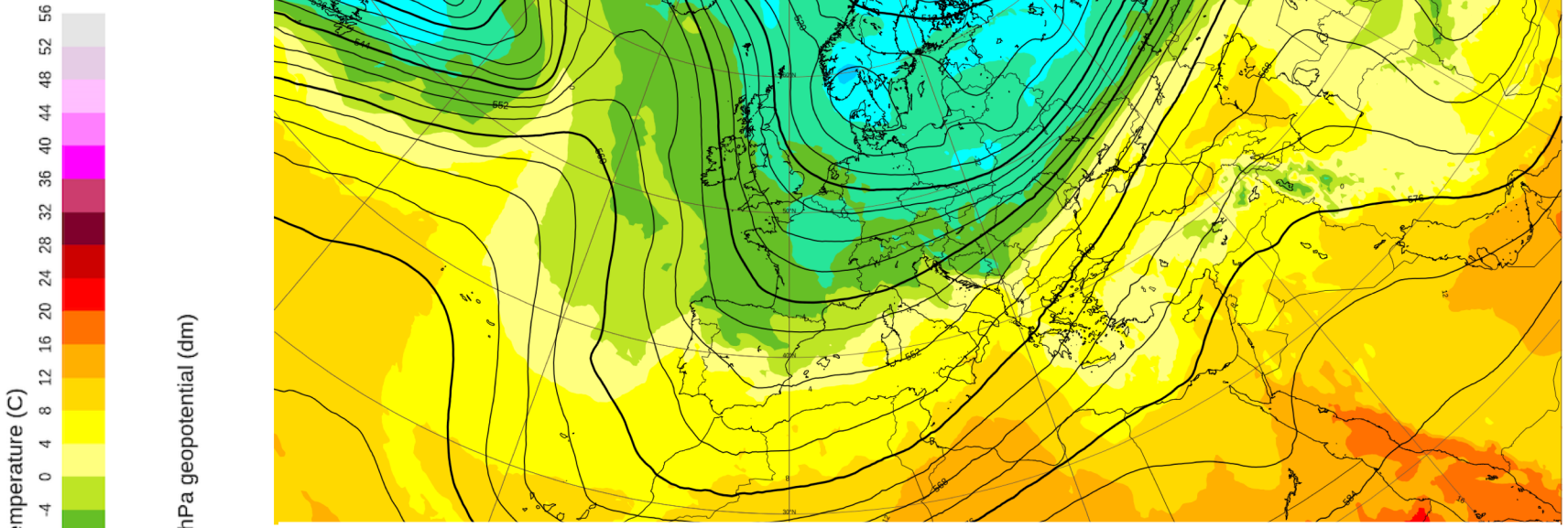
Base time: Mon 05 Jan 2026 00 UTC Valid time: Mon 05 Jan 2026 00 UTC (+0h) Area : Europe

HRES



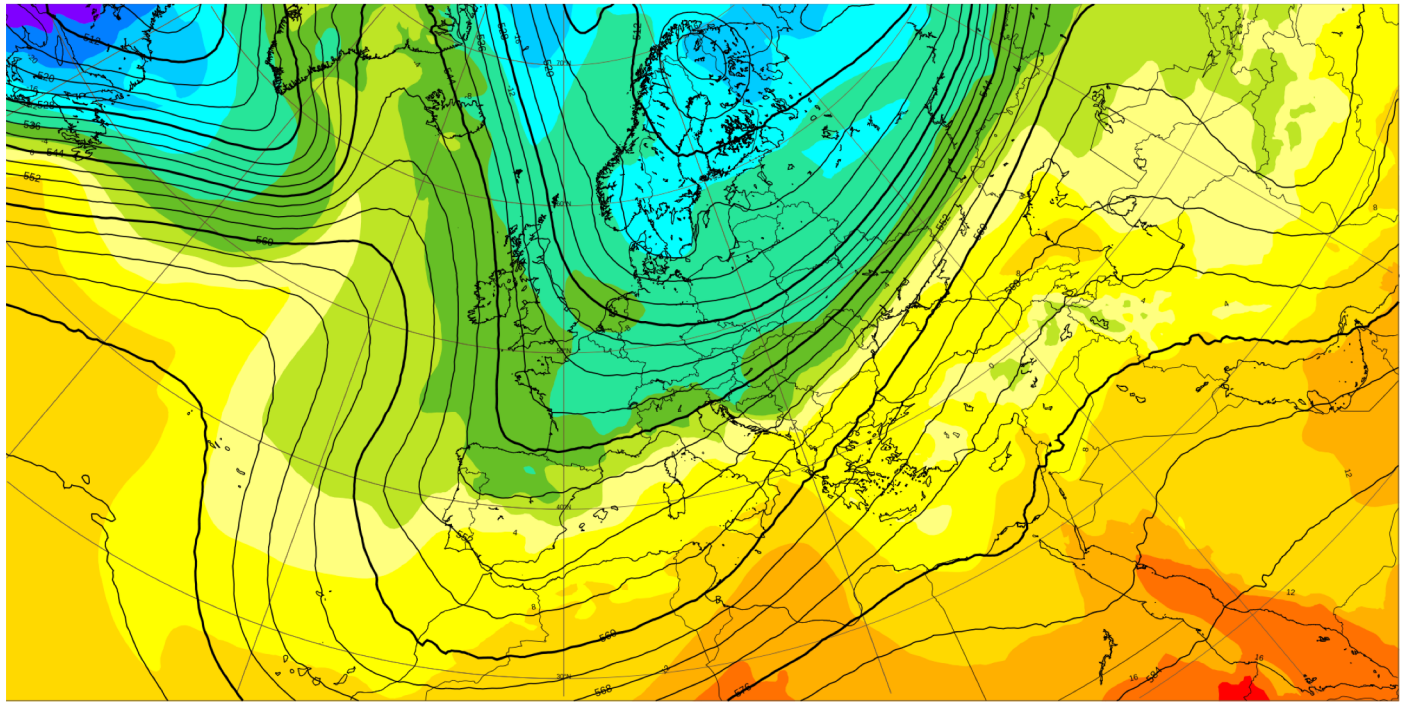
ECMWF HRES

Base time: Wed 31 Dec 2025 00 UTC Valid time: Mon 05 Jan 2026 00 UTC (+120h) Area : Europe



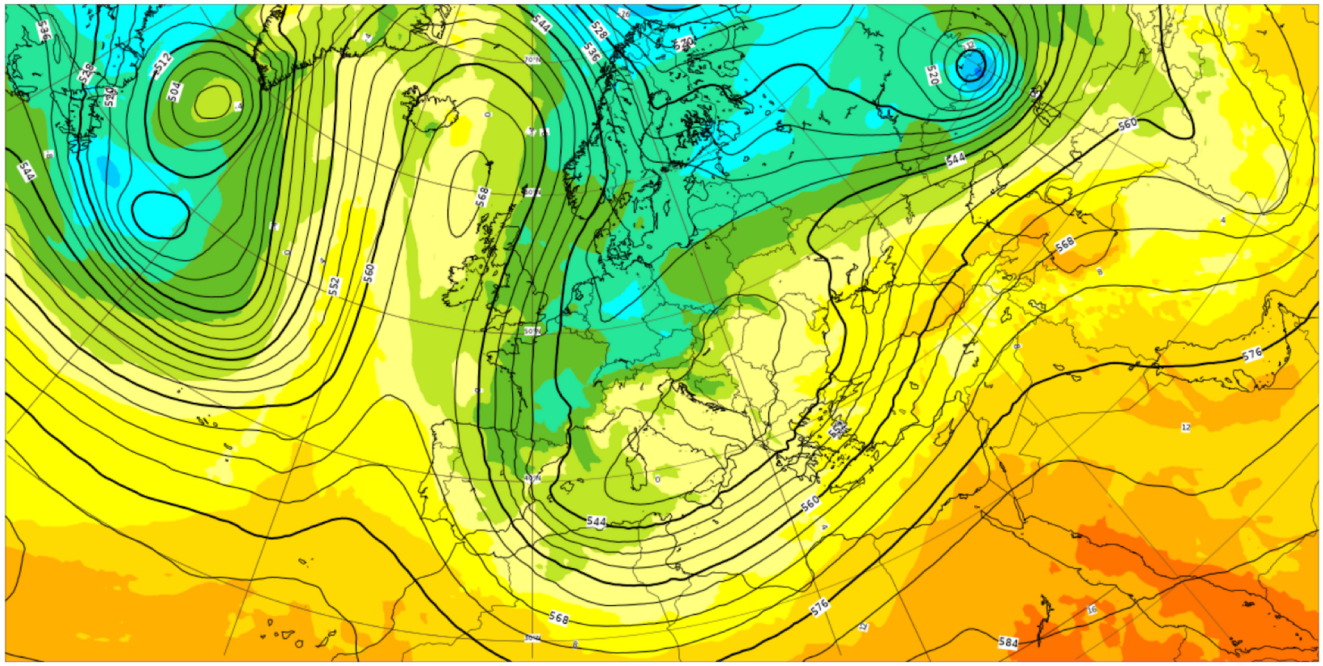
AIFS ML

Base time: Wed 31 Dec 2025 00 UTC Valid time: Mon 05 Jan 2026 00 UTC (+120h) Area : Europe



ECMWF

Base time: Tue 02 Jan 2024 00 UTC Valid time: Mon 08 Jan 2024 00 UTC (+144h) Area : Europe

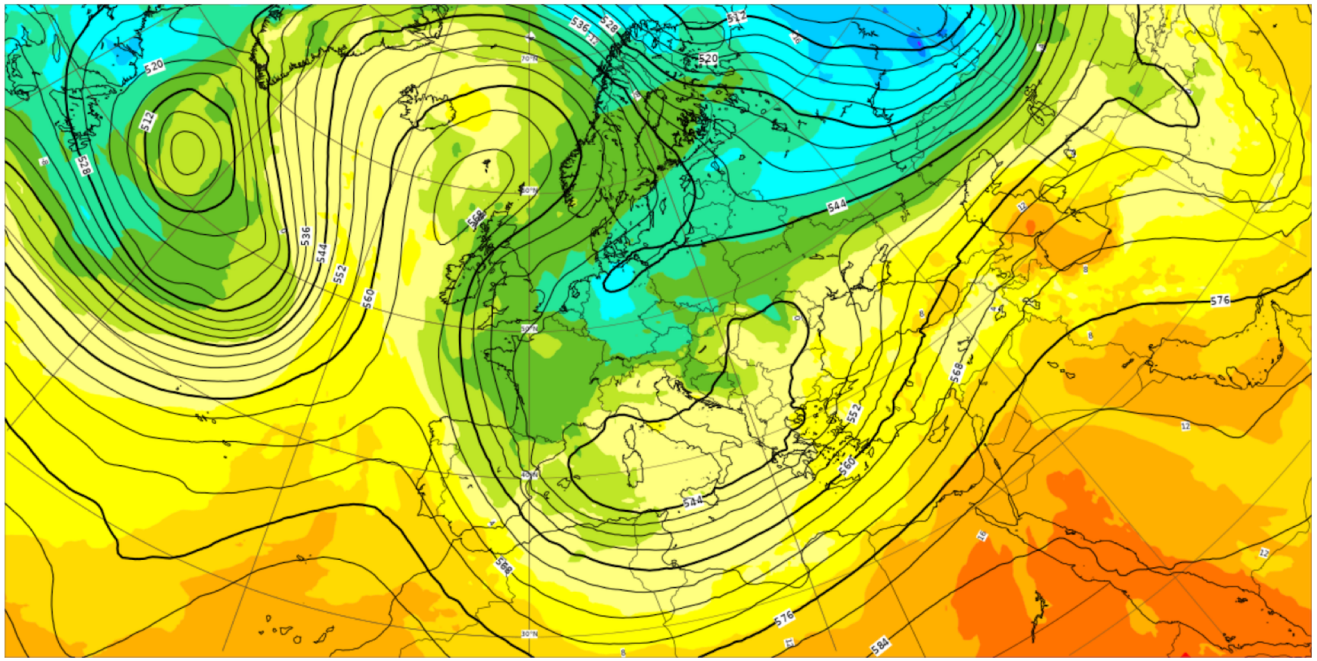


850 hPa temperature (C)

56 52 48 44 40 36 32 28 24 20 16 12 8 4 0 -4 -8 -12 -16 -20 -24 -28 -32 -36 -40 -44 -48 -52 -60 -70 -80

500 hPa geopotential (dm)

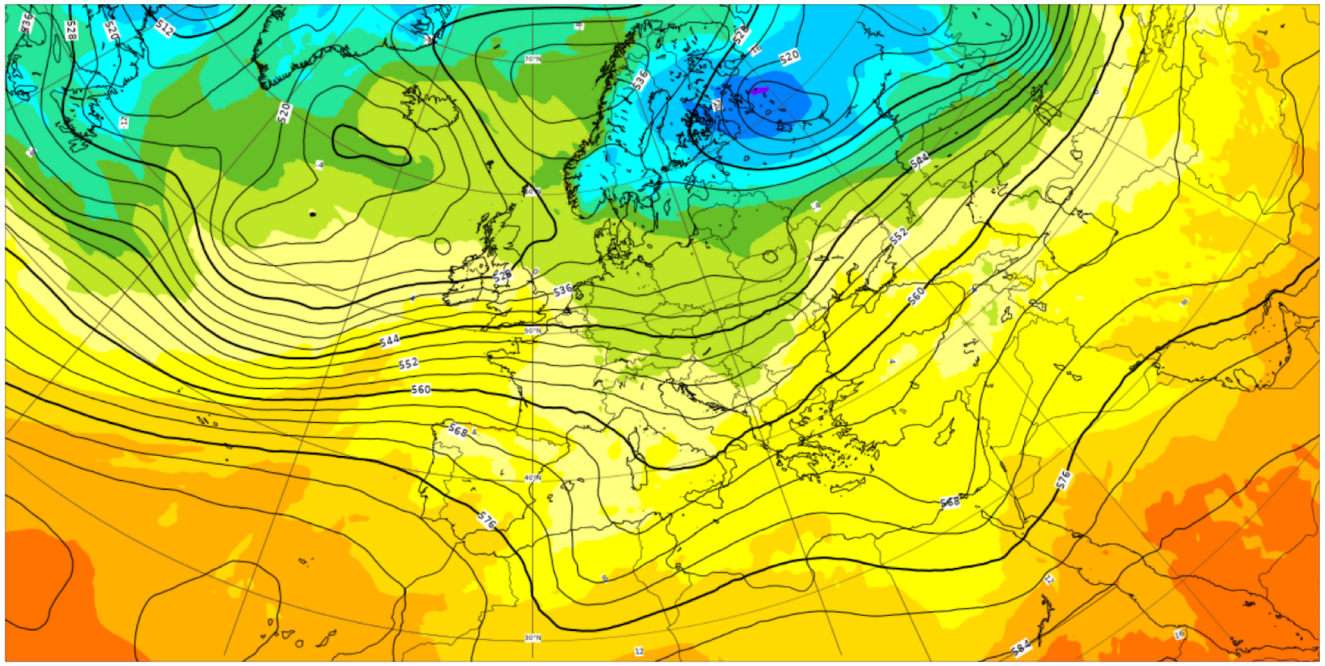
Base time: Mon 08 Jan 2024 00 UTC Valid time: Mon 08 Jan 2024 00 UTC (+0h) Area : Europe



HRES

ECMWF

Base time: Tue 02 Jan 2024 00 UTC Valid time: Tue 02 Jan 2024 00 UTC (+0h) Area : Europe

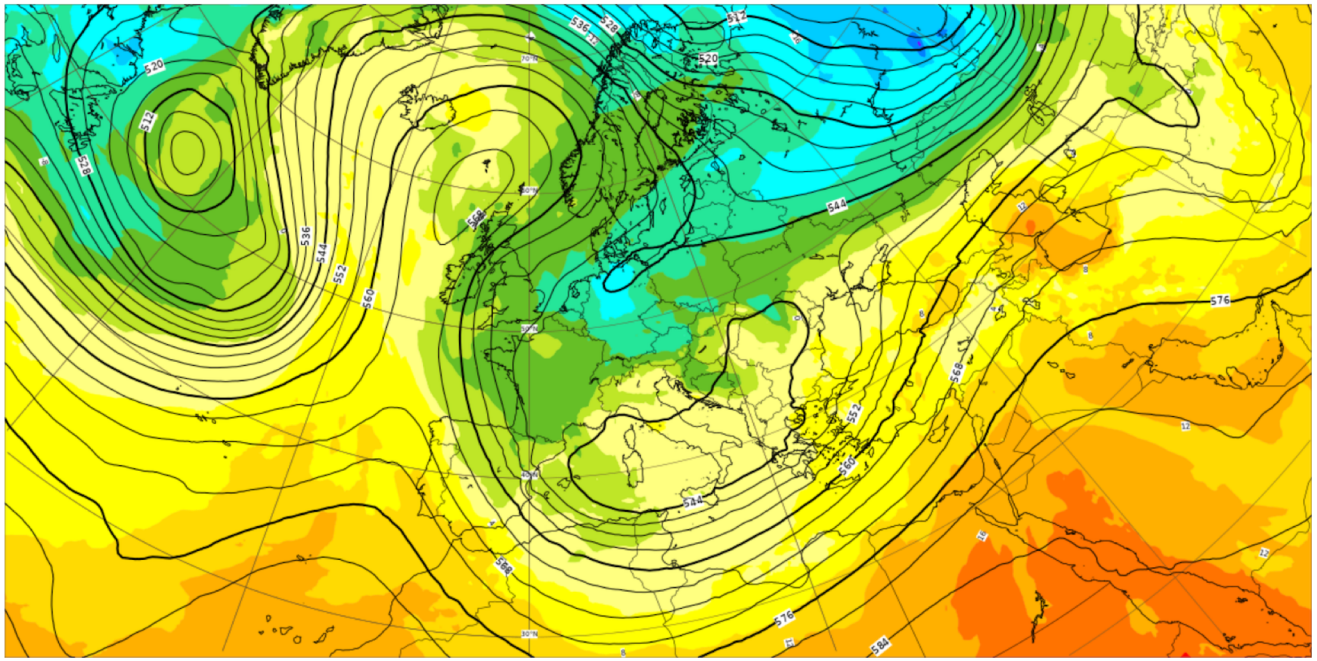


850 hPa temperature (C)

56 52 48 44 40 36 32 28 24 20 16 12 8 4 0 -4 -8 -12 -16 -20 -24 -28 -32 -36 -40 -44 -48 -52 -60 -70 -80

500 hPa geopotential (dm)

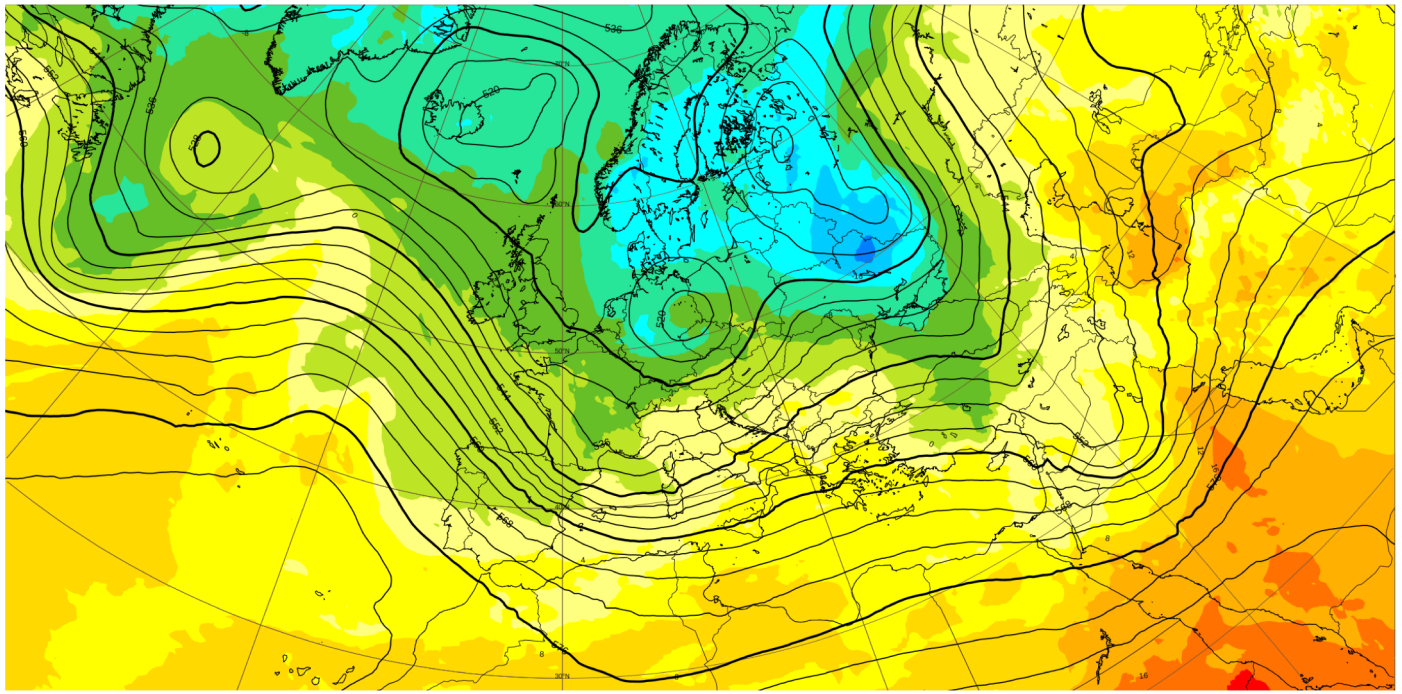
Base time: Mon 08 Jan 2024 00 UTC Valid time: Mon 08 Jan 2024 00 UTC (+0h) Area : Europe



HRES

ECMWF

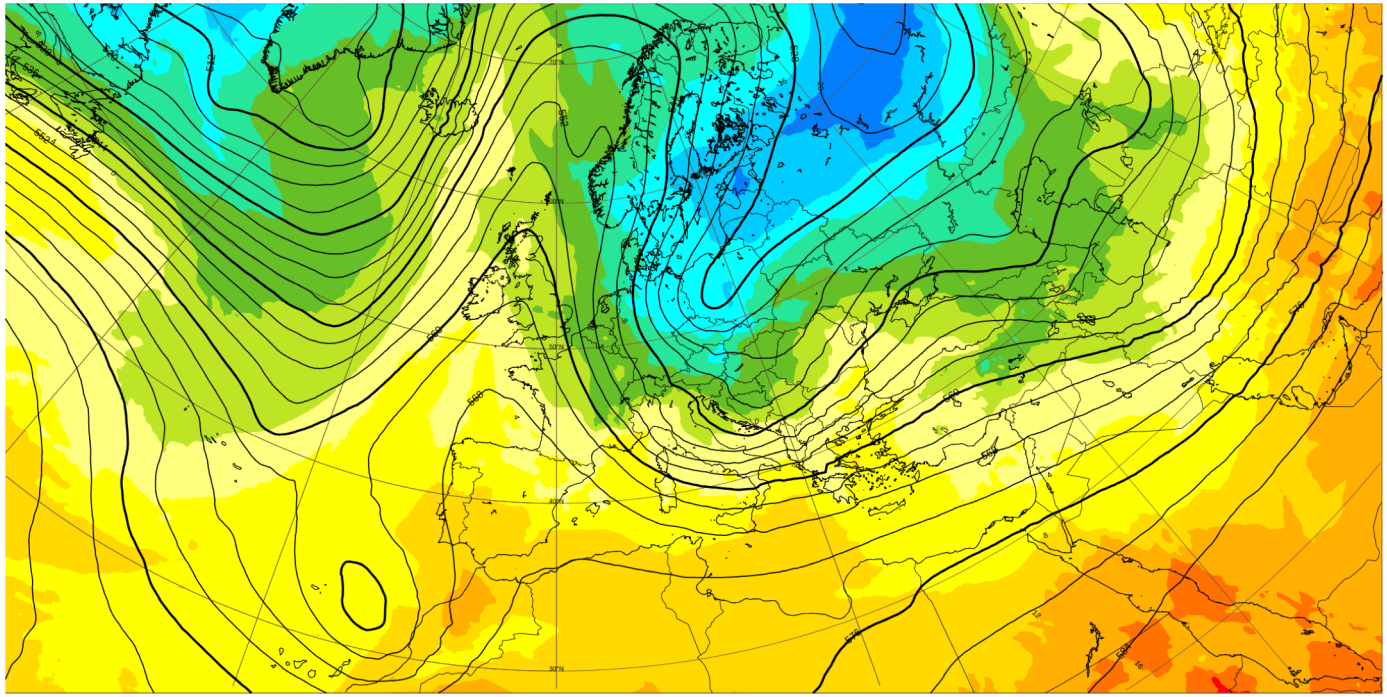
Base time: Mon 05 Jan 2026 00 UTC Valid time: Sat 10 Jan 2026 00 UTC (+120h) Area : Europe



