Sorbonne Université
Université Paris-Saclay
Master 2 Océan, Atmosphère, Climat et Observations Spatiales (OACOS)
Master 2 Water, Air, Pollution and Energy at local and regional scales (WAPE)

Année 2017-2018

Course *Introduction to data assimilation*

From numerical modelling to data assimilation

Olivier Talagrand

9 January 2018

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.

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 difficulties 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 is quite natural to pray for them, though they would consider it ridiculous to ask for an eclipse by prayer ? […] *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 that 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.*

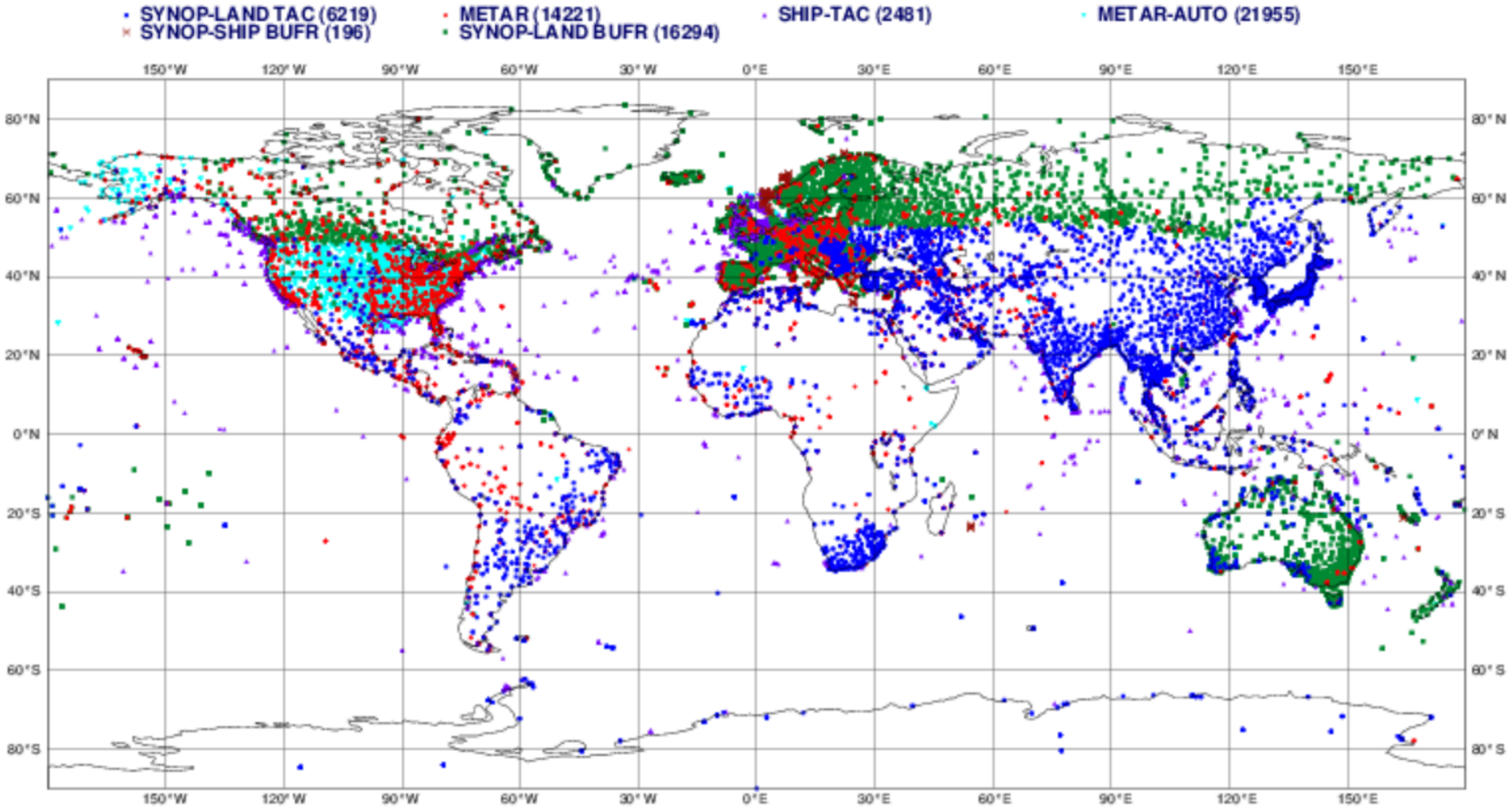
H. Poincaré, *Science et Méthode*, Paris, 1908