Forecasts

5-day range

Base time: Mon 05 Jan 2026 00 UTC Valid time: Thu 15 Jan 2026 00 UTC (+240h) Area : Europe



10-day range

HRES

850 hPa temperature (C)

56 52 48 44 40 36 32 28 24 20 16 12 8 4 0 -4 -8 -12 -16 -20 -24 -28 -32 -36 -40 -44 -48 -52 -60 -70 -80

500 hPa geopotential (dm)

ECMWF

Base time: Mon 05 Jan 2026 00 UTC Valid time: Sat 10 Jan 2026 00 UTC (+120h) Area : Europe

Forecasts

5-day range

Base time: Mon 05 Jan 2026 00 UTC Valid time: Thu 15 Jan 2026 00 UTC (+240h) Area : Europe

10-day range

HRES

Snow depth (cm)

10000 500 200 100 70 50 40 30 20 15 10 5 2 1

Sea ice cover as fraction of a grid box ((0-1))

1 0.99 0.9 0.5 0.1

Ensemble forecast

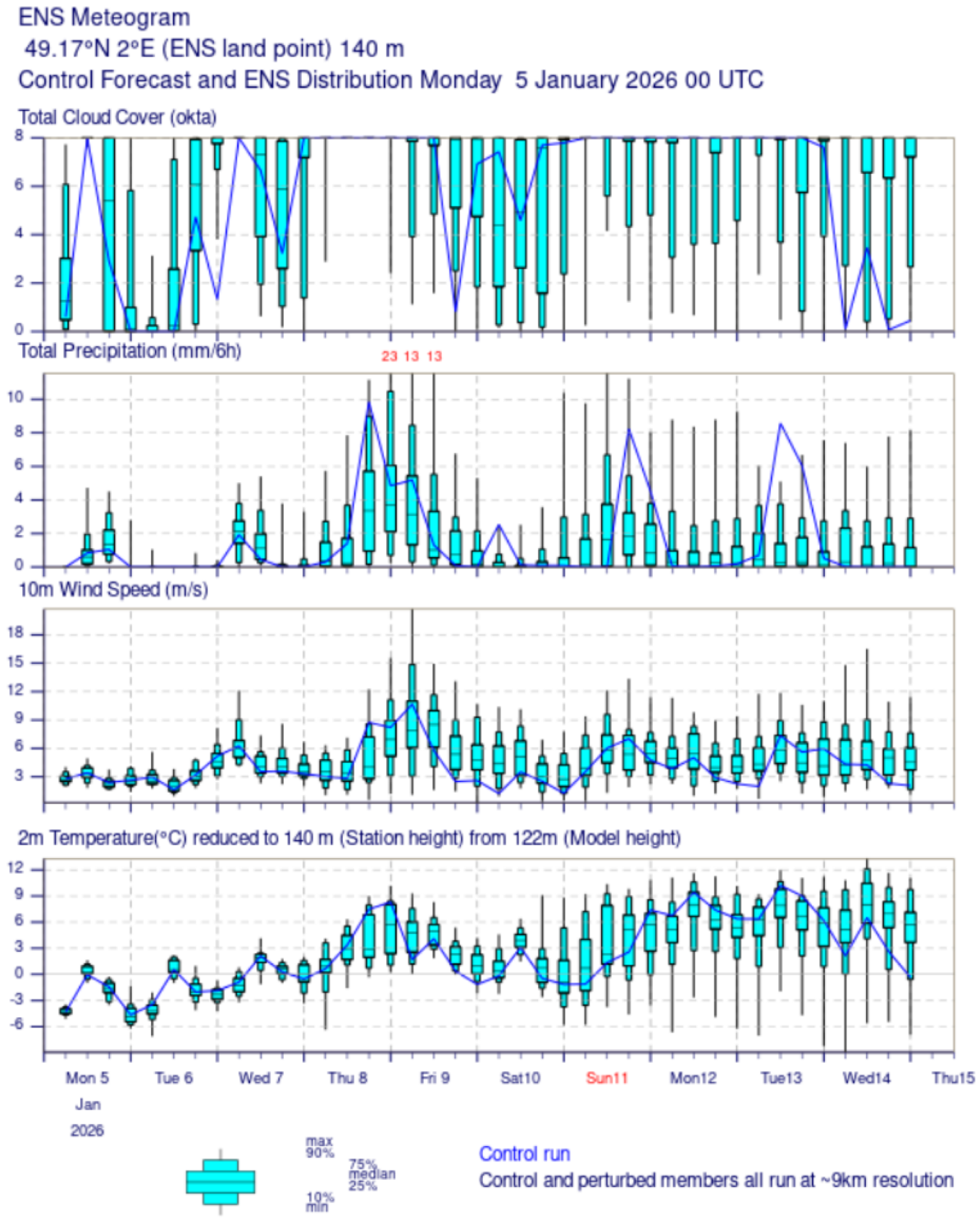
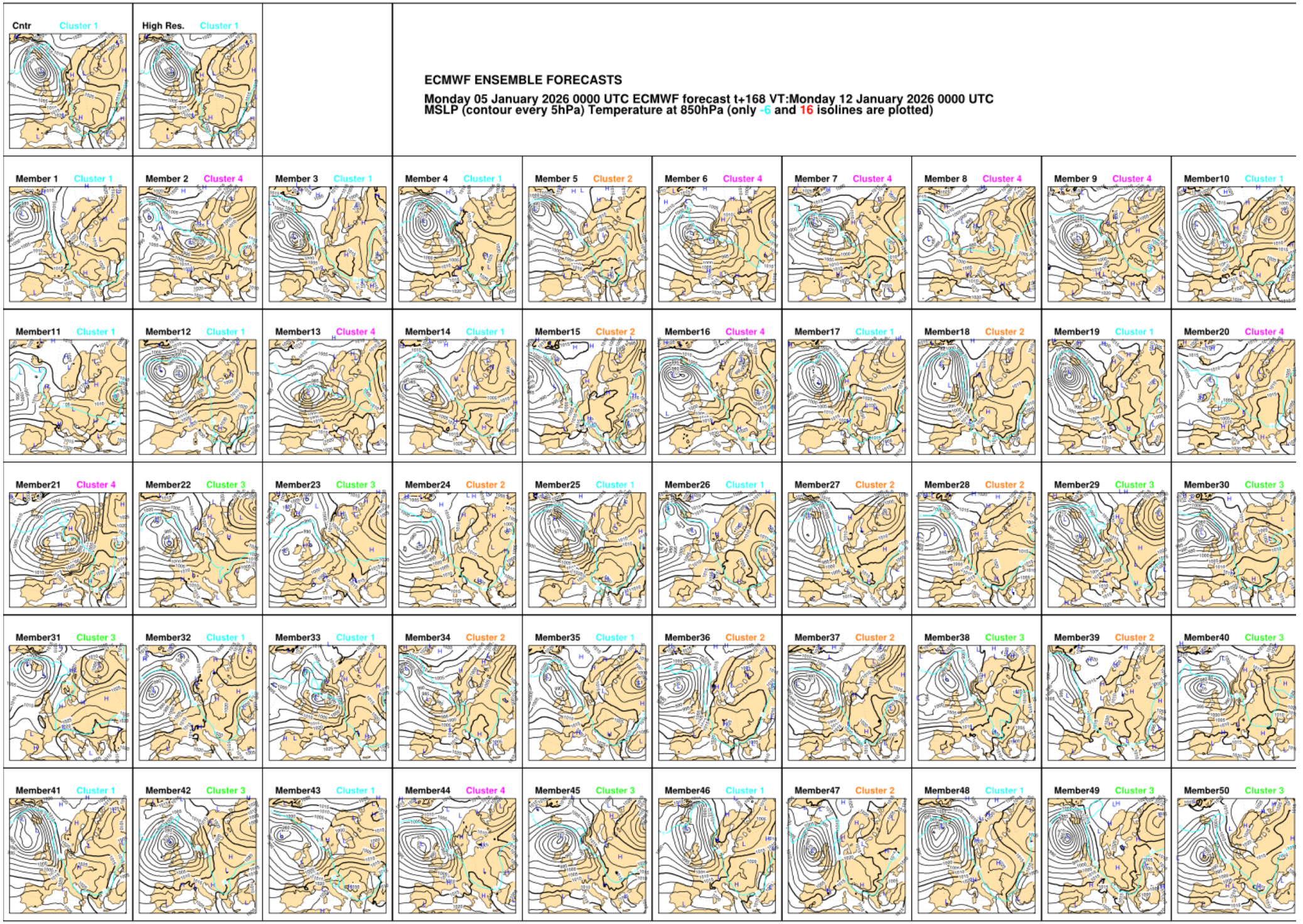


Fig. 1: Members of day 7 forecast of 500 hPa geopotential height for the ensemble originated from 25 January 1993.

Figure 6 Hurricane Katrina mean-sea-level-pressure (MSLP) analysis for 12 UTC of 29 August 2005 and t+84h high-resolution and EPS forecasts started at 00 UTC of 26 August:

1st row: *1st panel: MSLP analysis for 12 UTC of 29 Aug*
2nd panel: MSLP t+84h $T_L511L60$ forecast started at 00 UTC of 26 Aug
3rd panel: MSLP t+84h EPS-control $T_L255L40$ forecast started at 00 UTC of 26 Aug
Other rows: 50 EPS-perturbed $T_L255L40$ forecast started at 00 UTC of 26 Aug.

The contour interval is 5 hPa, with shading patters for MSLP values lower than 990 hPa.


Cntr Cluster 1
High Res. Cluster 1
ECMWF ENSEMBLE FORECASTS
Monday 05 January 2026 0000 UTC ECMWF forecast t+168 VT:Monday 12 January 2026 0000 UTC
MSLP (contour every 5hPa) Temperature at 850hPa (only -6 and 16 isolines are plotted)
Member 1 Cluster 1
Member 2 Cluster 4
Member 3 Cluster 1
Member 4 Cluster 1
Member 5 Cluster 2
Member 6 Cluster 4
Member 7 Cluster 4
Member 8 Cluster 4
Member 9 Cluster 4
Member10 Cluster 1
Member11 Cluster 1
Member12 Cluster 1
Member13 Cluster 4
Member14 Cluster 1
Member15 Cluster 2
Member16 Cluster 4
Member17 Cluster 1
Member18 Cluster 2
Member19 Cluster 1
Member20 Cluster 4
Member21 Cluster 4
Member22 Cluster 3
Member23 Cluster 3
Member24 Cluster 2
Member25 Cluster 1
Member26 Cluster 1
Member27 Cluster 2
Member28 Cluster 2
Member29 Cluster 3
Member30 Cluster 3
Member31 Cluster 3
Member32 Cluster 1
Member33 Cluster 1
Member34 Cluster 2
Member35 Cluster 1
Member36 Cluster 2
Member37 Cluster 2
Member38 Cluster 3
Member39 Cluster 2
Member40 Cluster 3
Member41 Cluster 1
Member42 Cluster 3
Member43 Cluster 1
Member44 Cluster 4
Member45 Cluster 3
Member46 Cluster 1
Member47 Cluster 2
Member48 Cluster 1
Member49 Cluster 3
Member50 Cluster 3

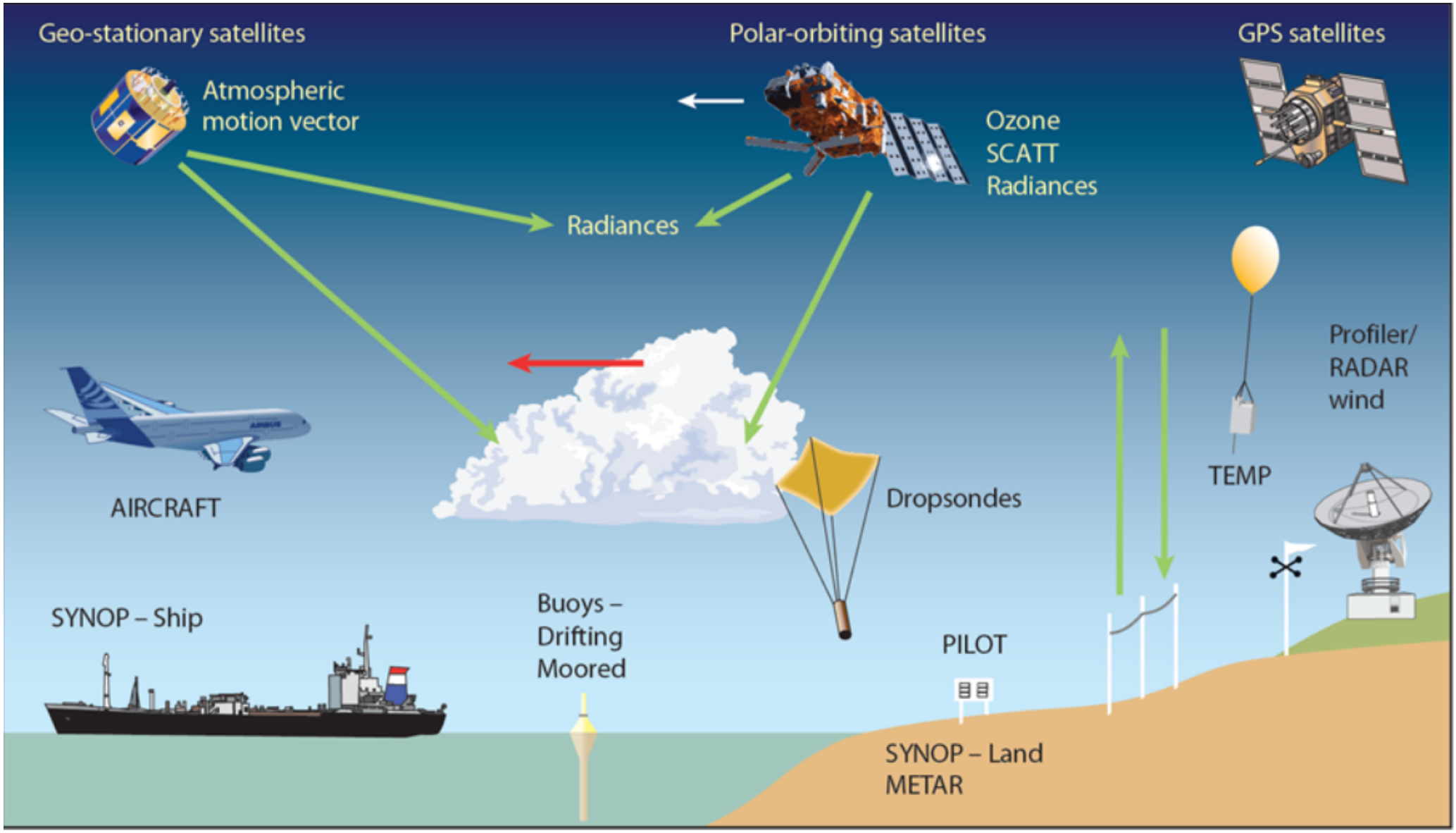
Pourquoi les météorologistes ont-ils tant de peine à prédire le temps avec quelque certitude ? Pourquoi les chutes de pluie, les tempêtes elles-mêmes nous semblent-elles arriver au hasard, de sorte que bien des gens trouvent tout naturel de prier pour avoir la pluie ou le beau temps, alors qu'ils jugeraient ridicule de demander une éclipse par une prière ? Nous voyons que les grandes perturbations se produisent généralement dans les régions où l'atmosphère est en équilibre instable. Les météorologistes voient bien que cet équilibre est instable, qu'un cyclone va naître quelque part ; mais où, ils sont hors d'état de le dire ; un dixième de degré en plus ou en moins en un point quelconque, le cyclone éclate ici et non pas là, et il étend ses ravages sur des contrées qu'il aurait épargnées. Si on avait connu ce dixième de degré, on aurait pu le savoir d'avance, mais les observations n'étaient ni assez serrées, ni assez précises, et c'est pour cela que tout semble dû à l'intervention du hasard.

H. Poincaré*, Science et Méthode,* Paris, 1908

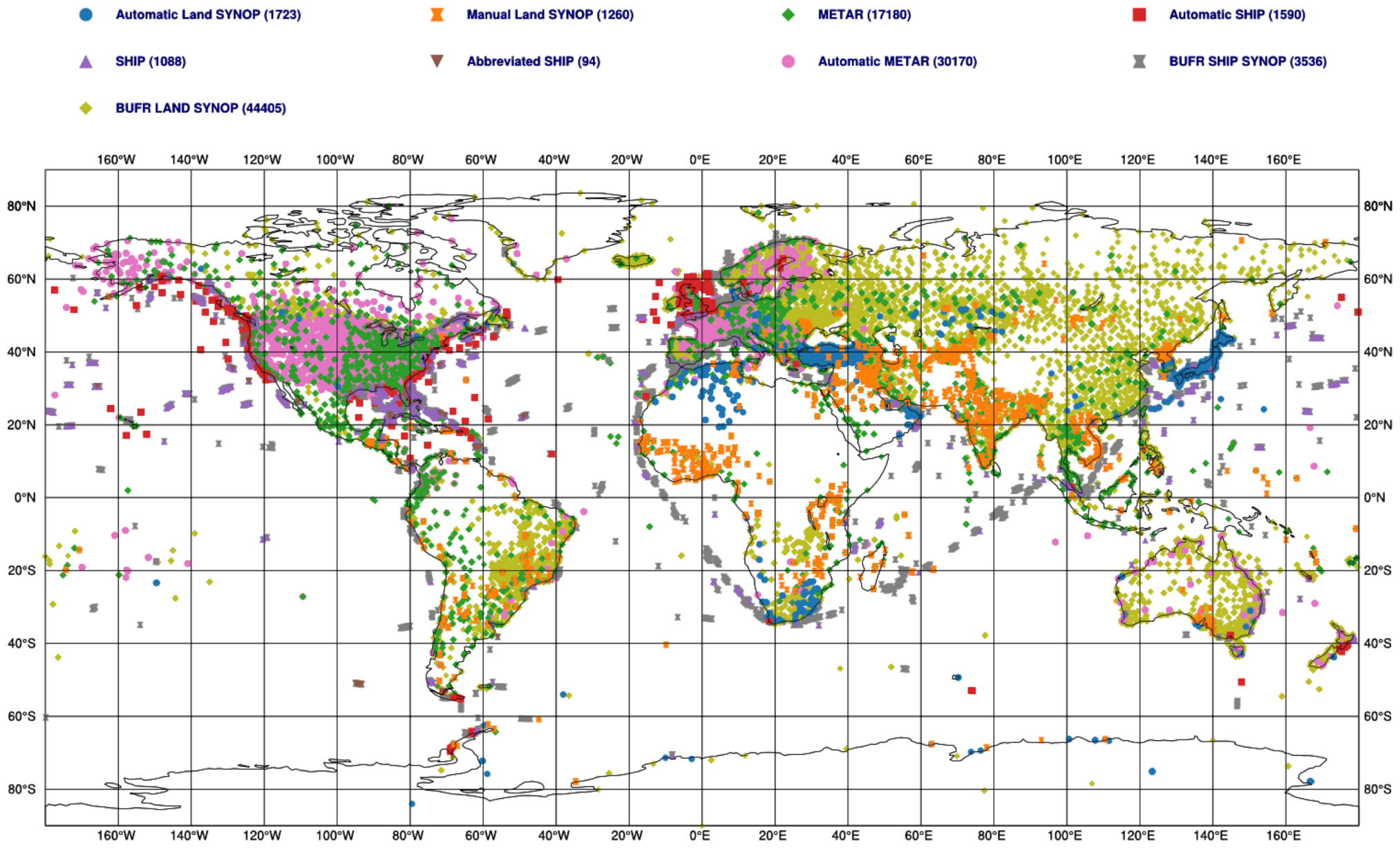
Why have meteorologists such difficulty in predicting the weather with any certainty? Why is it that showers and even storms seem to come by chance, so that many people think it quite natural to pray for rain or fine weather, though they would consider it ridiculous to ask for an eclipse by prayer? We see that great disturbances are generally produced in regions where the atmosphere is in unstable equilibrium. The meteorologists see very well that the equilibrium is unstable, that a cyclone will be formed somewhere, but exactly where they are not in a position to say; a tenth of a degree more or less at any given point, and the cyclone will burst here and not there, and extend its ravages over districts it would otherwise have spared. If they had been aware of this tenth of a degree they could have known it beforehand, but the observations were neither sufficiently comprehensive nor sufficiently precise, and that is the reason why it all seems due to the intervention of chance.

H. Poincaré, *Science et Méthode*, Paris, 1908

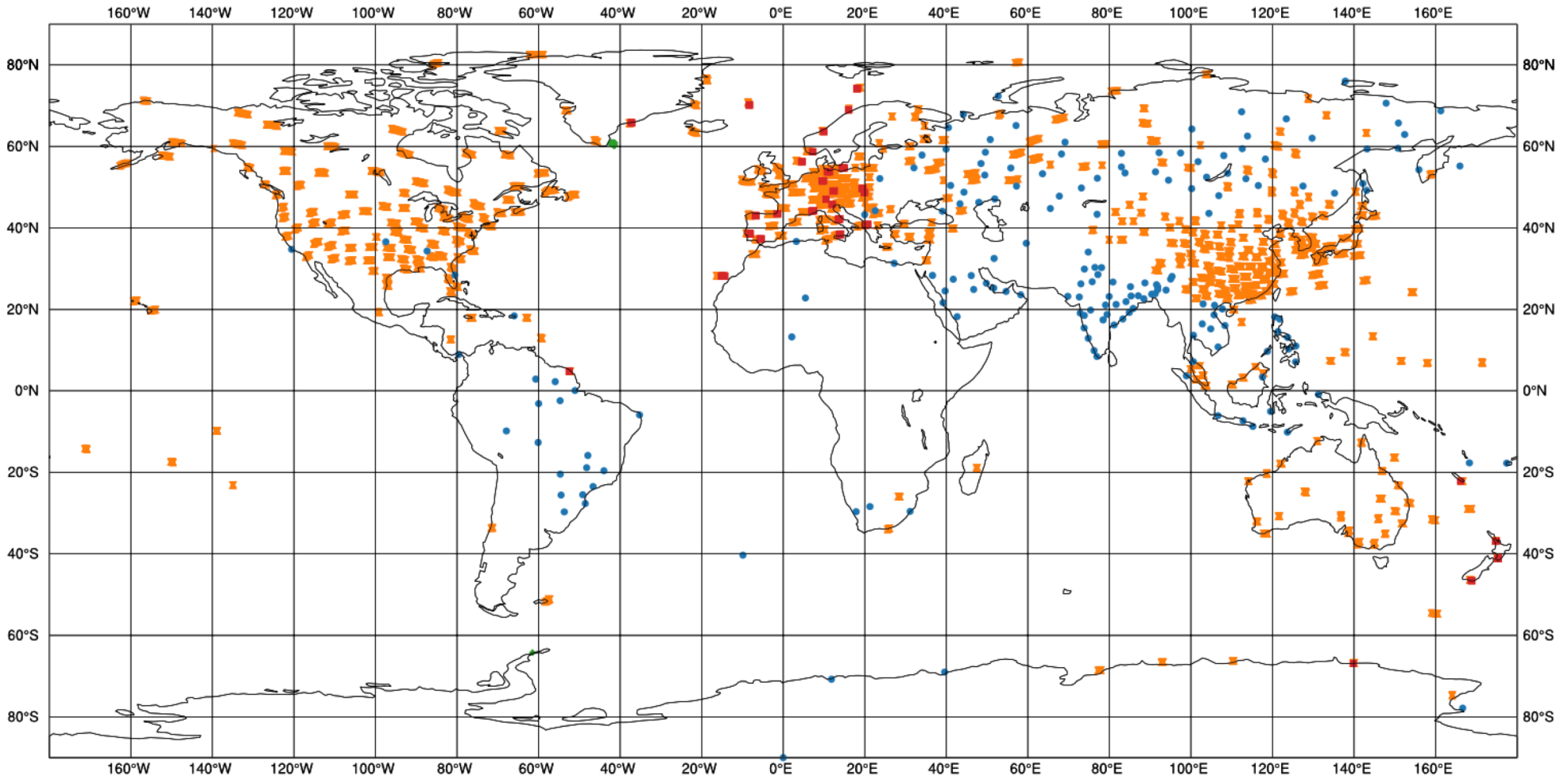
(English transl. by F. Maitland, *Science and Method*,
T. Nelson and Sons, London, 1914)


Geo-stationary satellites
Atmospheric motion vector
Polar-orbiting satellites
Ozone SCATT Radiances
GPS satellites
Radiances
Profiler/ RADAR wind
TEMP
AIRCRAFT
Dropsondes
SYNOP – Ship
Buoys – Drifting Moored
PILOT
SYNOP – Land METAR

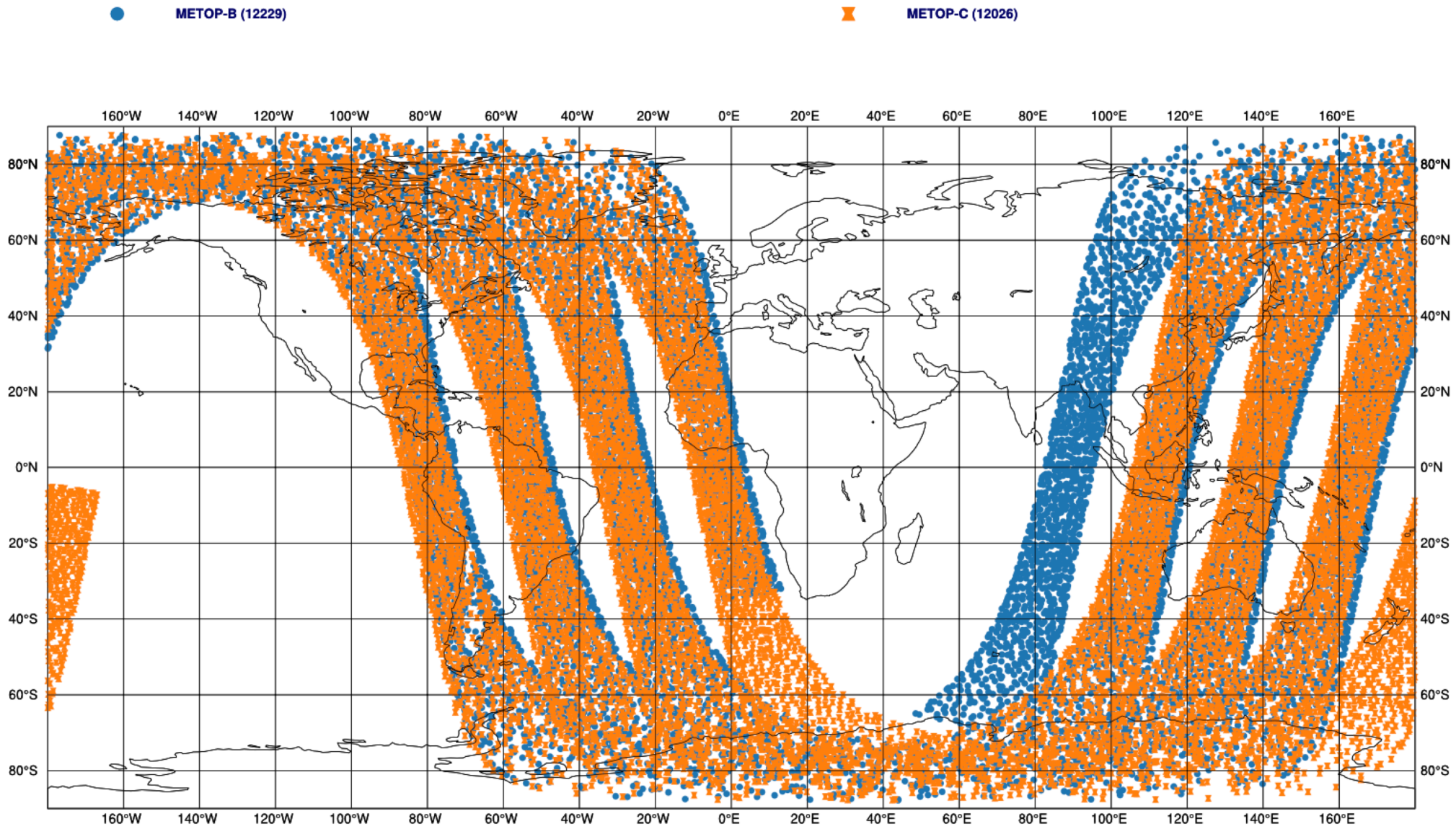
ECMWF data coverage (used observations) - SYNOP-SHIP-METAR
2026010421 to 2026010503
Total number of obs = 101046



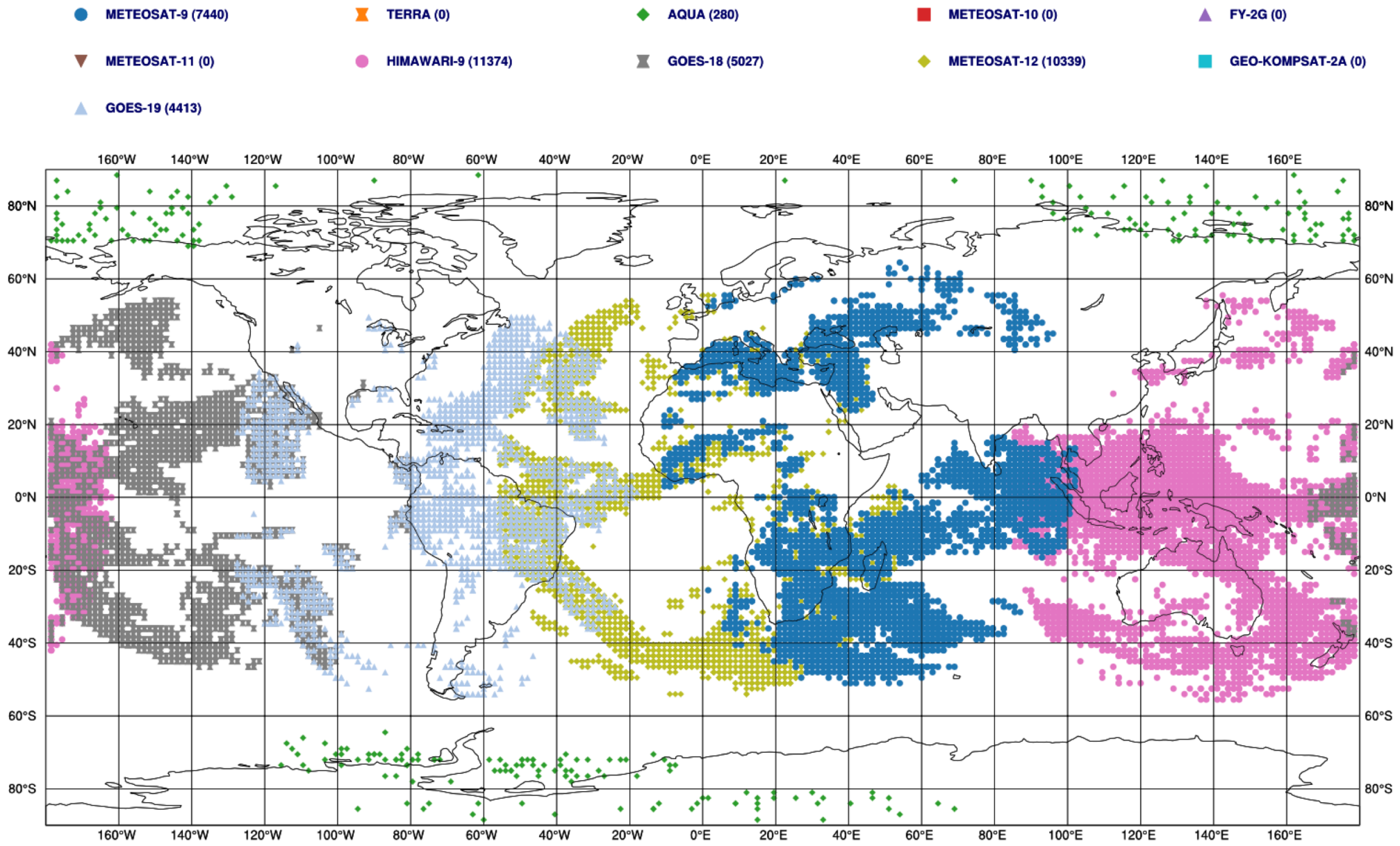
ECMWF data coverage (used observations) - RADIOSONDE
2026010421 to 2026010503
Total number of obs = 622



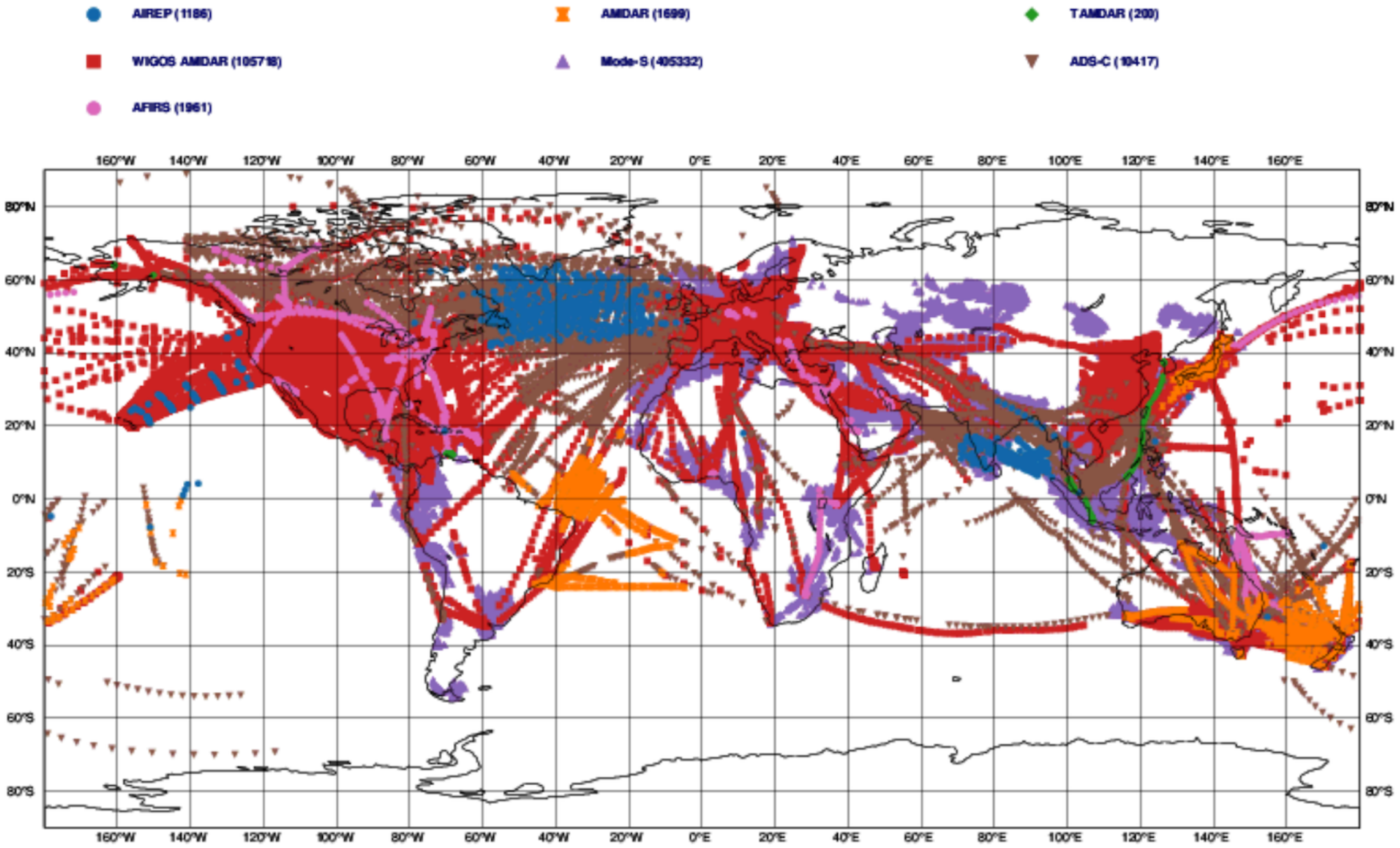
ECMWF data coverage (used observations) - IASI
2026010421 to 2026010503
Total number of obs = 24255



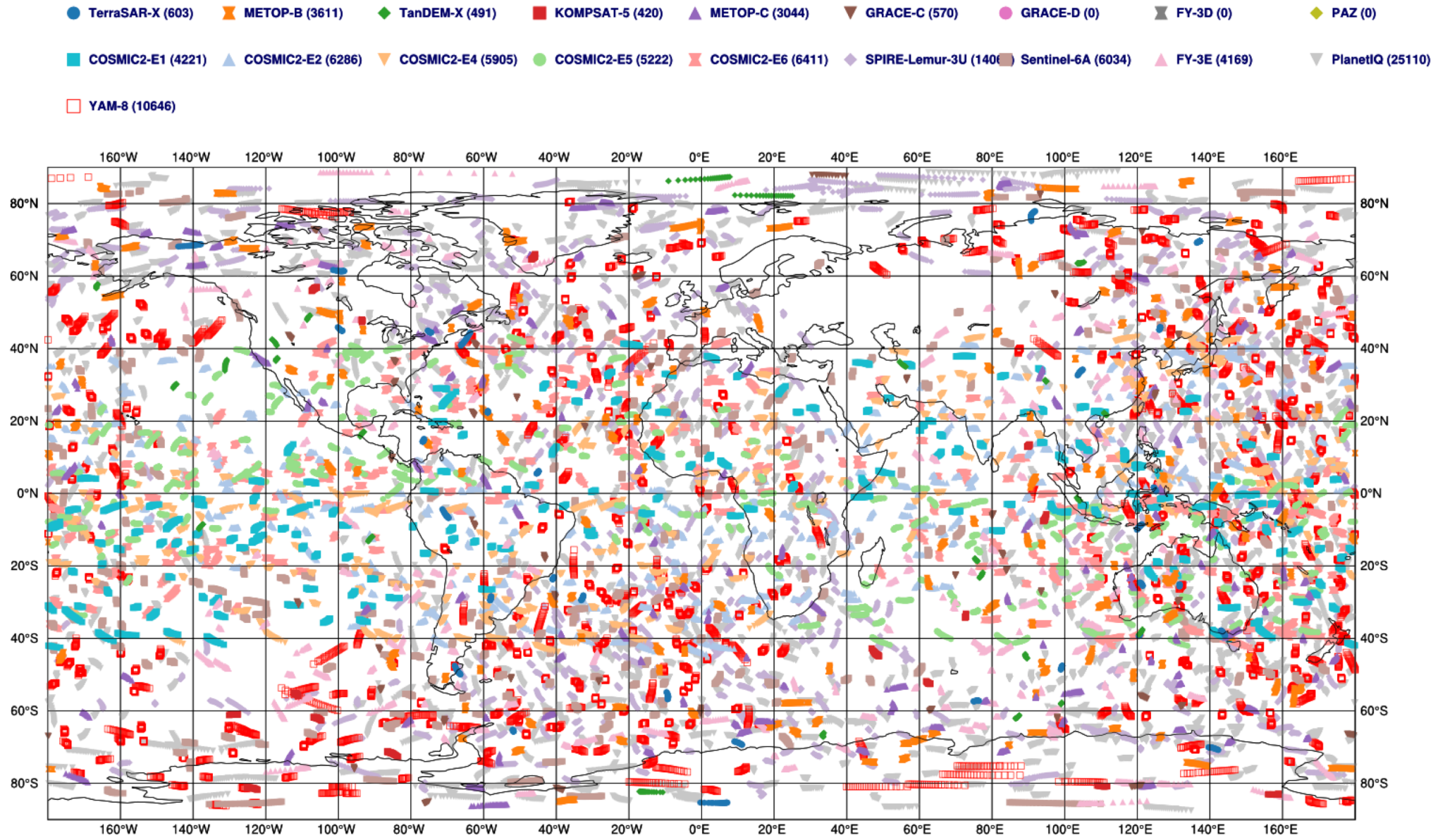
ECMWF data coverage (used observations) - AMV WV
2026010421 to 2026010503
Total number of obs = 38873



ECMWF data coverage (used observations) - AIRCRAFT
2026010421 to 2026010503
Total number of obs = 526513



ECMWF data coverage (used observations) - GPSRO
2026010421 to 2026010503
Total number of obs = 96810



As of 2023

We receive 800 million observations daily, and 60 million quality-controlled observations are available daily for use in the Integrated Forecasting System (IFS); the vast majority of these are satellite measurements, but ECMWF also benefits from all available observations from non-satellite sources, including surface-based and aircraft reports.

- *Synoptic* observations (ground observations, radiosonde observations), performed simultaneously, by international agreement, in all meteorological stations around the world (00:00, 06:00, 12:00, 18:00 UTC), and are in practice concentrated over continents.

- *Asynoptic* observations (satellites, aircraft), performed more or less continuously in time.

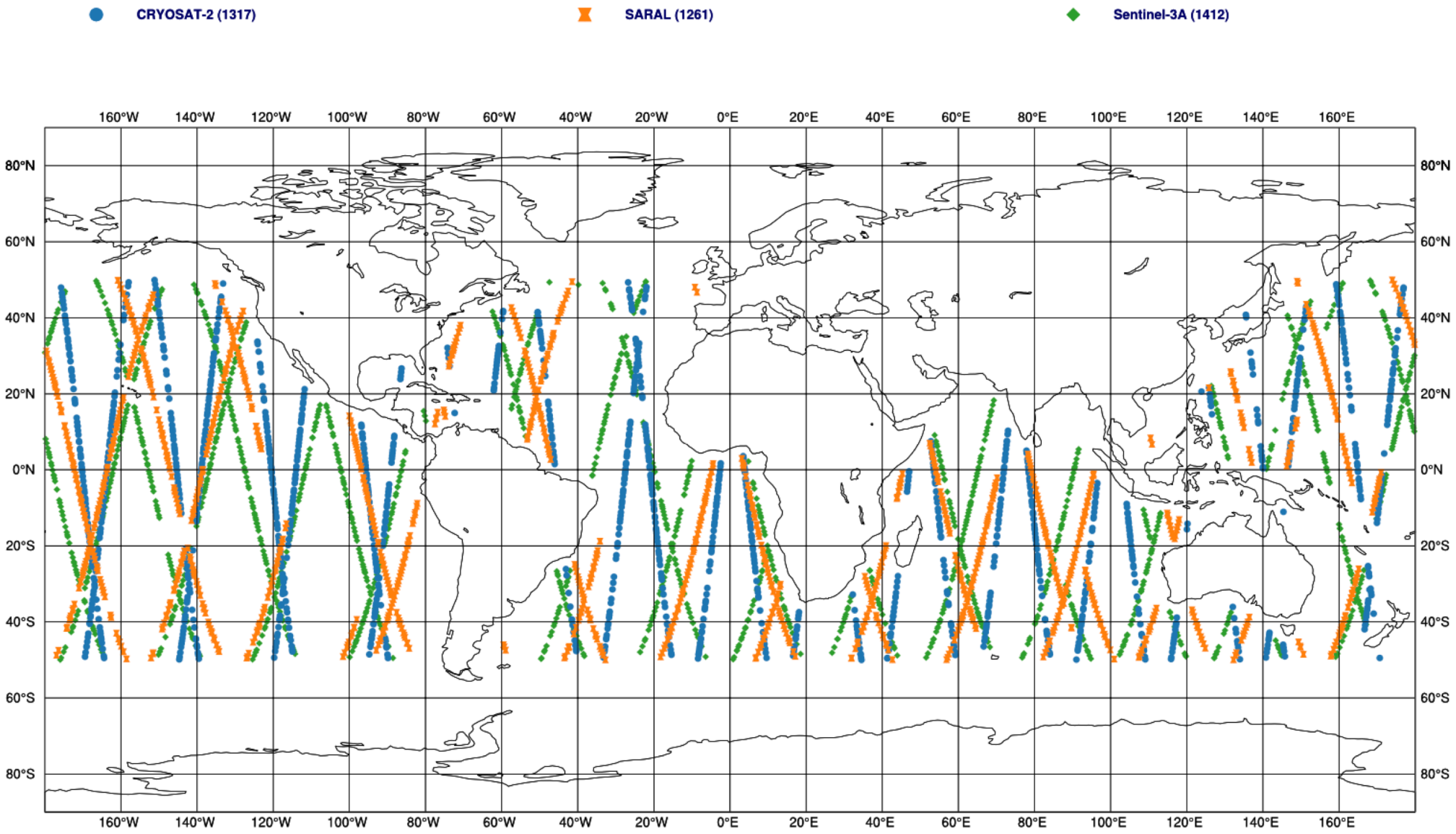
- *Direct* observations (temperature, pressure, horizontal components of the wind, moisture), which are local and bear on the variables used for describing the flow in numerical models.