(translated Dover Publ., 1952)

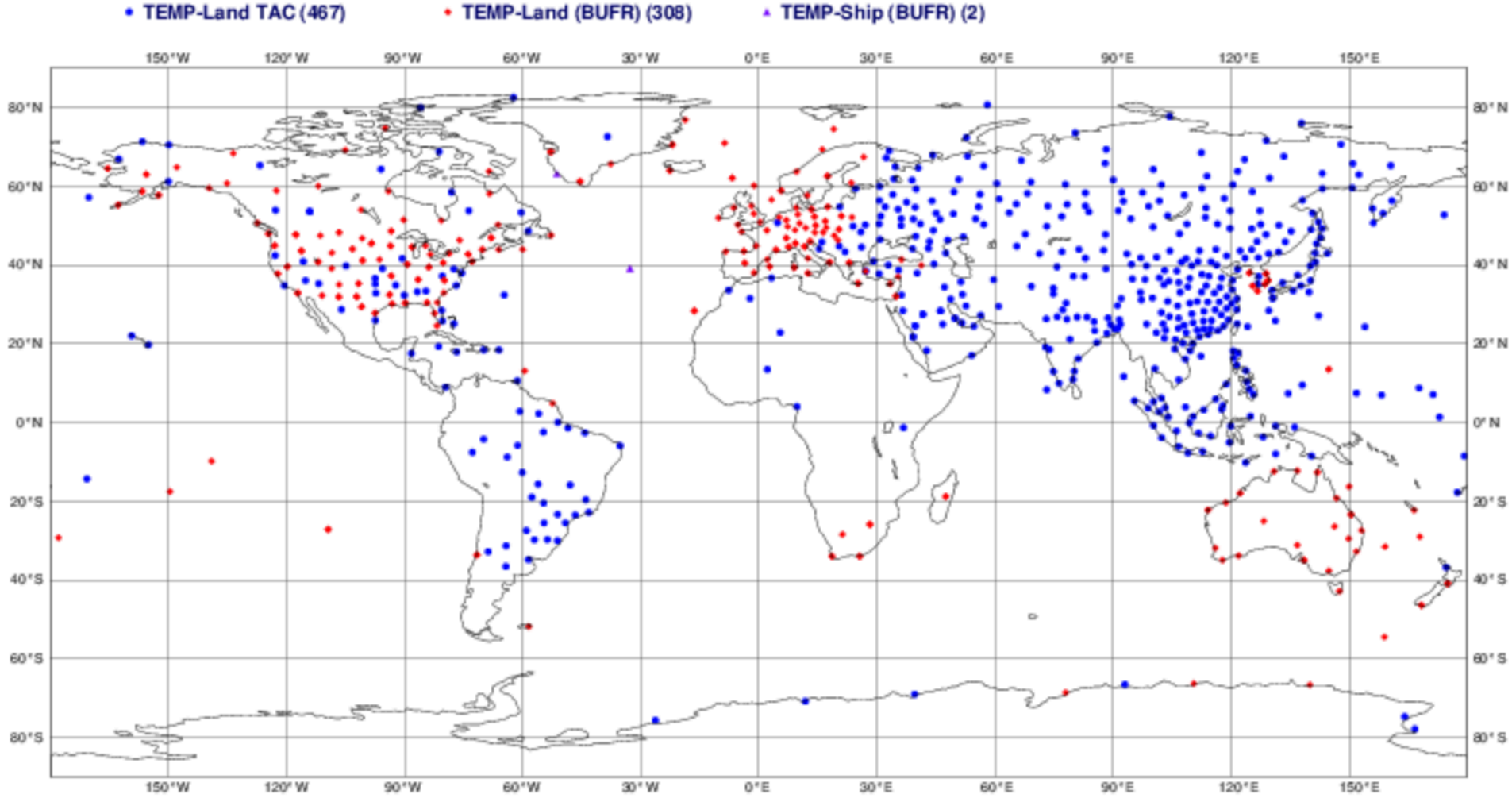
ECMWF data coverage (used observations) - SYNOP-SHIP-METAR