- *Indirect* observations (radiometric observations, …), which bear on some more or less complex combination (most often, a one-dimensional spatial integral) of variables used for for describing the flow

$$y = H(\boldsymbol{x})$$

H : *observation operator (*for instance, radiative transfer equation)

ECMWF data coverage (used observations) - SEA LEVEL ANOMALY
20260103 00
Total number of obs = 3990



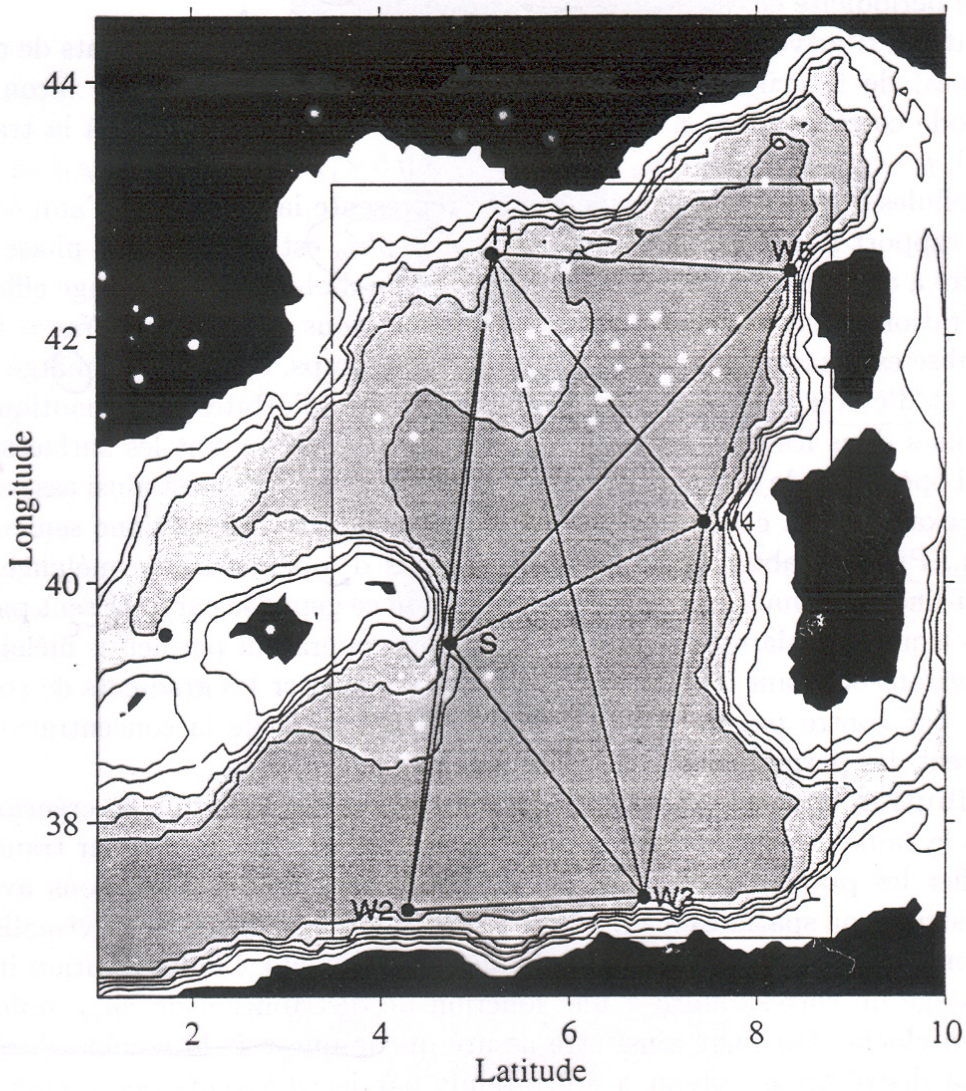


FIG. 1 – *Bassin méditerranéen occidental: réseau d'observation tomographique de l'expérience Thétis 2 et limites du domaine spatial utilisé pour les expériences numériques d'assimilation.*

E. Rémy, Doctoral Dissertation, 1999

Weather prediction, at the occasional exception of very short range (a few hours at most), is performed by using appropriate numerical models. There exist at present two forms of numerical models

- *Physical models*, built on an explicit formulation of the physical laws which govern the evolution of the state of the atmosphere.

- In recent years, models based on *Machine Learning* (or *Artificial Intelligence*) have been developed. They are built on training sets of past atmospheric states.

Physical models

Basic physical laws

- Conservation of mass

  $D\rho/Dt + \rho\,\mathrm{div}\underline{U} = 0$

- Conservation of energy

  $De/Dt - (p/\rho^2)\, D\rho/Dt = Q$

- Conservation of momentum

  $D\underline{U}/Dt + (1/\rho)\,\underline{\mathrm{grad}}p - \underline{g} + 2\,\underline{\Omega} \wedge \underline{U} = \underline{F}$

- Equation of state

  $f(p, \rho, e) = 0$ (for a perfect gas $p/\rho = rT$, $e = C_v T$)

- Conservation of mass of secondary components (water in the atmosphere, salt in the ocean, chemical species, …)

  $Dq/Dt + q\,\mathrm{div}\underline{U} = S$

These physical laws must be expressed in practice in discretized (and necessarily imperfect) form, both in space and time ⇒ *numerical model*

Parlance of the trade :

- Adiabatic and inviscid, and therefore thermodynamically reversible, processes (everything except Q, $\underline{F}$ and S) make up *'dynamics'*

- Processes described by terms Q, $\underline{F}$ and S make up *'physics'*

All presently existing physical models are built on more or less simplified forms of the general physical laws. Global numerical models, used either for large-scale meteorological prediction or for climate simulation, are at present built on the so-called *primitive equations*. Those equations rely on several approximations, the most important of which being the *hydrostatic approximation*, which expresses balance, in the vertical direction, of the gravity and pressure gradient forces. This forbids explicit description of thermal convection, which must be 'parameterized' in some appropriate way.

More and more *limited-area models* have been developed over time. They require appropriate definition of lateral boundary conditions (not a simple problem). Many of them are non-hydrostatic, and therefore allow description of convection.

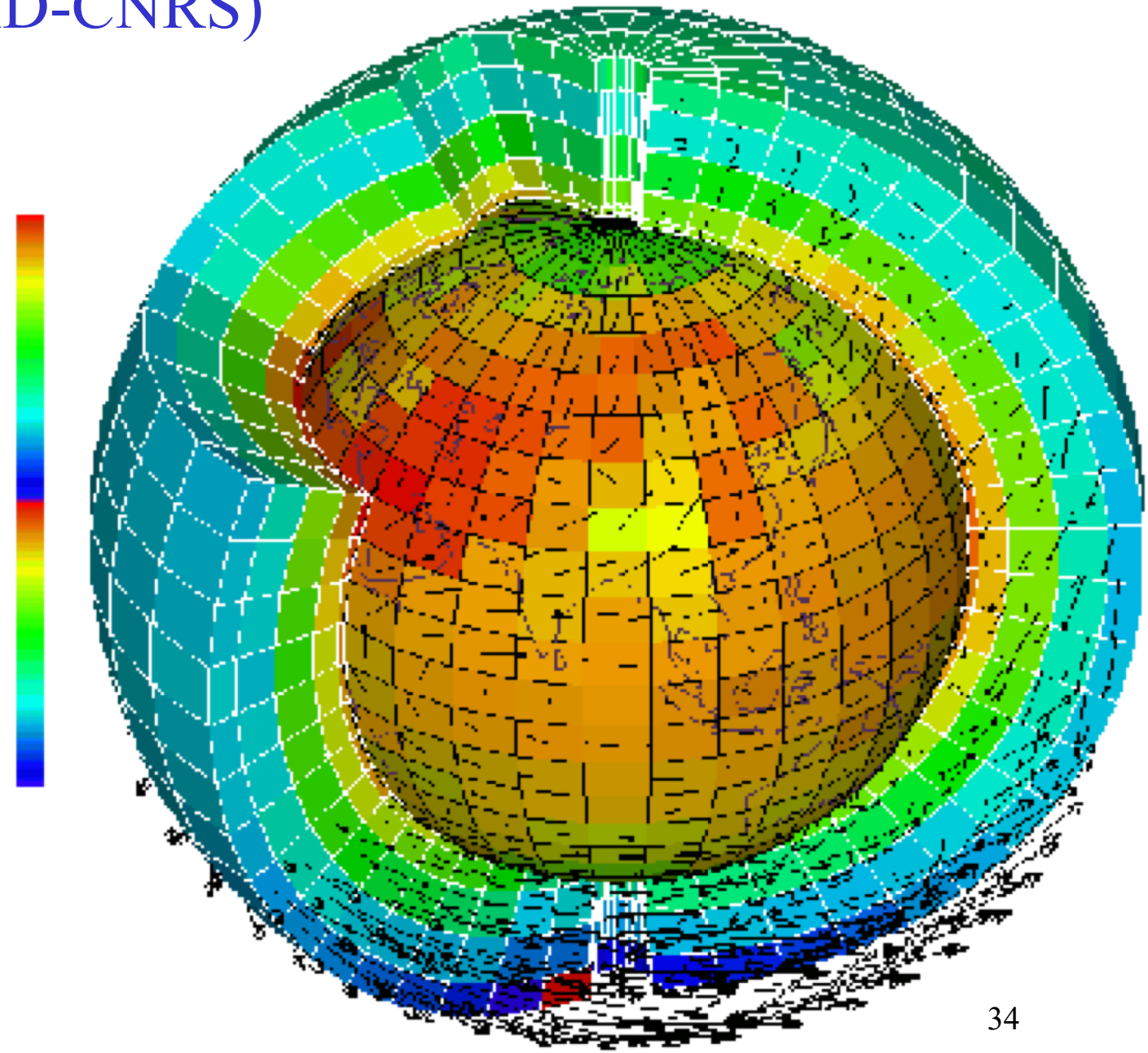
There exist at present two forms of horizontal spatial discretization

- Gridpoint discretization

- (Semi-)spectral discretization (mostly for global models, and most often only in the horizontal direction)

Finite element discretization, which is very common in many forms of numerical modelling, is rarely used for modelling of the atmosphere. It is more frequently used for oceanic modelling, where it allows to take account of the complicated geometry of coast-lines.

Schematic of a gridpoint atmospheric model (L. Fairhead /LMD-CNRS)



In gridpoint models, meteorological fields are defined by values at the nodes of the grid. Spatial and temporal derivatives are expressed by finite differences.

In spectral models, fields are defined by the coefficients of their expansion along a prescribed set of basic functions. In the case of global meteorological models, those basic functions are the *spherical harmonics* (eigenfunctions of the laplacian at the surface of the sphere).

Modèles (semi-)spectraux

$$T(\mu=\sin(\text{latitude}),\ \lambda=\text{longitude}) = \sum_{\substack{0\le n<\infty \\ -n\le m\le n}} T_n^m Y_n^m(\mu,\lambda)$$

où les $Y_n^m(\mu,\lambda)$ sont les *harmoniques sphériques*

$$Y_n^m(\mu,\lambda) \propto P_n^m(\mu)\exp(im\lambda)$$

$P_n^m(\mu)$ est la *fonction de Legendre* de deuxième espèce

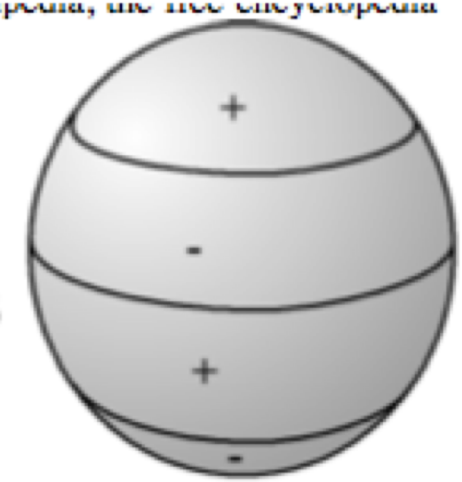
$$P_n^m(\mu) \propto (1-\mu^2)^{\frac{m}{2}} \frac{d^{n+m}}{d\mu^{n+m}}(\mu^2-1)^n$$

n et m sont respectivement le *degré* et l'*ordre* de l'harmonique $Y_n^m(\mu,\lambda)$

$n = 0, 1, \ldots$ $\quad - n \le m \le n$

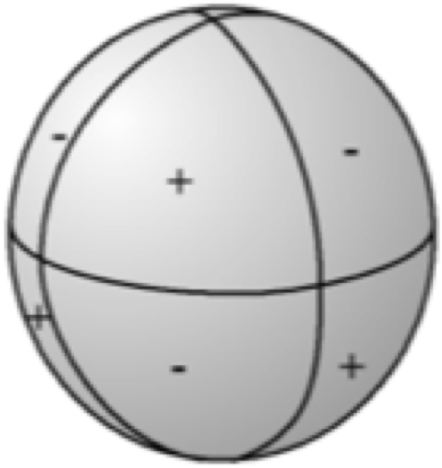
From Wikipedia, the free encyclopedia

$l = 3$
$m = 0$
$l\text{-}m = 3$

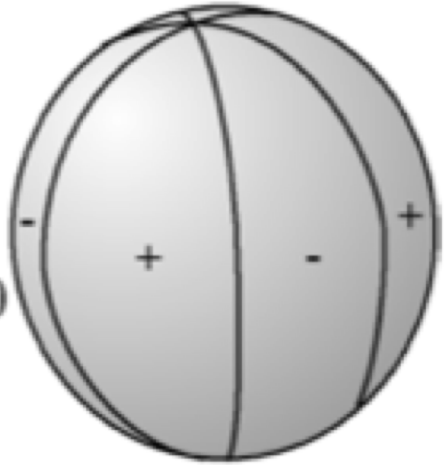


$l = 3$
$m = 1$
$l\text{-}m = 2$

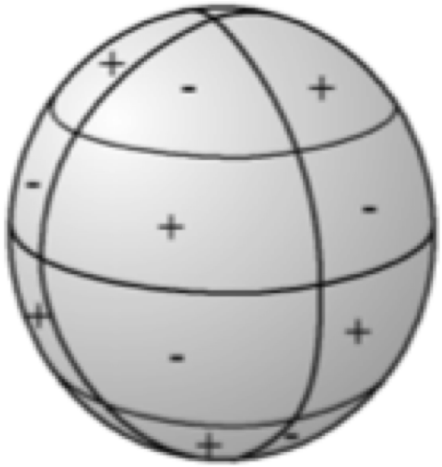
$l = 3$
$m = 2$
$l\text{-}m = 1$



$l = 3$
$m = 3$
$l\text{-}m = 0$



$l = 5$
$m = 2$
$l\text{-}m = 3$



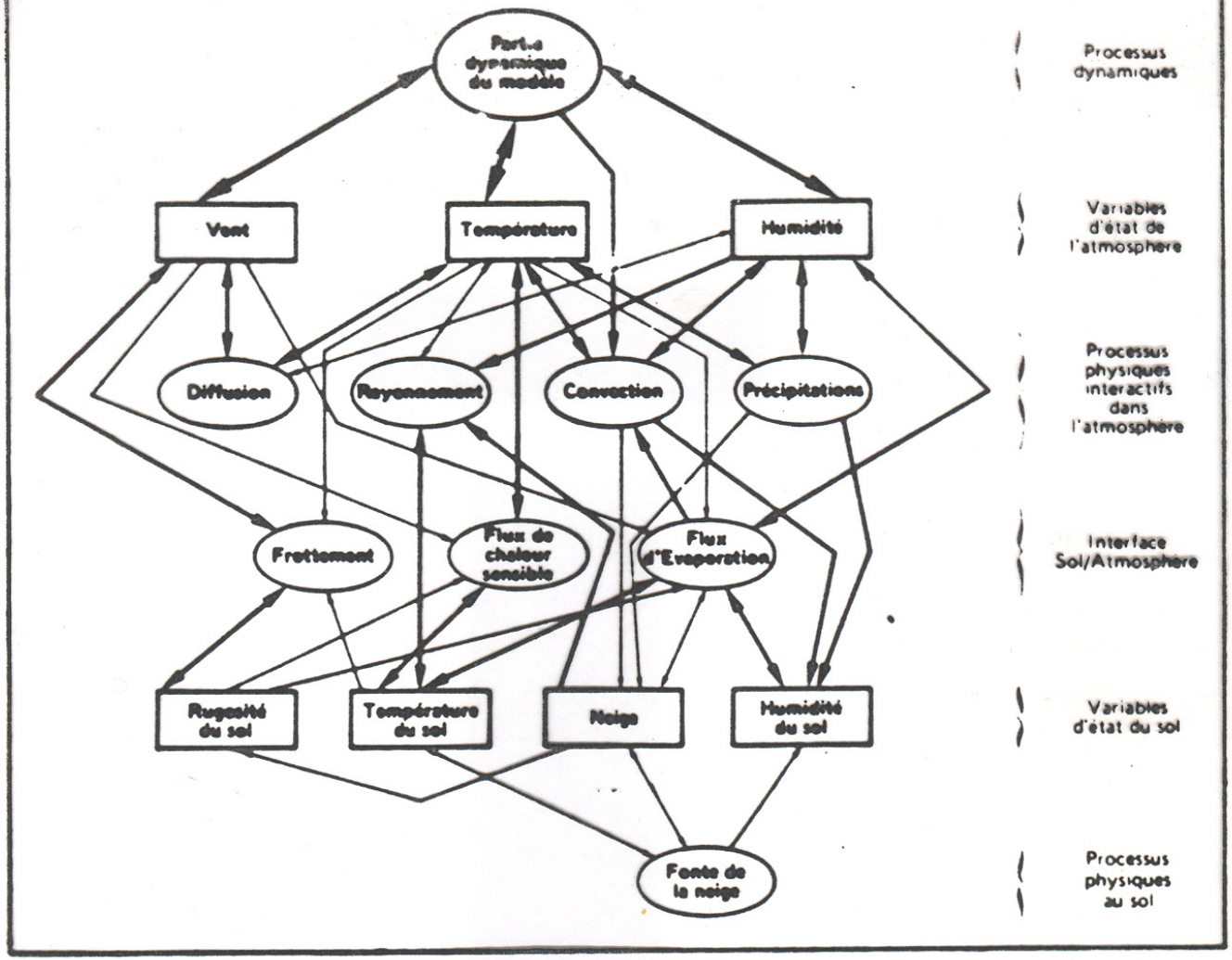
Linear operations, and in particular differentiation with respect to spatial variables, are performed in spectral space, while nonlinear operations and 'physical' computations (advection by the motion, diabatic heating and cooling, …) are performed in gridpoint physical space. This requires constant transformations from one space to the other, which are made possible at an acceptable cost through the systematic use of Fast Fourier Transforms.

For that reason, those models are called *semi-spectral*.

Numerical schemes have been gradually developed and validated for the ‘dynamics’ component of models, which are by and large considered now to work satisfactorily (although regular improvements are still being made).

The situation is different as concerns ‘physics’, where many problems remain (as concerns for instance the water cycle and the associated exchanges of energy, the interactions between the atmosphere and the underlying medium, or the statistical representation, or ‘parameterization’, of subgrid scale processes). ‘Physics’ as a whole remains the weaker point of models, and is still the object of active research.

5 - SCHEMA DES INTERACTIONS PHYSIQUES DANS LE MODELE



Many other processes have been introduced with time in numerical models

- different phases of water (in particular cloudiness, including various types of clouds)

- radiative impact of aerosols, either natural or anthropogenic

- water run-off (including underground run-off)

- secondary components (ozone, …)

- …

- effect of solar eclipses (ECMWF Technical Memorandum 920, December 2024)

Machine Learning models

Machine Learning (aka ***Artificial Intelligence***)

Set of empirical (vector) data

$$(x_i , y_i) , i = 1, N$$

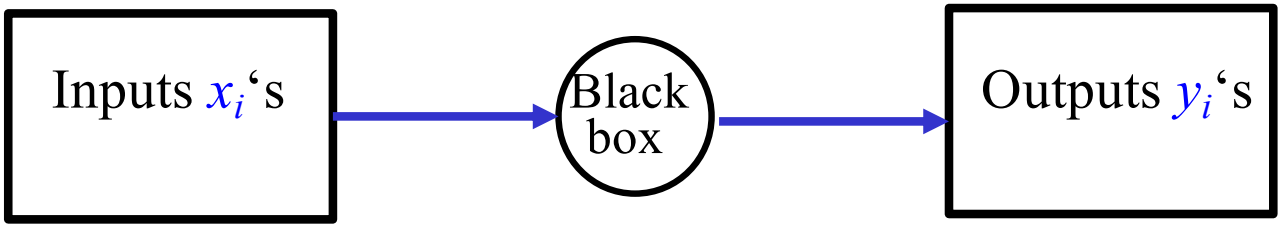
with no *a priori* explicitly known relationship between the inputs x_i's and the outputs y_i's.

Look for an explicit relationship of the form

$$y \approx f(x)$$

at least over a practically useful domain of variation of x.

Machine Learning (2)



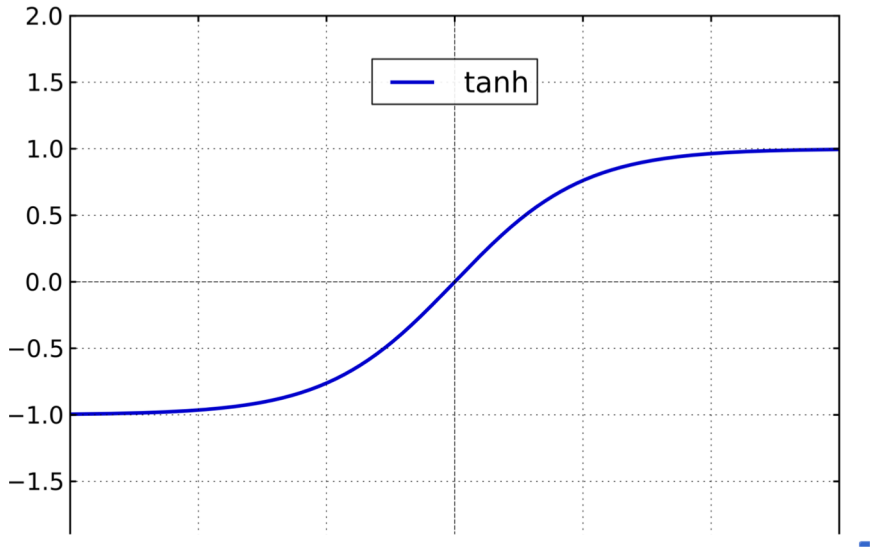
Replace black box with (possibly approximate) function $y \approx f(x)$

Neural networks define the function f as a composition of basic 'simple' functions, called *activation functions*.