05/01/2018 00

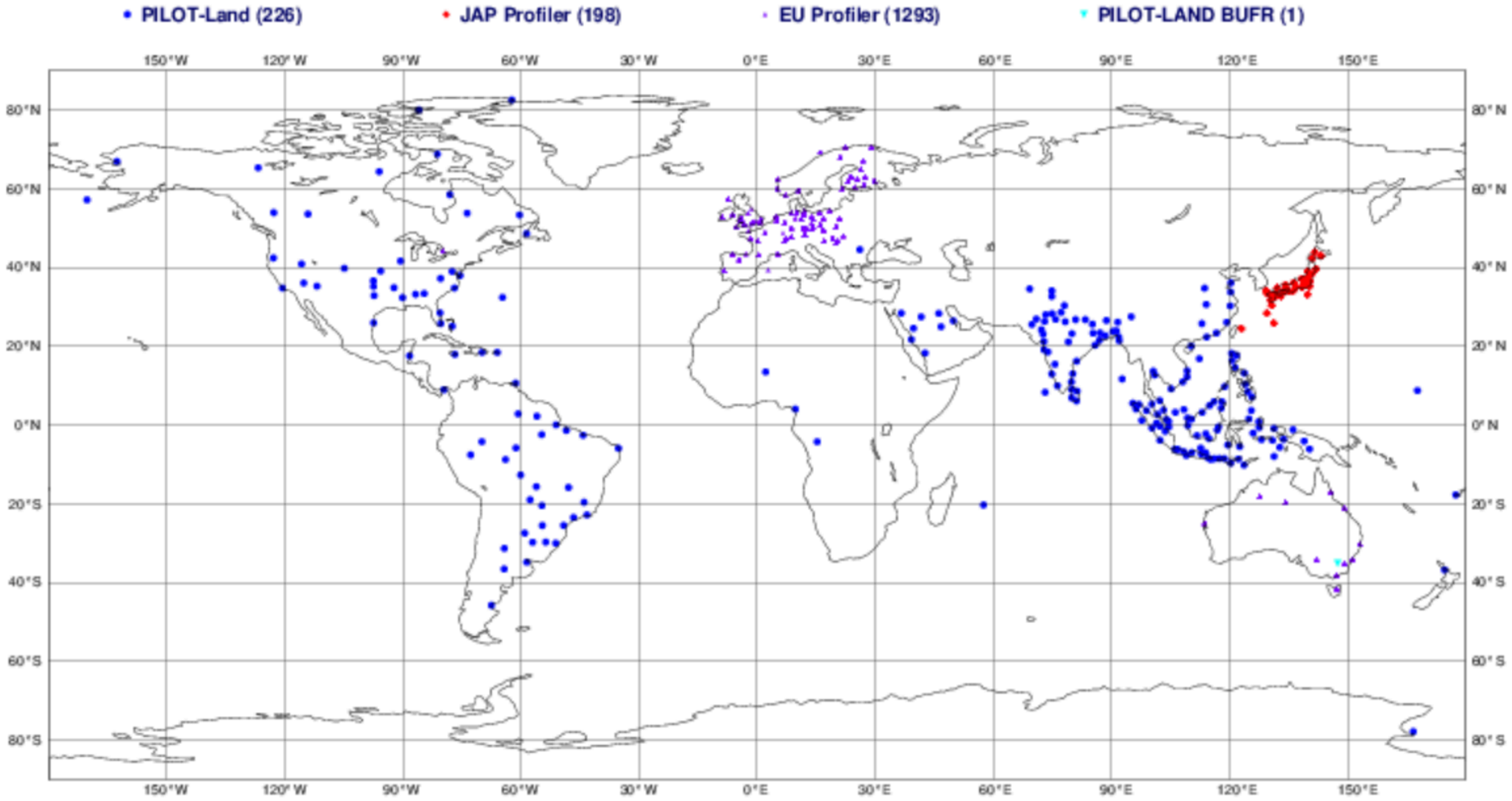
Total number of obs = 61366



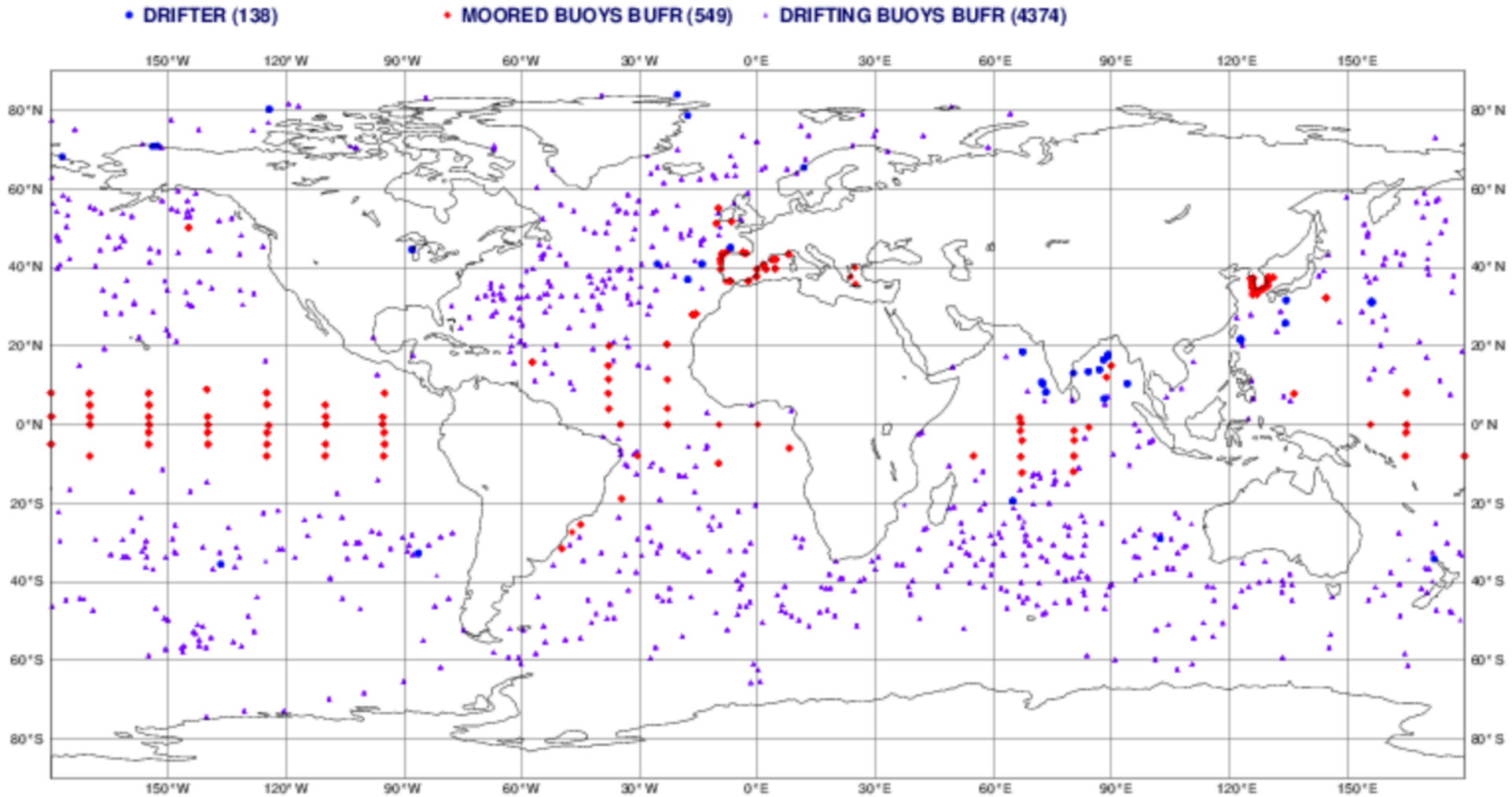
ECMWF data coverage (used observations) - RADIOSONDE
05/01/2018 00
Total number of obs = 777