Machine Learning (3)

Sigmoid functions, *e.g.* the *hyperbolic tangent* function $\tanh(x)$, are very commonly used as activation functions

$\tanh(x) = (e^{2x} - 1) / (e^{2x} + 1)$

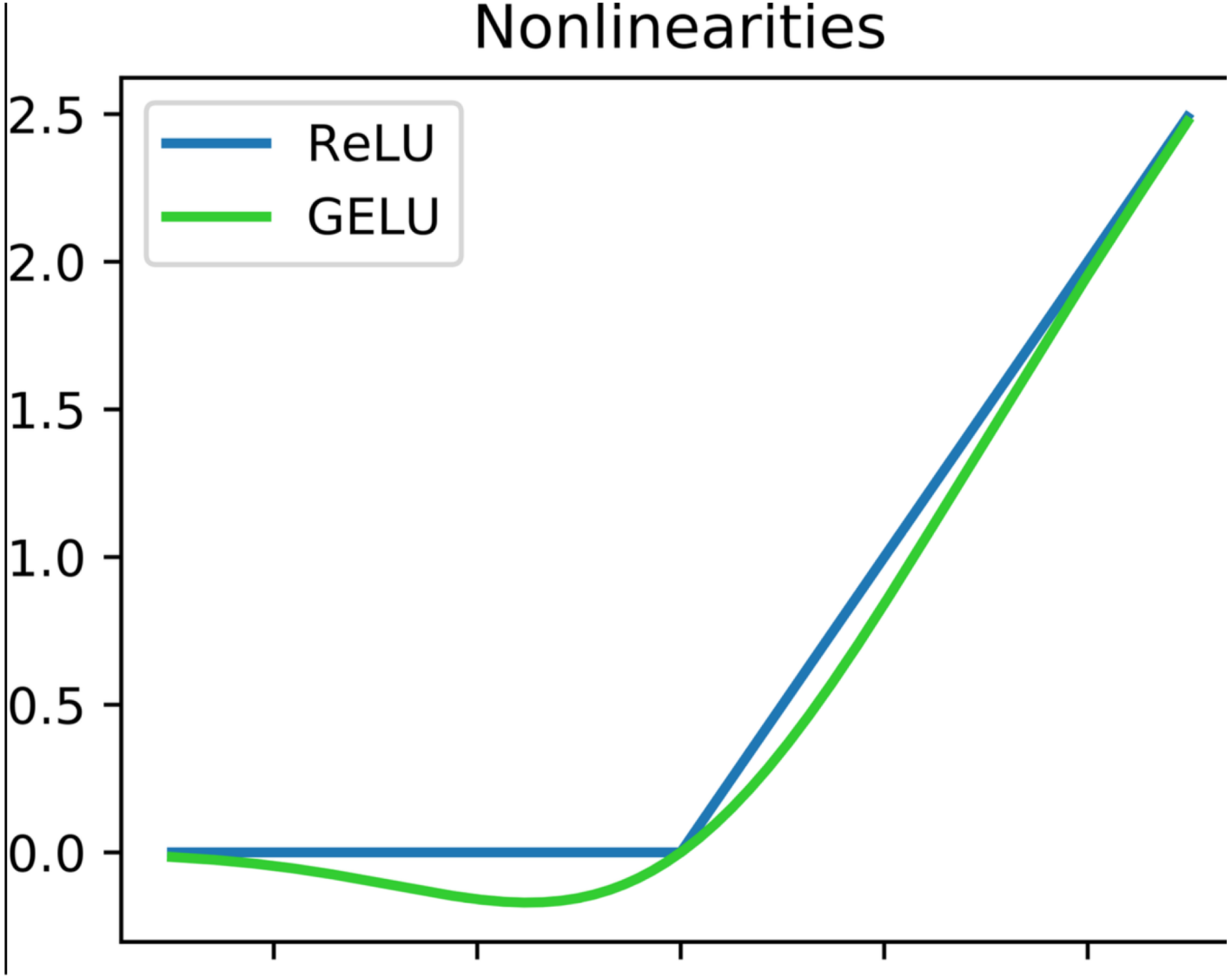


Affine change of coordinates. Four degrees of freedom : two for the coordinates of the central point, and one for the range of variation in each direction.

Machine Learning (4)

A typical network consists of layers of elementary activation functions, or *neurons,* the neurons in each layer being compositions of neurons in previous layers. Weights given to the neurons are determined by minimization of a *loss function* that measures the misfit between the original and computed outputs (for example a quadratic difference).

Machine Learning (5)



Other commonly used activation functions

ReLU (rectified linear unit) (blue)

Gaussian-error linear unit (GELU) (green)

Machine Learning (6)

Neural networks

This approach, with many variants, has proved to be extremely efficient, and is now used for innumerable applications in many different domains.

It has been applied to numerical weather prediction. A number of recently developed ML softwares are

GraphCast

Pangu-Weather

FourCastNet

FuXi

AIFS (ECMWF)

Machine Learning (6). *Neural networks*

FuXi (伏羲) has been trained on 39 years of ERA5. It has a spatial resolution of 0.25° (28 km, against 9 km for ECMWF HRES) and produces forecasts for a number of meteorological variables

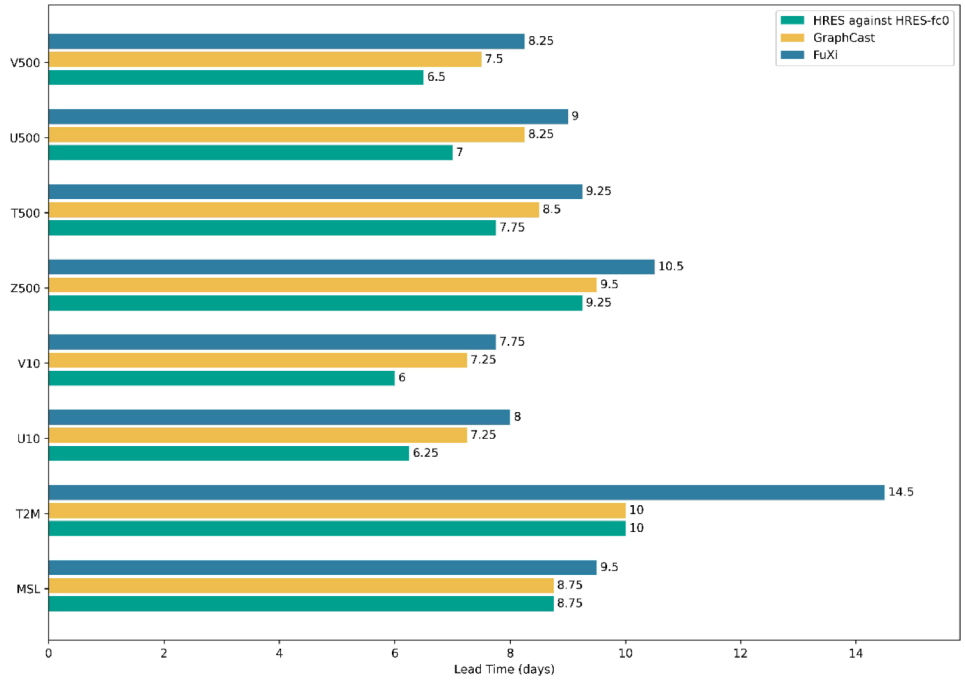


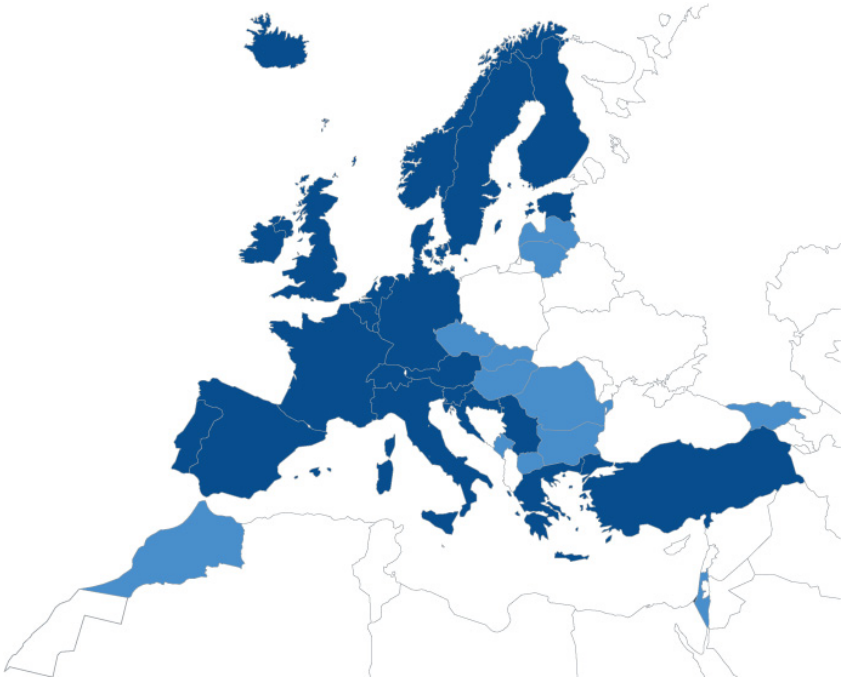
Fig. B.2: Skillful forecast lead times (ACC >0.6) of ECMWF HRES, Graphcast, and FuXi for 4 surface variables (MSL, $T2M$, $U10$, and $V10$) and 4 upper-air variables ($Z500$, $T500$, $U500$, and $V500$) at 500 hPa pressure level.

European Centre for Medium-range Weather Forecasts (ECMWF)

(Centre Européen pour les Prévisions Météorologiques à Moyen Terme, CEPMMT)

Centre is a common service to 23 European countries, with cooperative agreements with 12 other countries.

Located in Reading (UK), Bologna (Italy) and Bonn (Germany)



ECMWF's core mission is to:[12]

- Produce numerical weather forecasts and monitor planetary systems that influence weather
- Carry out scientific and technical research to improve forecast skill
- Maintain an archive of meteorological data

Produces regular (and in particular daily) forecasts, at various ranges, for various aspects of the weather system

Integrated Forecasting System (IFS)

June 2023 High-resolution model (HRES, now IFS-CF)

Triangular semi-spectral truncation TCO1279 / O1280 (horizontal resolution ≈ 9 kilometres)

Hydrostatic primitive equations. 137 vertical levels (0 - 80 km)

Finite-element vertical discretisation (hybrid coordinate)

Dimension of corresponding state vector > 10^9

Integration timestep (semi-Lagrangian semi-implicit scheme): 450 seconds

October 2024 Integrated in ensemble form (1 control + 50 perturbed forecasts) four times a day (from 00, 06, 12 and 18 UTC) to 6-day range and twice a day (from 00 and 12 UTC) to 15-day range.

Integrated Forecasting System (IFS) (contd)

- ECMWF Reanalysis v5 (ERA5). Covers the period from January 1940 to present, provides hourly estimates of a large number of atmospheric, land and oceanic climate variables at 32-km horizontal resolution and 137 pressure levels. Includes estimates of uncertainty for all variables. It is produced using 4D-Var data assimilation over 12-hour assimilation windows, and is constantly updated.

- Subseasonal and seasonal forecasts.

Artificial Intelligence/Integrated Forecasting System (AIFS)

AIFS Single Implemented operationally on 25 February 2025

Developed on Machine Learning through *graph neural networks*. Trained in-house on 1979–2018 ERA5 reanalyses. Loss function defined on *Continuous Ranked Probability Score* (*CRPS*)

Horizontal grid spacing of 32 km, 13 pressure-levels

90% decrease in computing time wrt IFS, and reduction of approximately 1,000 times in energy use for making a forecast.

Run twice a day, with ‘timestep’ of 6 hours, to 15-day range

Ensemble version *AIFS ENS* implemented operationally on 1 July 2025
(50 + 1 runs at 32-km resolution)

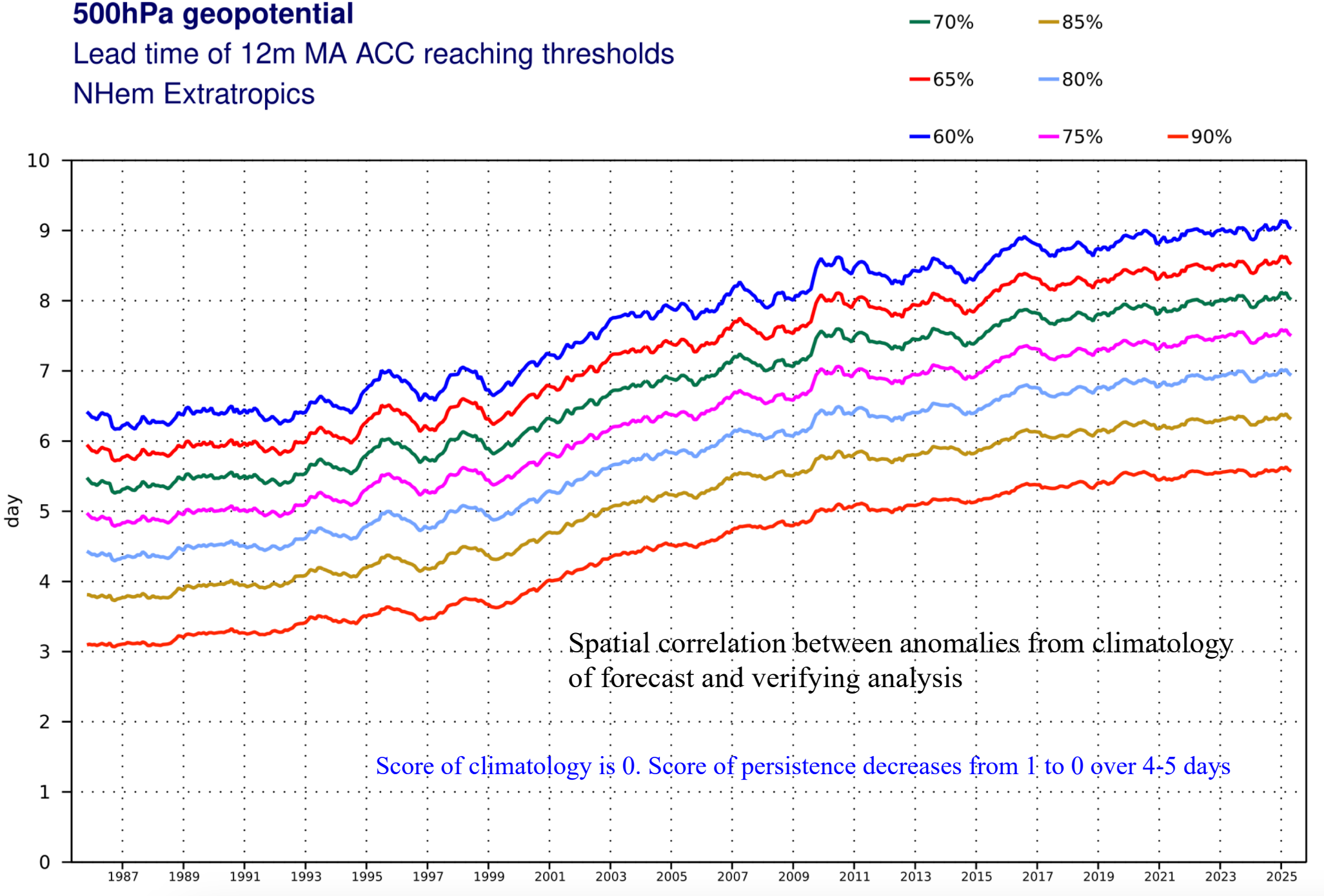
Results on site of ECMWF *www.ecmwf.int*

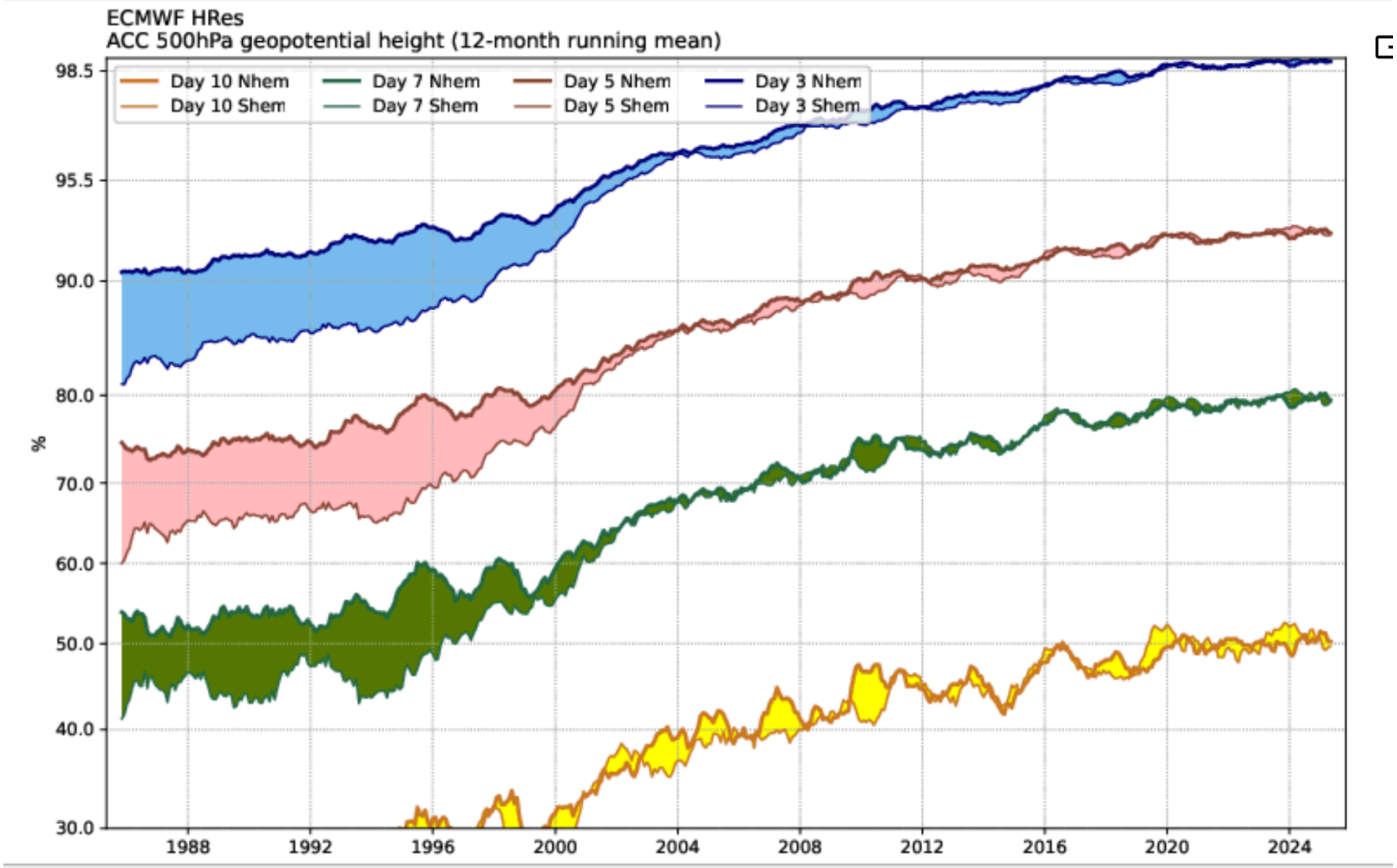
In particular

T. Haiden *et al., Evaluation of ECMWF forecasts,* Technical Memorandum 931, September 2025, ECMWF, Reading, UK.

Available at the address :

81680-evaluation-of-ecmwf-forecasts.pdf


500hPa geopotential
Lead time of 12m MA ACC reaching thresholds
NHem Extratropics
70%
85%
65%
80%
60%
75%
90%
day
10
9
8
7
6
5
4
3
2
1
0
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
2021
2023
2025
Spatial correlation between anomalies from climatology
of forecast and verifying analysis
Score of climatology is 0. Score of persistence decreases from 1 to 0 over 4-5 days


ECMWF HRes
ACC 500hPa geopotential height (12-month running mean)
Day 10 Nhem
Day 10 Shem
Day 7 Nhem
Day 7 Shem
Day 5 Nhem
Day 5 Shem
Day 3 Nhem
Day 3 Shem
%
98.5
95.5
90.0
80.0
70.0
60.0
50.0
40.0
30.0
1988
1992
1996
2000
2004
2008
2012
2016
2020
2024

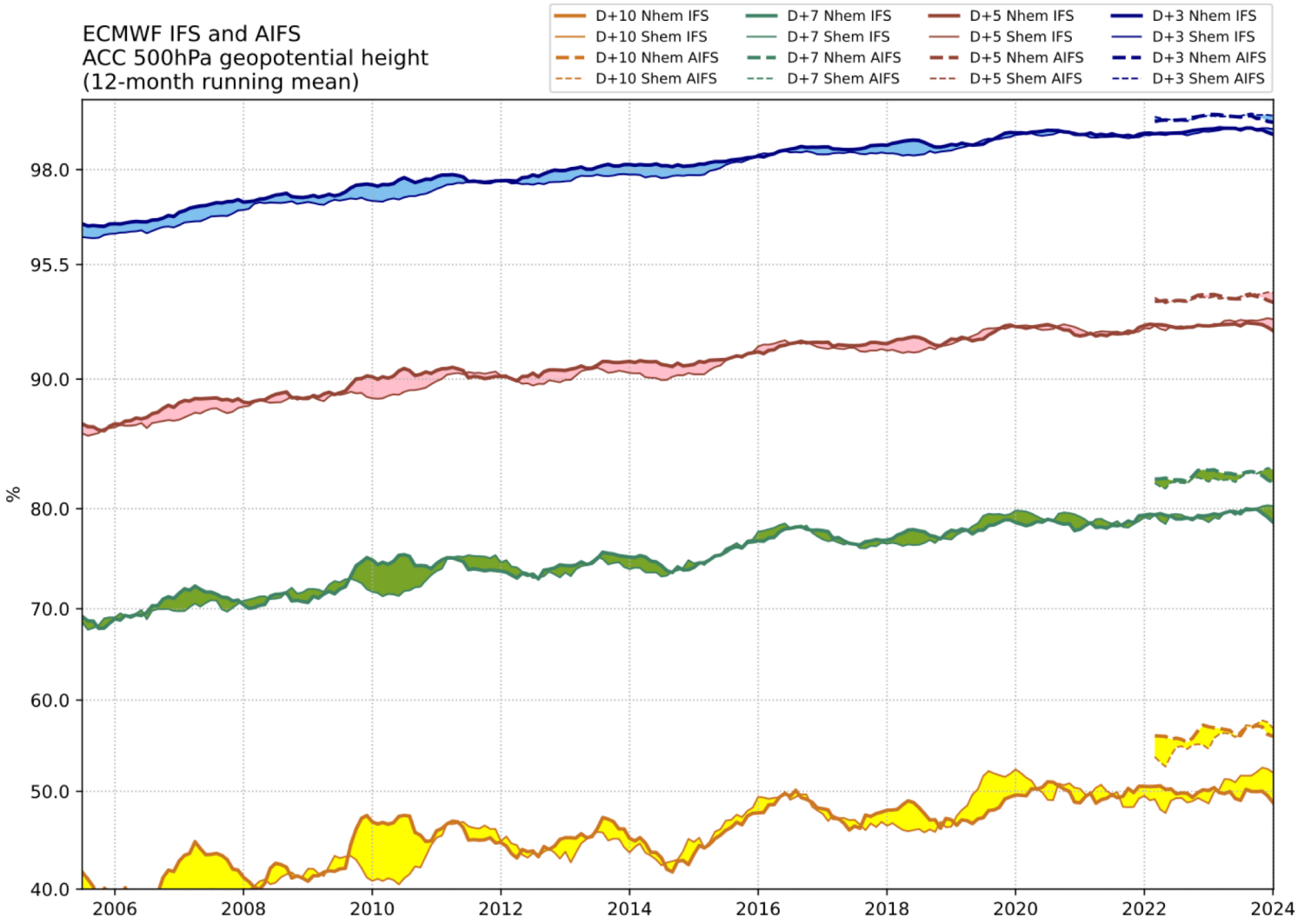
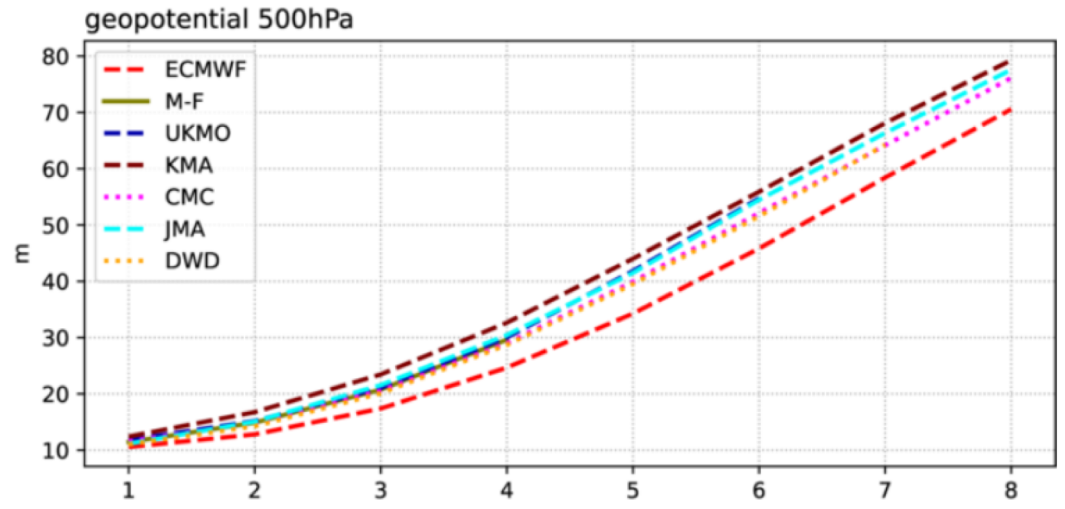


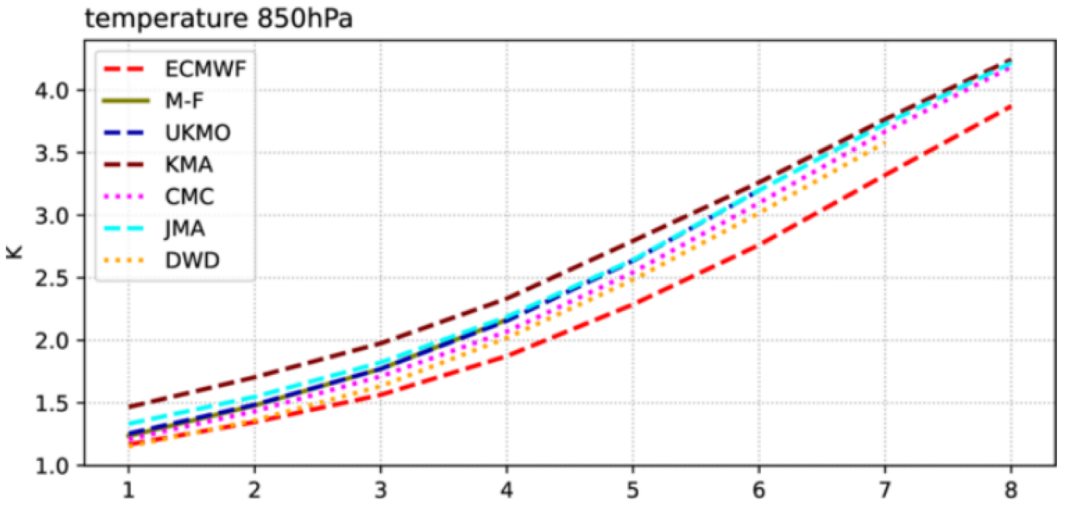
Fig. 1. Time series of 12-month running mean forecast skill of the ECMWF high-resolution deterministic Integrated Forecasting System (IFS, solid lines) and ML IFS (AIFS, dashed lines). Skill is expressed as the anomaly correlation (= correlation between forecasts and verifying analyses, normalised by climatological signal) for the 500 hPa height. Thick (thin) curves show northern (southern) hemisphere averages for day-3 (blue), day-5 (red), day-7 (green) and day-10 (yellow). (Figure courtesy Martin Janousek, ECMWF).

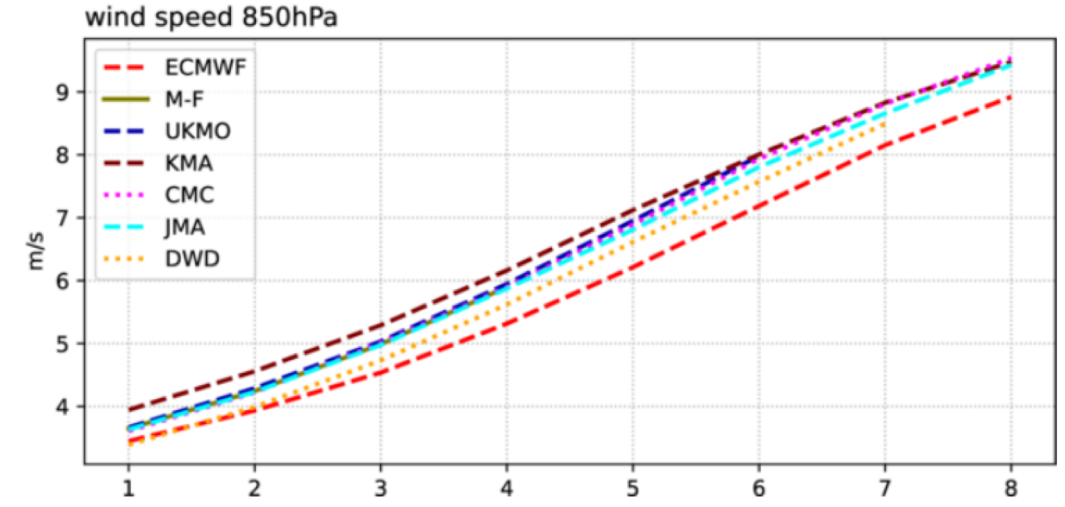
August 2024
– July 2025

Northern hemisphere extratropics

Verification against radiosondes







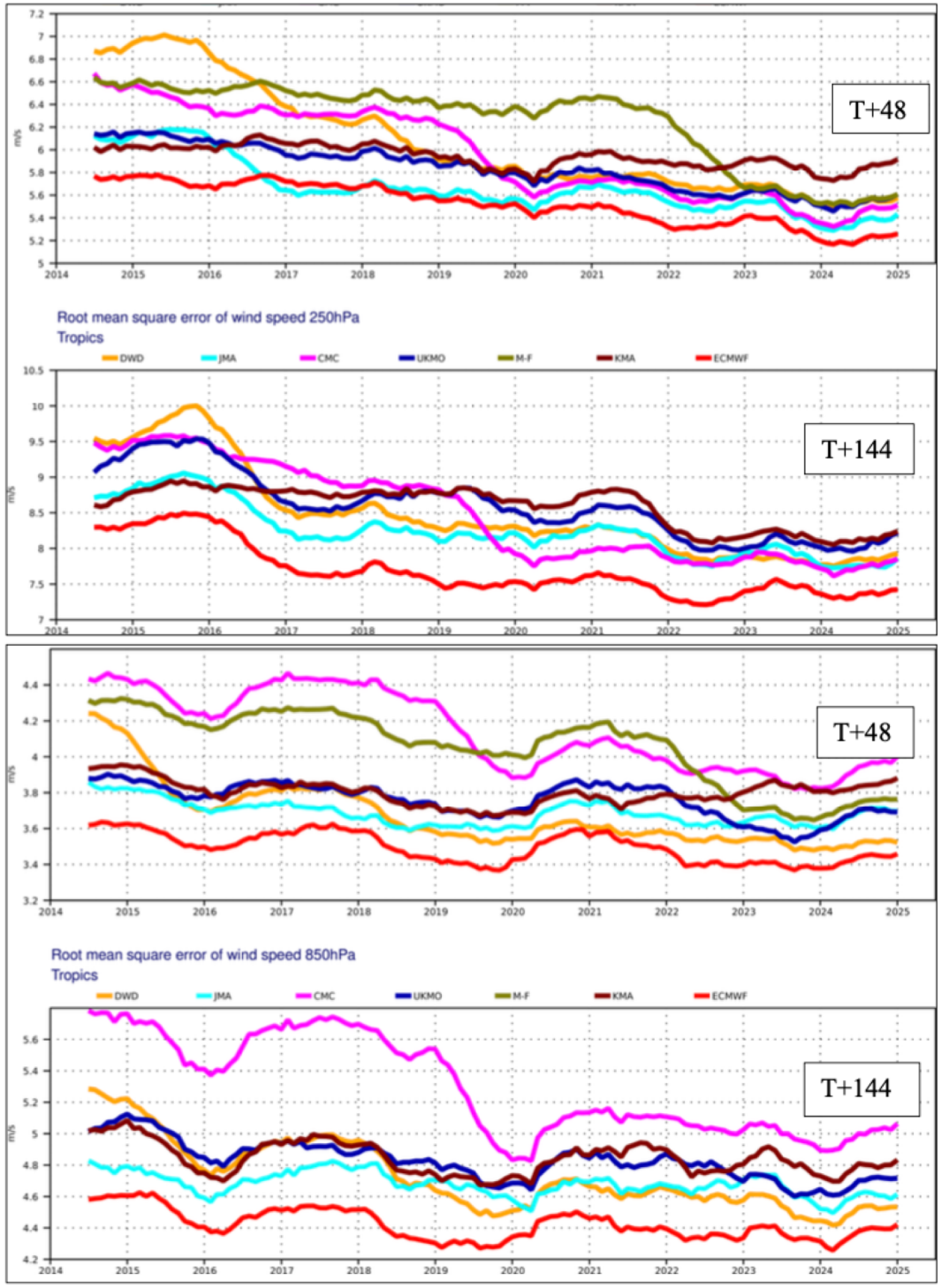


Figure 18: As Figure 17 but for verification against radiosonde observations.

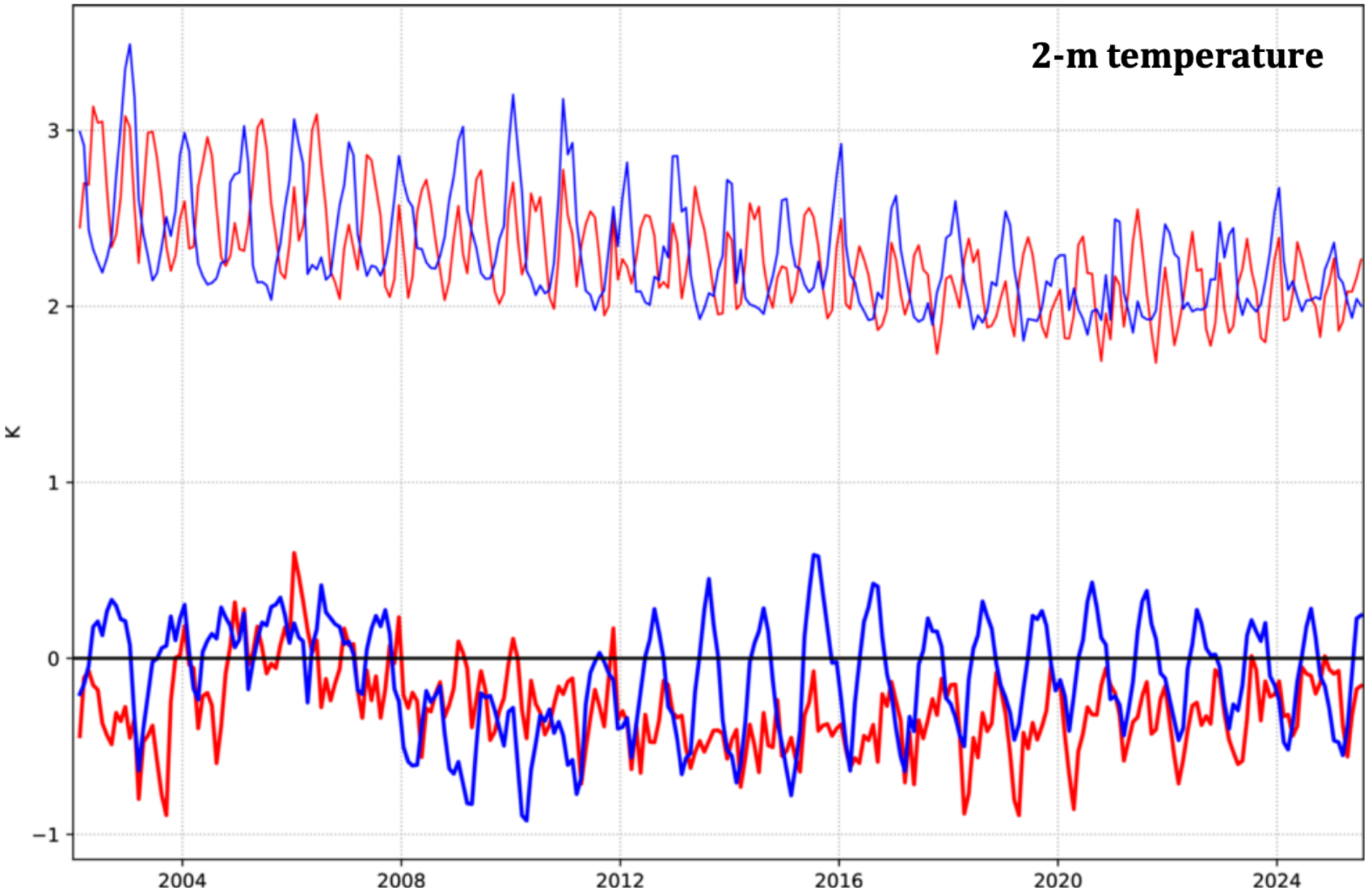


Figure 28: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

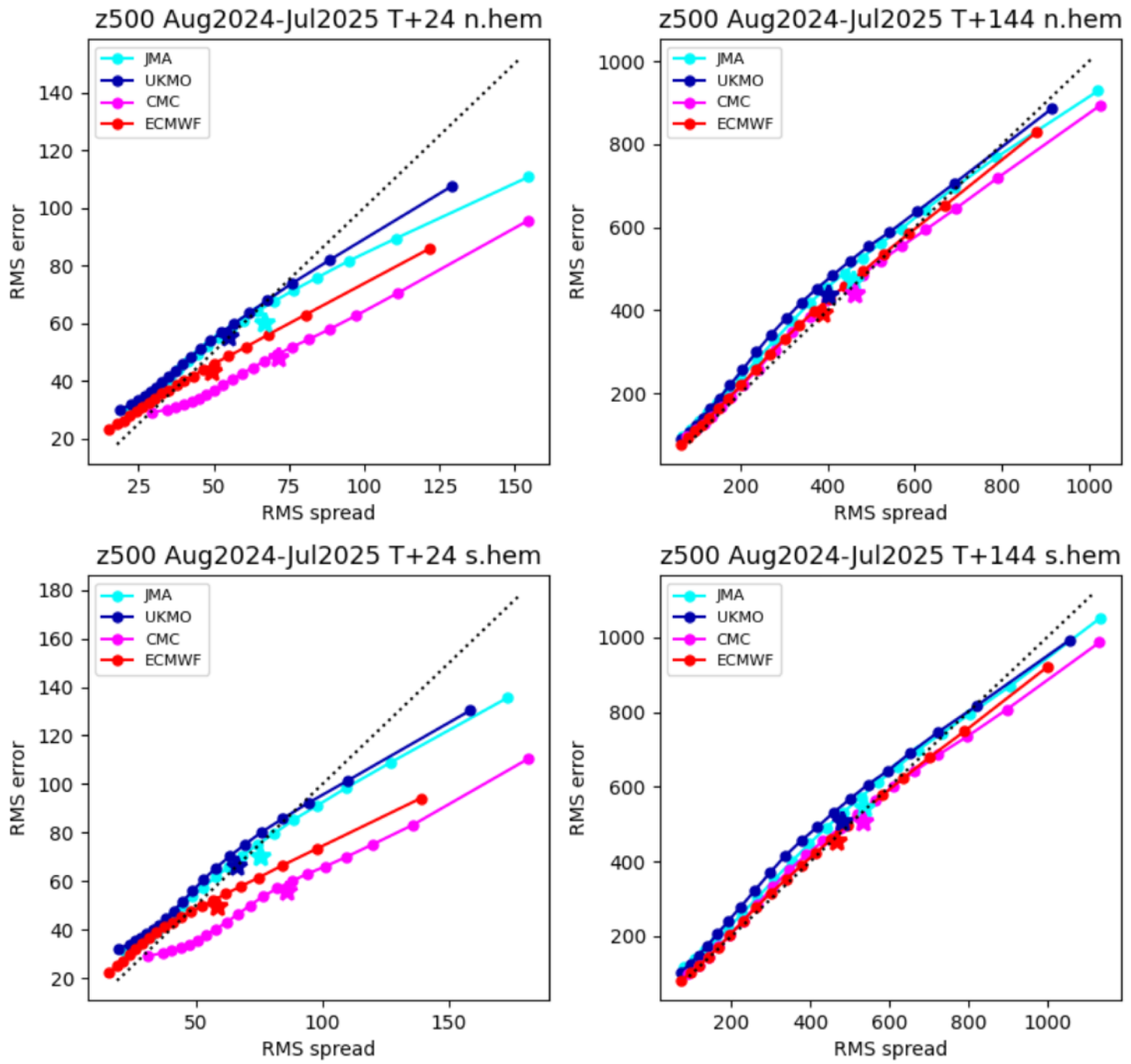


Figure 10: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2024–July 2025 in the northern (top) and southern (bottom) hemisphere extra-tropics for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, and stars show average error-spread relationship.

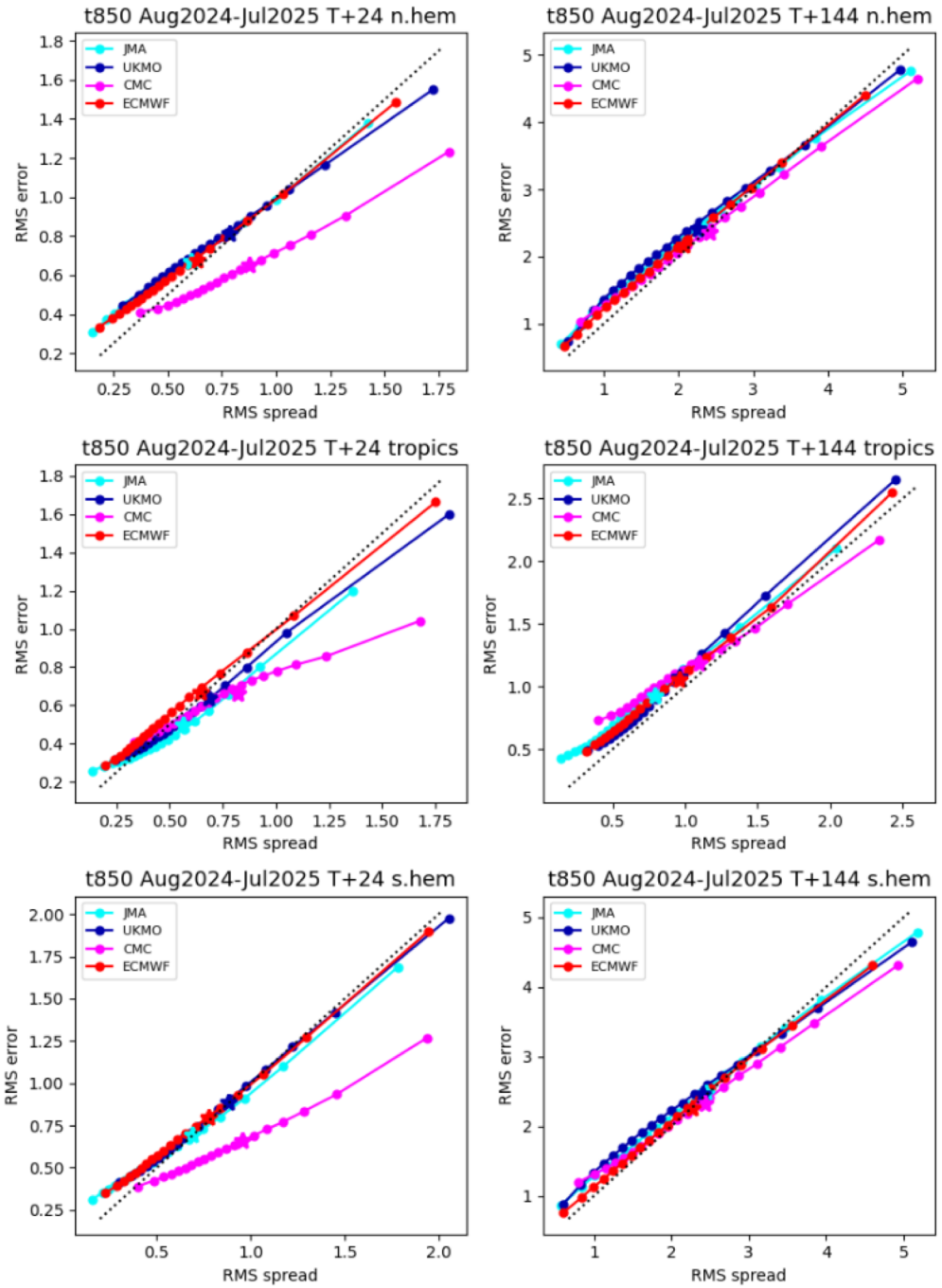
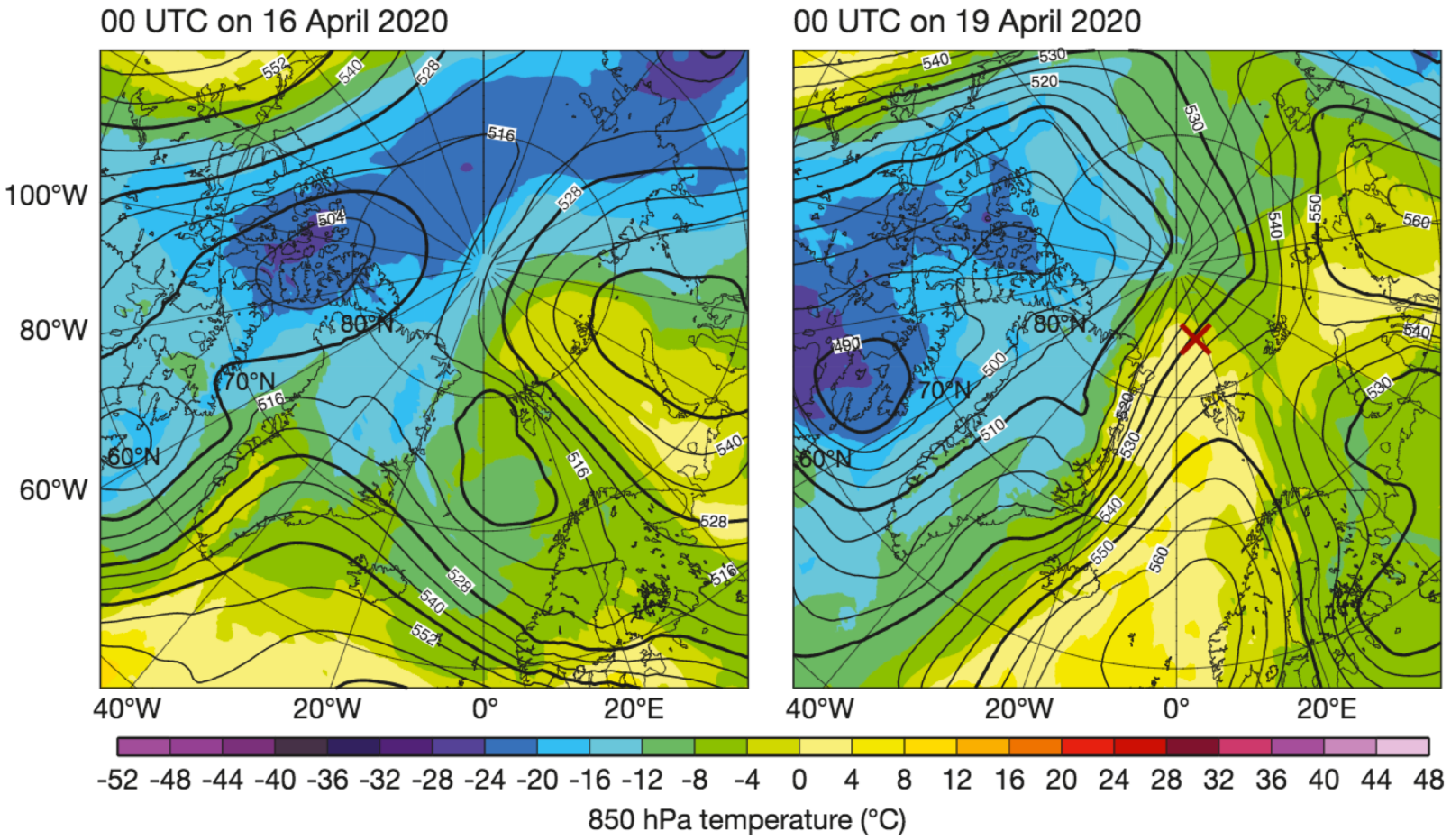
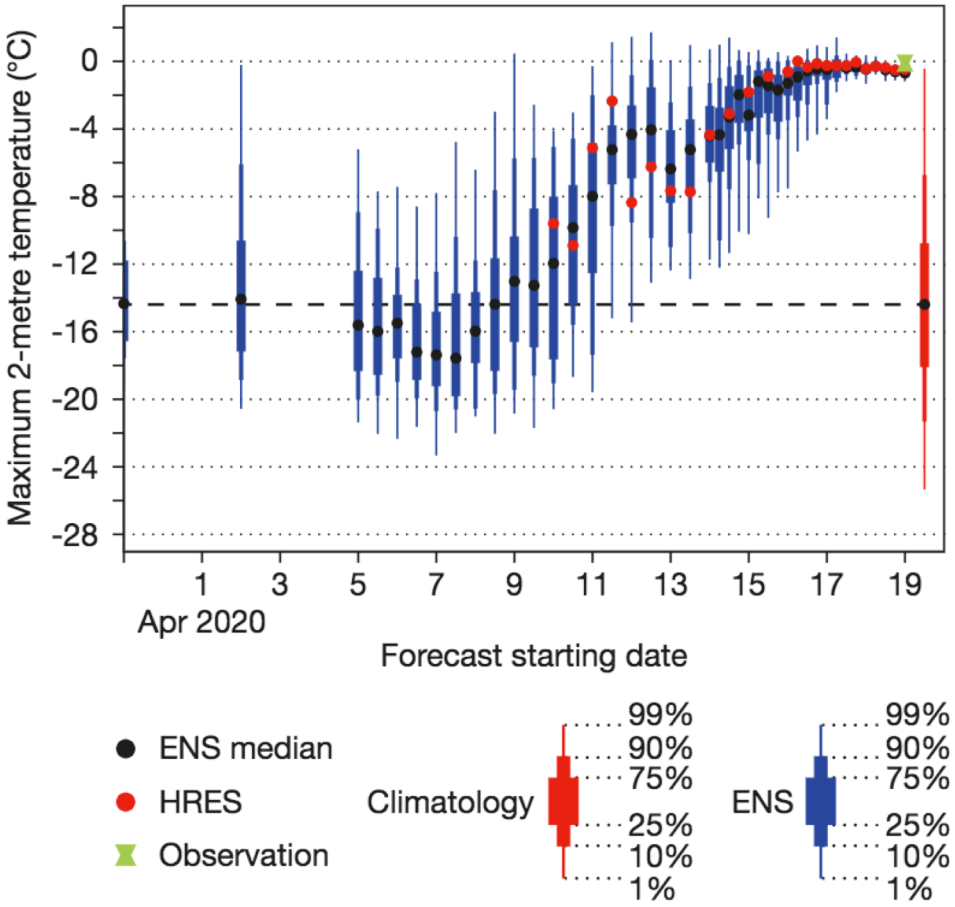


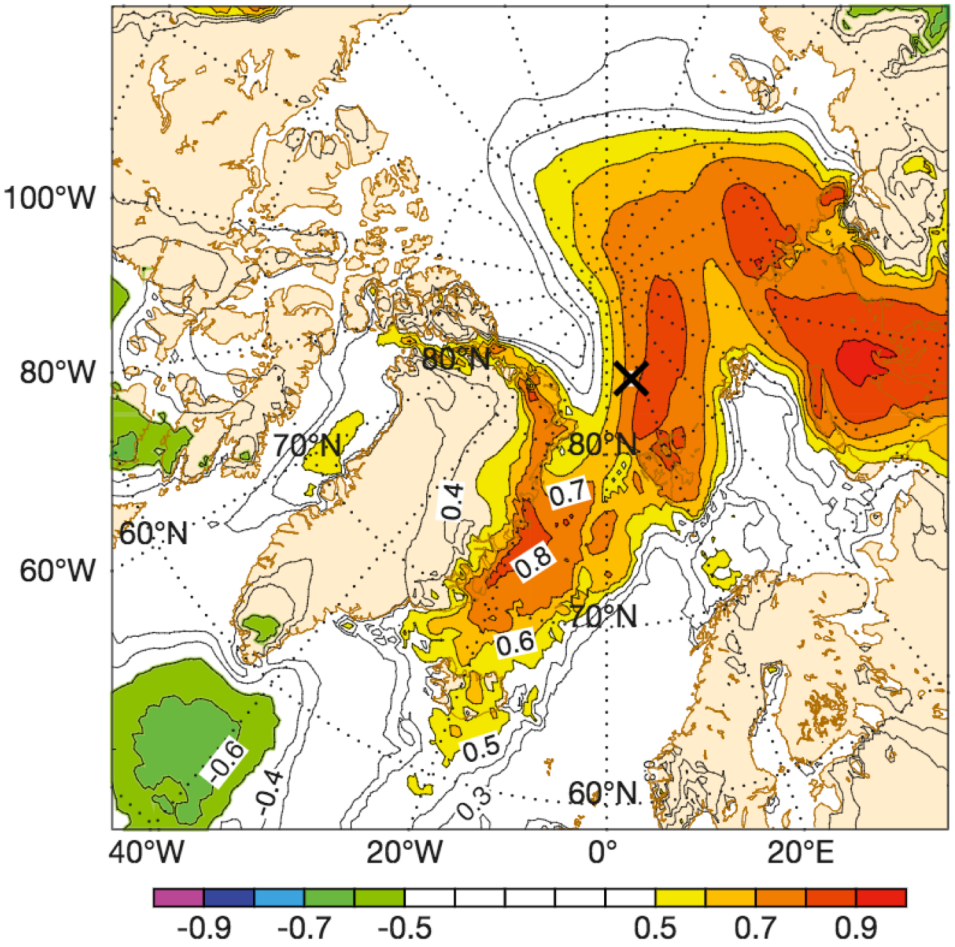
Figure 11: As Figure 8 but for 850 hPa temperature and including the tropics.



Synoptic situation on 16 and 19 April. Analysis of geopotential height at 500 hPa (contours) and temperature at 850 hPa (shading) for 00 UTC on 16 April 2020 (left) and 00 UTC on 19 April (right). The cross shows the approximate location of Polarstern on 19 April.



Evolution of forecasts for the 19 April warm air intrusion. The plot shows ensemble forecasts with different starting times for maximum 2-metre temperature for the Polarstern location on 19 April.



Two-metre maximum temperature EFI. The chart shows the 5-day forecast from 00 UTC on 15 April 2020 of the EFI for 2-metre maximum temperature on 19 April.

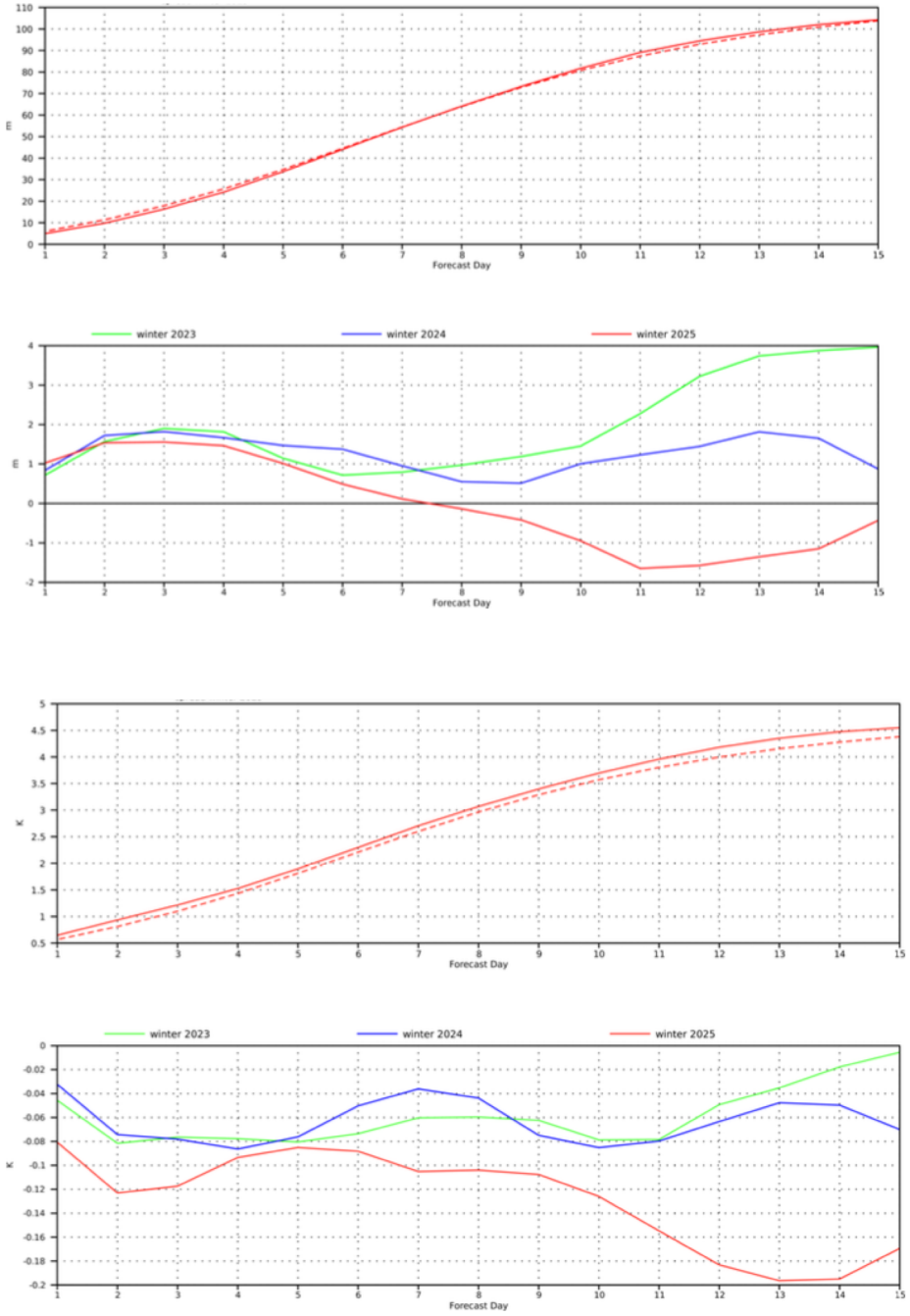


Figure 9: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2024–2025 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate spread is too small); verification is against analysis, plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extratropical northern hemisphere for forecast days 1 to 15.

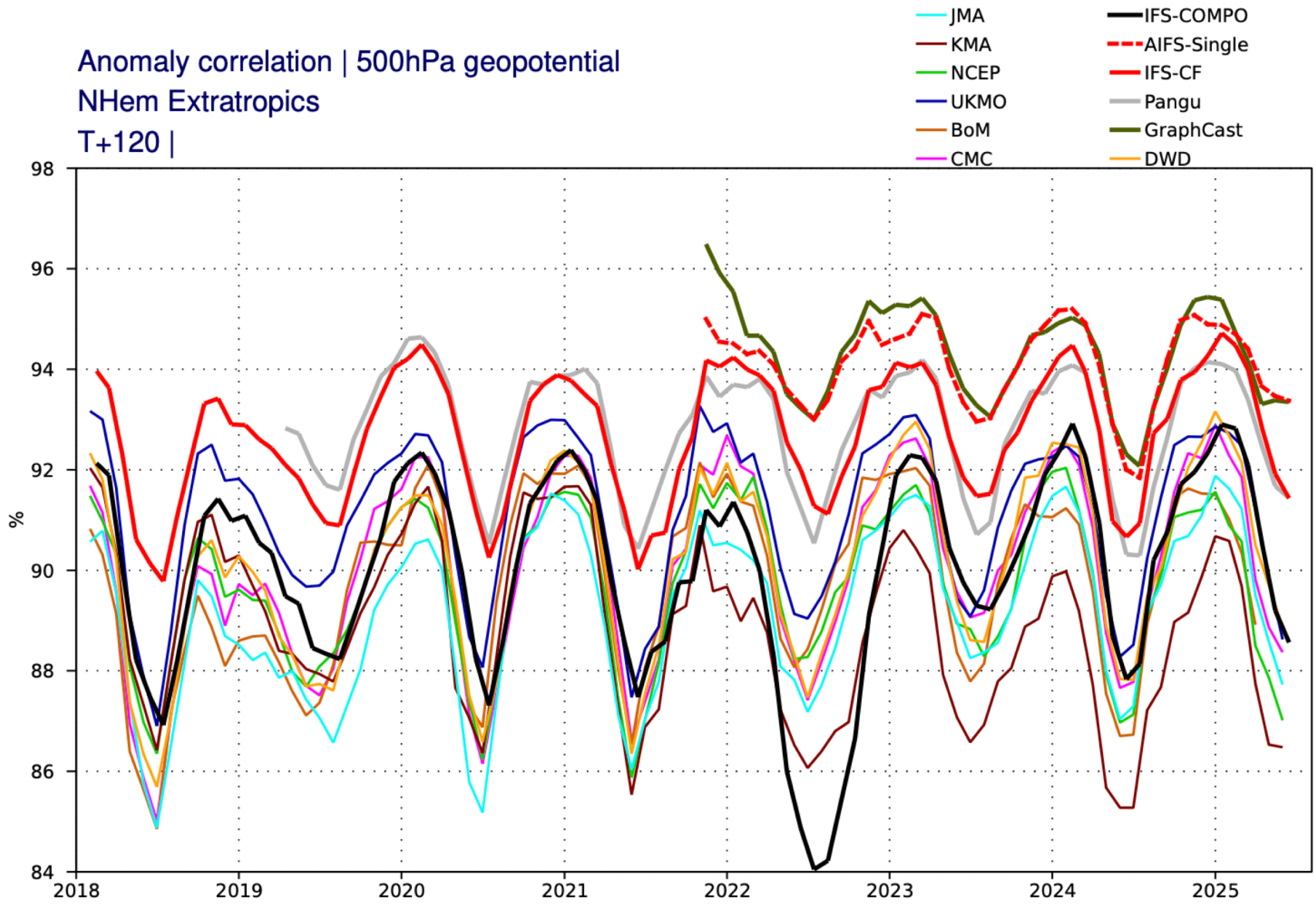


Figure 19: Anomaly correlation of 500 hPa geopotential in the northern hemisphere extratropics at day 5. CAMS forecast IFS-COMPO (black) shown in comparison to the IFS-CF (red) and forecasts from other global centres (thin lines). Also shown are forecasts from machine learning (ML) models: GraphCast (olive), Pangu (grey), and AIFS-Single (red dashes).

IFS-COMPO lower resolution + aerosols

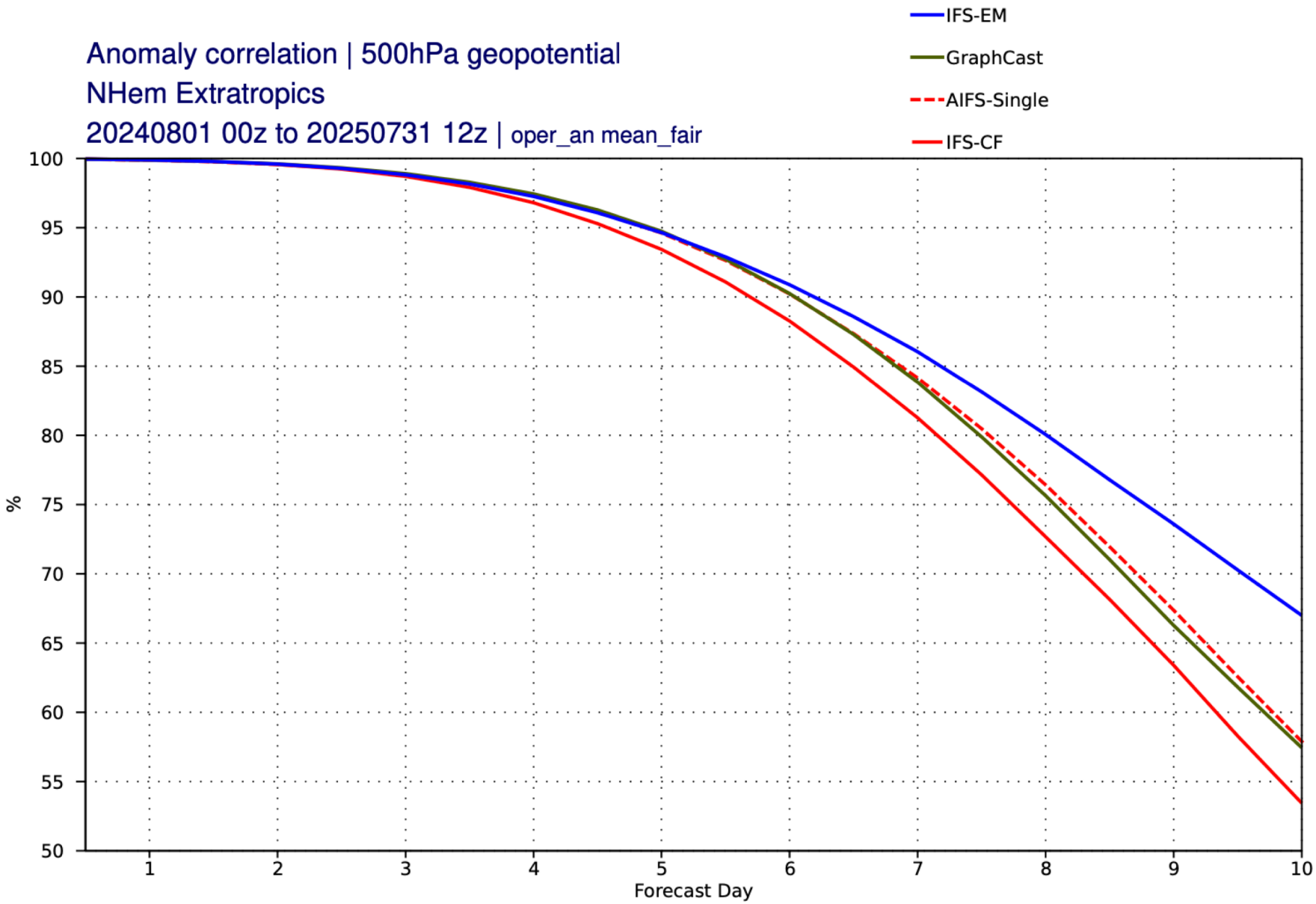


Figure 20: Anomaly correlation of 500 hPa geopotential in the northern extratropics for the 12-month period Aug 2024 to July 2025. Blue: IFS-ENS mean, olive: GraphCast, red dashes: AIFS-Single, red: IFS-CF.

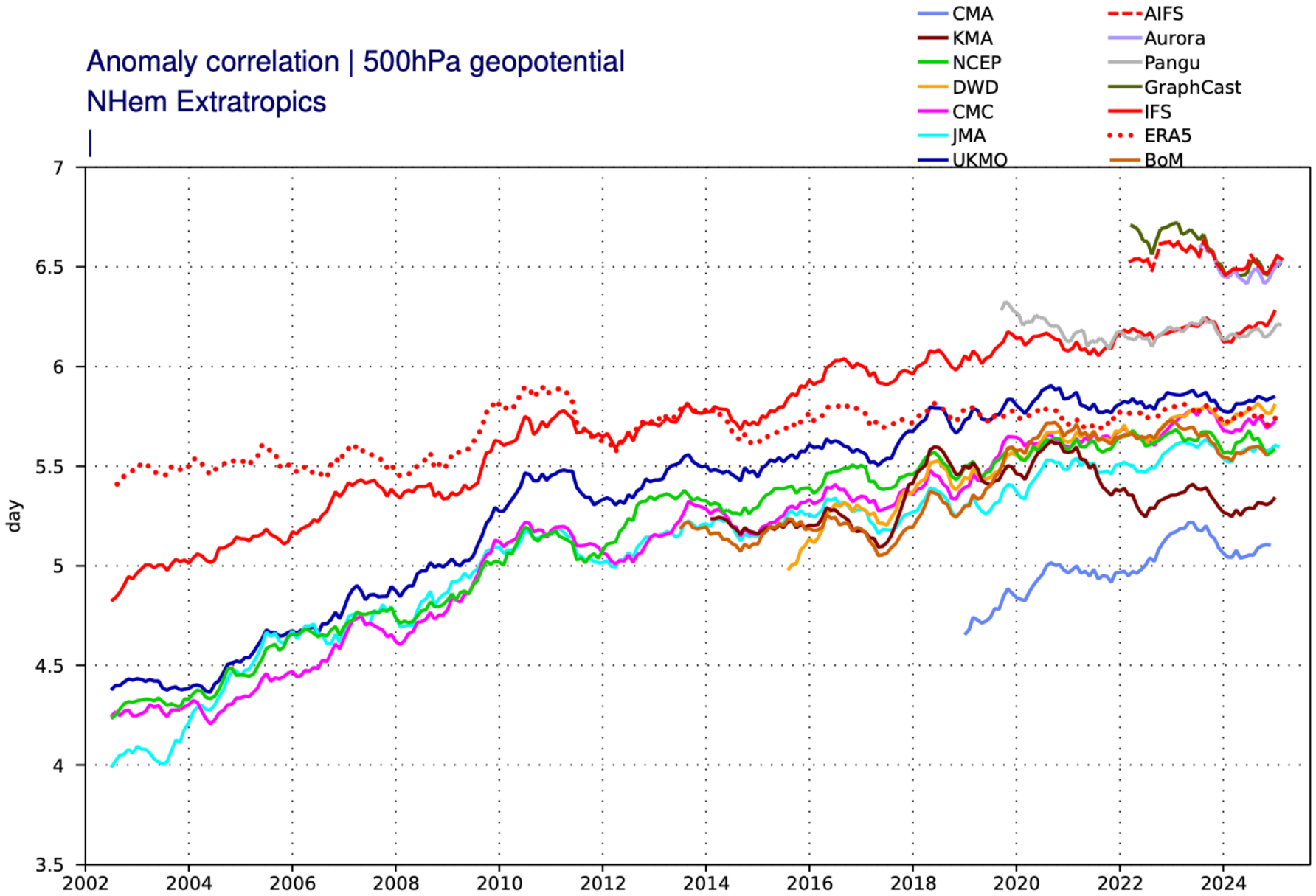


Figure 21: Lead time at which the 500 hPa anomaly correlation in the northern extratropics drops to 86% (specific threshold chosen such that lead time does not exceed forecast range of any centre). Comparison of physics-based forecasts from other global centres, IFS-CF (red), AIFS-Single (red dashes), ERA5 (red dots) for reference, and other ML models run at ECMWF.

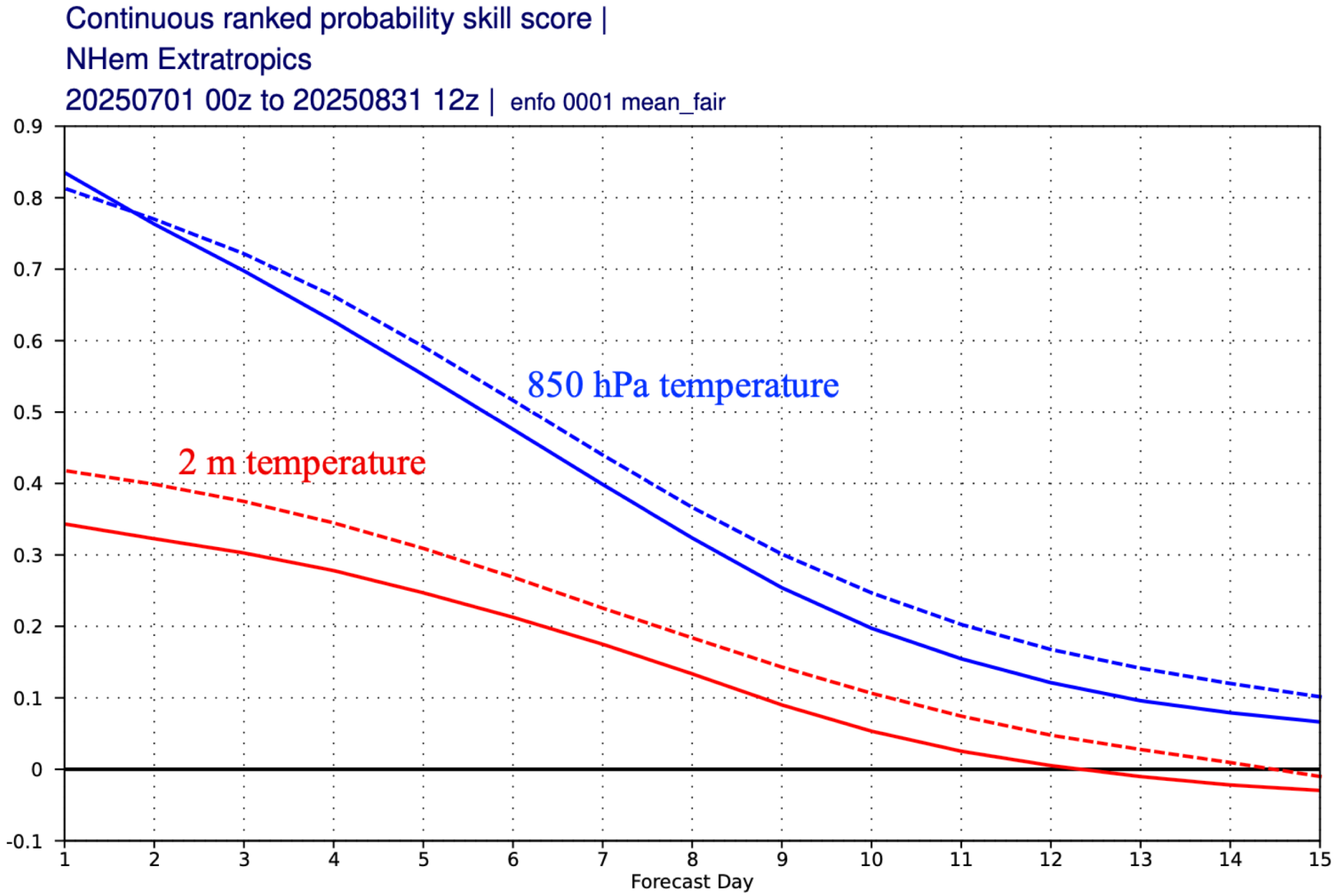
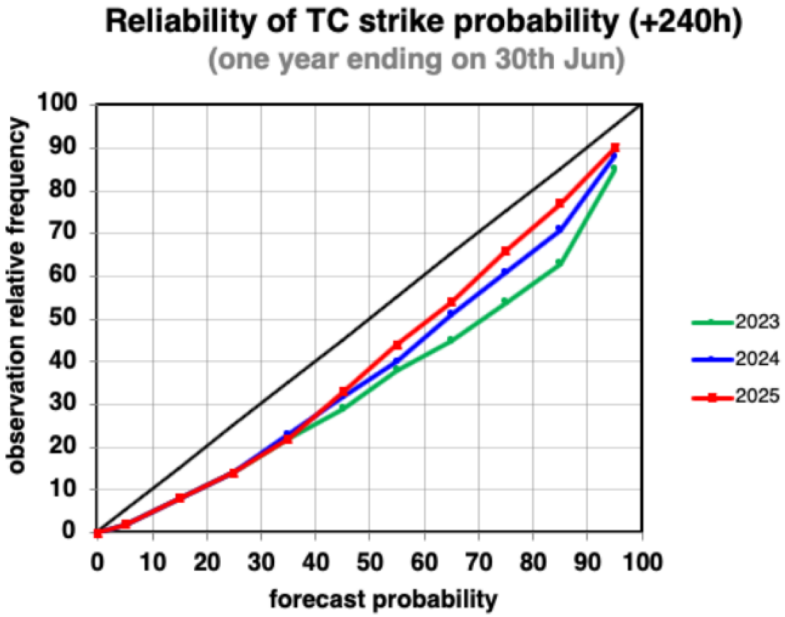
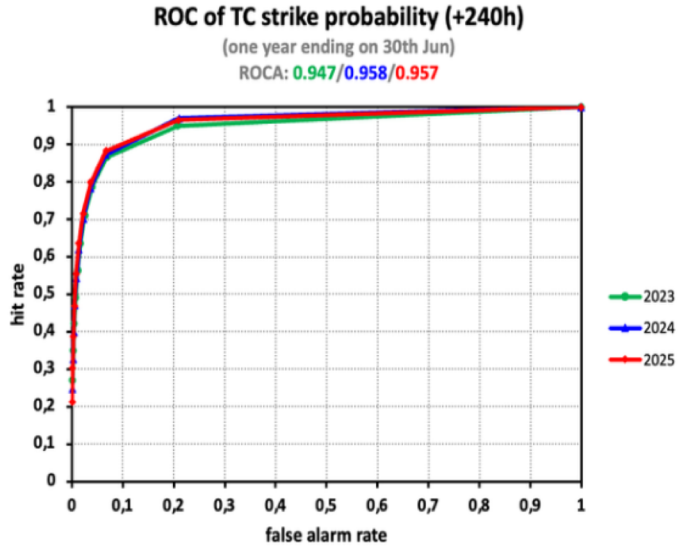


Figure 22: Comparison of IFS-ENS (solid) and AIFS-ENS (dashed) scores for 850 hPa temperature (blue) and 2 m temperature (red) in the northern extratropics. Shown is the CRPSS evaluated against analysis for 850 hPa temperature and evaluated against SYNOP for 2 m temperature. Verification period is July-August 2025.

Probability at any location that a reported tropical cyclone will pass within 120 km during the next 240 hours





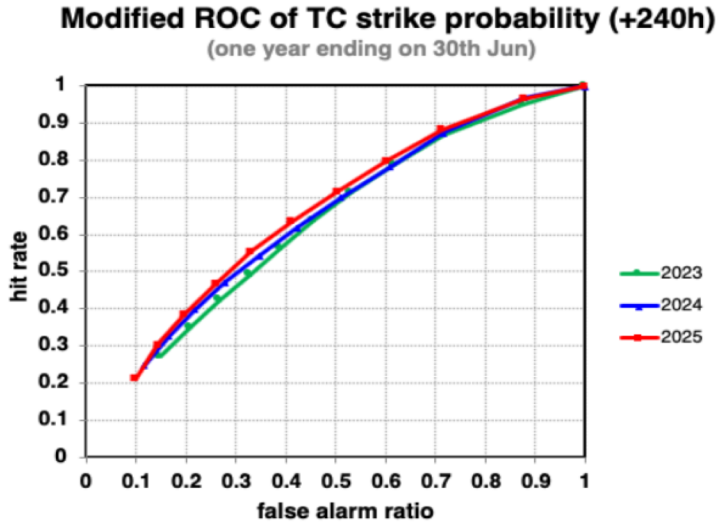


Figure 45: Probabilistic verification of IFS-ENS tropical cyclone forecasts at day 10 for three 12-month periods: July 2022–June 2023 (green), July 2023–June 2024 (blue) and July 2024–June 2025 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) the standard ROC diagram and (right) a modified ROC diagram, where the false alarm ratio is used instead of the false alarm rate. For both ROC and modified ROC, the closer the curve is to the upper-left corner, the better, indicating a greater proportion of hits, and fewer false alarms.

ECMWF

500 hPa geopotential

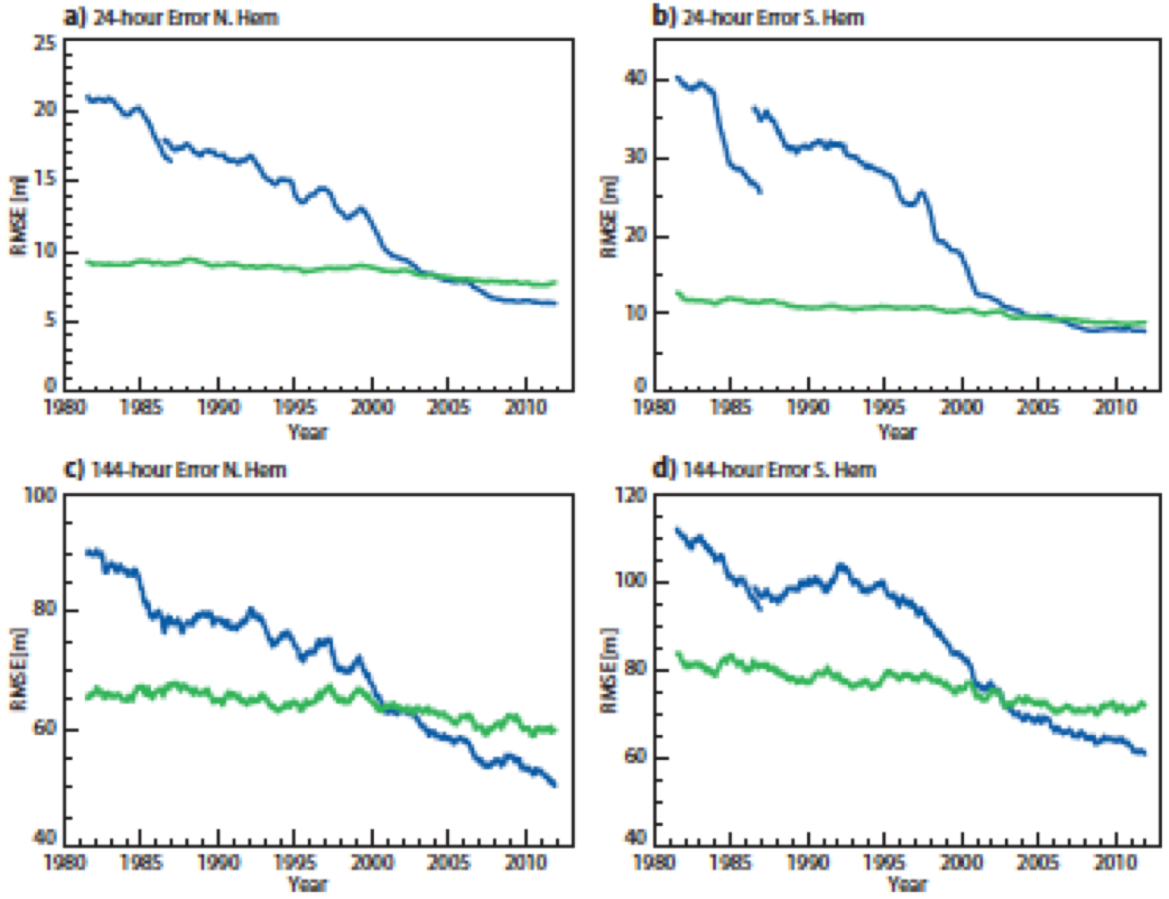


FIG. 3. Evolution of forecast errors from 1981 to 2012 for N.Hem (a and c) and S.Hem (b and d). Operational forecasts (blue) and ERA Interim (green). Note that before 1986 the operational analysis is used to verify the operational forecasts, after 1986 ERA Interim is used for the verification (with an overlap of 6 months present).

Magnusson and Källén, 2013, *Mon. Wea. Rev.*, **141**, 3142–3153

Remaining Problems

Mostly in the 'physics' of models (Q and $\underline{F}$ terms in basic equations)

- Water cycle (evaporation, condensation, influence on radiation absorbed or emitted by the atmosphere)

- Exchanges with ocean or continental surface (heat, water, momentum, …)

- …

Remaining Problems

Mostly in the 'physics' of models (Q and $\underline{F}$ terms in basic equations)

- Water cycle (evaporation, condensation, influence on radiation absorbed or emitted by the atmosphere)

- Exchanges with ocean or continental surface (heat, water, momentum, …)

- …

To sum up

Both physical and Machine Learning models are now used for weather prediction.

The latter are much more economical (by one order of magnitude). They produce forecasts that are more accurate in terms of RMS error and correlation coefficients, but bear on a rather limited amount of predicted parameters. The forecasts are also spatially smoother than physical forecasts

ML models are certainly going to evolve (and improve) significantly.

They are at the present stage trained on data sets produced by well-established and thoroughly validated physical models. How will that evolve ?

To sum up (contd)

Will there be some day global end-to-end weather prediction, based only on machine learning from observations ? Research work is being performed in that direction. What about new observing devices or instruments ? And on prediction of quantities which are not observed, but must be predicted anyway ?

Research is active on all those points.

What must we expect ? Impossible to tell to-day.

-

- What is assimilation ?

  - *Definition of initial conditions*

Purpose of assimilation : reconstruct as accurately as possible the state of the atmospheric or oceanic flow, using all available appropriate information. The latter essentially consists of

- The observations proper, which vary in nature, resolution and accuracy, and are distributed more or less regularly in space and time.

- Numerical models that produce (approximate) forecasts of the evolution of the flow from given initial and boundary conditions. These models can be built on either explicit description of the physical laws governing the flow or, now, on Machine Learning trained on past data (models of only the first type will be considered in the sequel).

- 'Asymptotic' properties of the flow, such as, *e. g.*, geostrophic balance of middle latitudes. Although they basically are necessary consequences of the physical laws which govern the flow, these properties can usefully be explicitly introduced in the assimilation process.

Both observations and 'model' are affected with some uncertainty $\Rightarrow$ uncertainty on the estimate.

For some reason, uncertainty is conveniently described by probability distributions (don't know too well why, but it works; see, *e.g.* Jaynes, 2007, *Probability Theory: The Logic of Science,* Cambridge University Press).

Assimilation is a problem in bayesian estimation.

Determine the conditional probability distribution for the state of the system, knowing everything we know (see Tarantola, A., 2005*, Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM).

Assimilation is one of many '*inverse problems*' encountered in many fields of science and technology

- solid Earth geophysics
- plasma physics
- 'nondestructive' probing
- navigation (spacecraft, aircraft, ….)
- …

Solution most often (if not always) based on Bayesian, or probabilistic, estimation. 'Equations' are fundamentally the same.

Difficulties specific to assimilation of meteorological observations :

- Very large numerical dimensions ($n \approx 10^6$-10^9 parameters to be estimated, $p \approx 10^7$-10^8 observations per 24-hour period). Difficulty aggravated in Numerical Weather Prediction by the need for the forecast to be ready in time.

- Non-trivial, actually chaotic, underlying dynamics

Proportion of computing resources devoted to assimilation of observations in the whole process of Numerical Weather Prediction has gradually increased over time.

Definition of initial conditions originally required a simple interpolation from observation stations to model gridpoints, with negligible cost. As of now, assimilation over 24 hours of observations requires about the same amount of resources as a 10-day physical forecast, including probabilistic forecast.

$z_1 = x + \zeta_1$ density function $p_1(\zeta) \propto \exp[- (\zeta^2)/2s_1]$

$z_2 = x + \zeta_2$ density function $p_2(\zeta) \propto \exp[- (\zeta^2)/2s_2]$

ζ_1 and ζ_2 mutually independent

$P(x = \xi \mid z_1, z_2)$?

$$x = \xi \;\Leftrightarrow\; \zeta_1 = z_1 - \xi \text{ and } \zeta_2 = z_2 - \xi$$

$$\begin{aligned} P(x = \xi \mid z_1, z_2) &\propto p_1(z_1 - \xi)\, p_2(z_2 - \xi) \\ &\propto \exp[-(z_1 - \xi)^2/2s_1] \exp[-(z_2 - \xi)^2/2s_2] \\ &= \exp[-A/2] \end{aligned}$$

with $A = (z_1 - \xi)^2/s_1 + (z_2 - \xi)^2/s_2$

$= (\xi - x^a)^2/p^a$ + *terms independent of* ξ

where $1/p^a = 1/s_1 + 1/s_2$, $x^a = p^a\,(z_1/s_1 + z_2/s_2)$

$$P(x = \xi \mid z_1, z_2) \propto \exp[-(\xi - x^a)^2/2p^a] = \mathcal{N}[x^a, p^a]$$

Conditional probability distribution of x, given z_1 and z_2 : $\mathcal{N}[x^a, p^a]$

Conditional probability distribution of x, given z_1 and z_2 : $\mathcal{N}[x^a, p^a]$

$1/p^a = 1/s_1 + 1/s_2$

$p^a <$ (s_1, s_2) independent of z_1 and z_2

$x^a = p^a\,(z_1/s_1 + z_2/s_2)$ is weighted average of z_1 and z_2, with respective weights $1/s_1$ and $1/s_2$. Larger weight is given to more accurate piece of data.

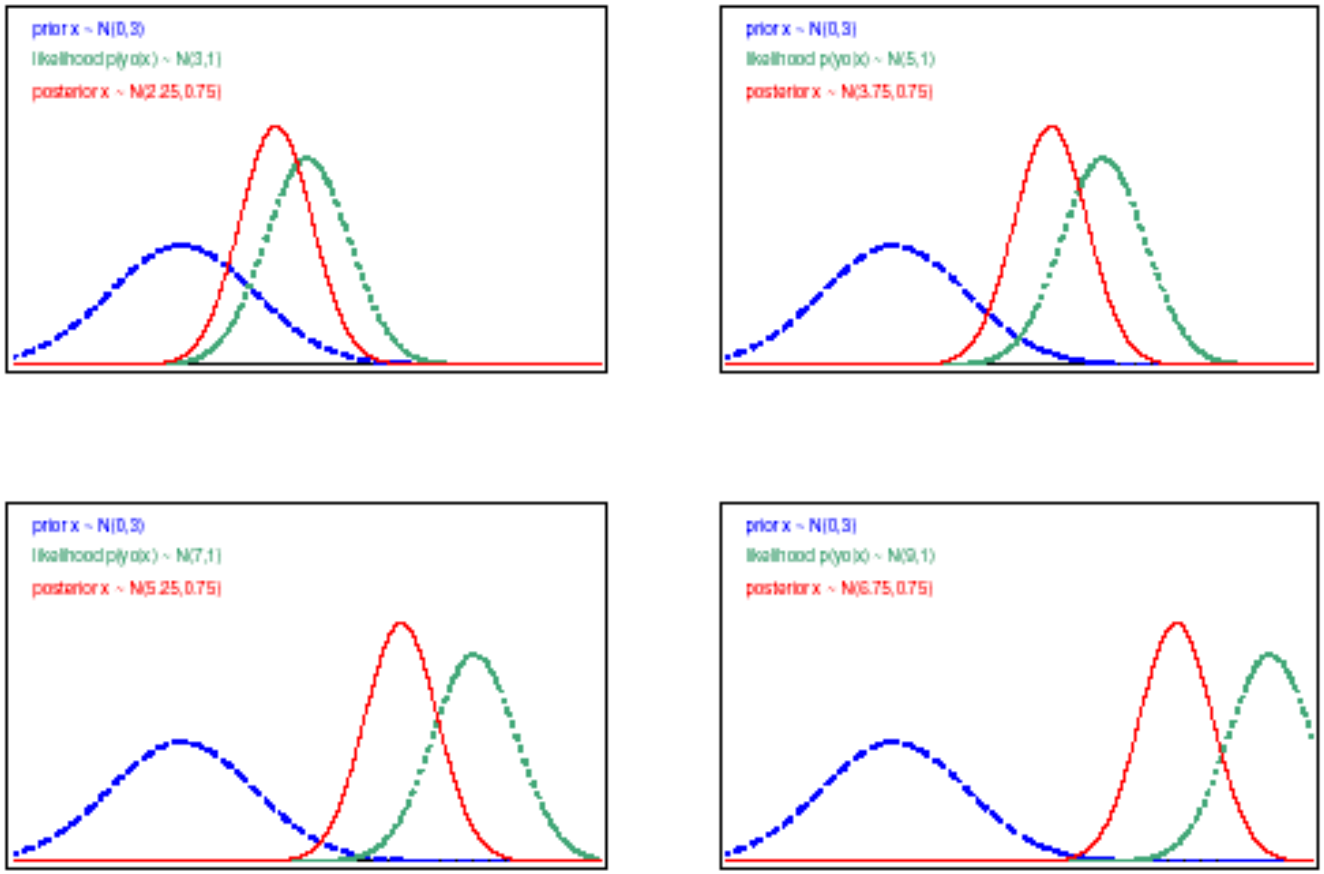


Fig. 1.1: Prior pdf $p(x)$ (dashed line), posterior pdf $p(x|y^o)$ (solid line), and Gaussian likelihood of observation $p(y^o|x)$ (dotted line), plotted against x for various values of y^o. (Adapted from Lorenc and Hammon 1988.)

Conditional expectation x^a minimizes following scalar *objective function*, defined on ξ-space

$$\xi \rightarrow \quad \mathcal{J}(\xi) \equiv \ (1/2)\ [(z_1 - \xi)^2 / s_1 + (z_2 - \xi)^2 / s_2\ ]$$

In addition

$$p^a = 1 / \mathcal{J}''(x^a)$$

Conditional probability distribution in Gaussian case

$$P(x = \xi \mid z_1, z_2) \propto \exp[\ -\ \underbrace{(\xi - x^a)^2 / 2p^a}_{\mathcal{J}(\xi) + Cst}\ ]$$

Estimate

$$x^a = p^a\,(z_1/s_1 + z_2/s_2)$$

with error p^a such that

$$1/p^a = 1/s_1 + 1/s_2$$

can also be obtained, independently of any Gaussian hypothesis, as simply corresponding to the linear combination of z_1 and z_2 that minimizes the error $E\,[(x^a - x)^2]$

Best Linear Unbiased Estimator (BLUE)