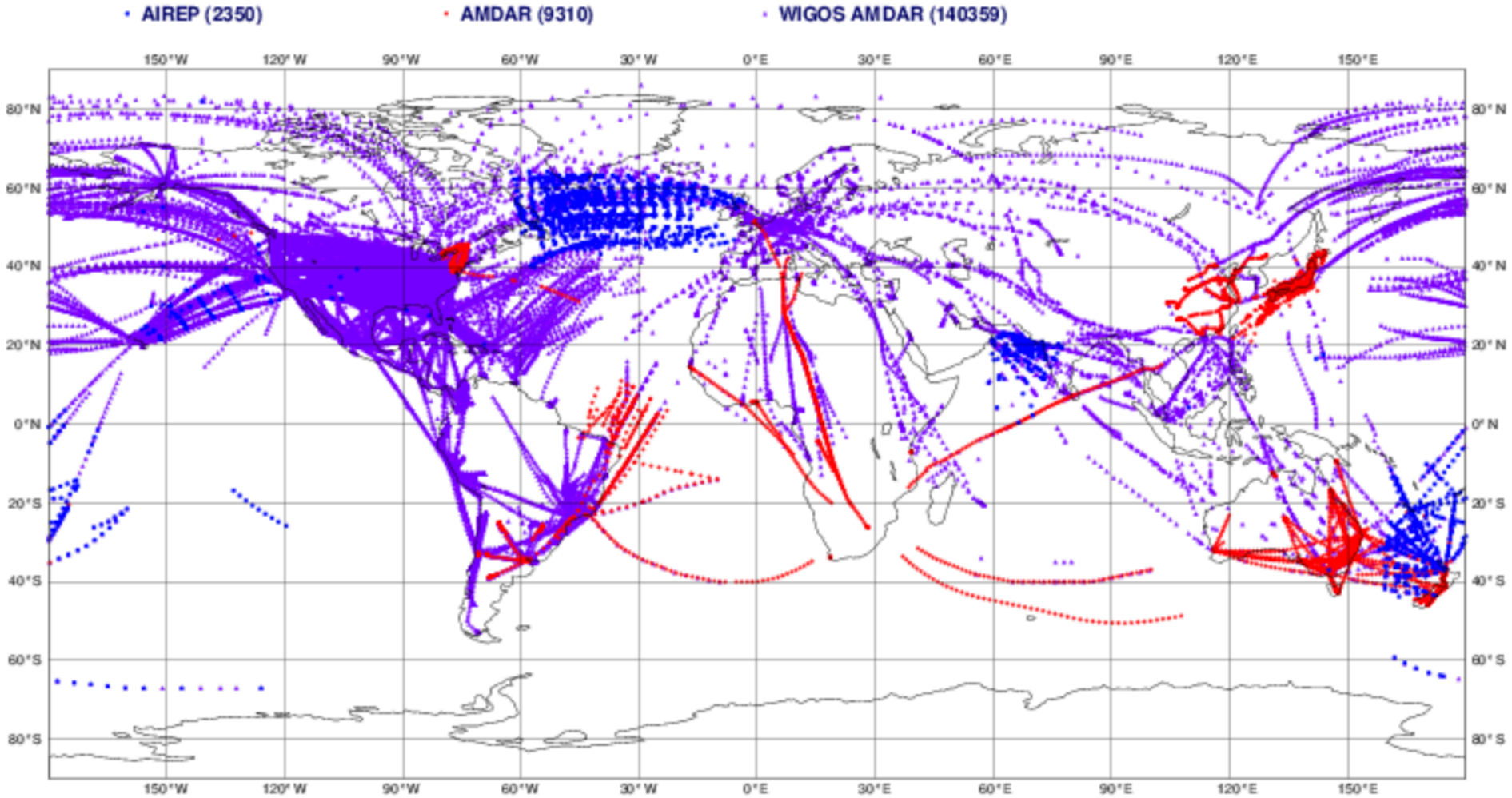
ECMWF data coverage (used observations) - PILOT
05/01/2018 00
Total number of obs = 1718



ECMWF data coverage (used observations) - BUOY
05/01/2018 00
Total number of obs = 5061



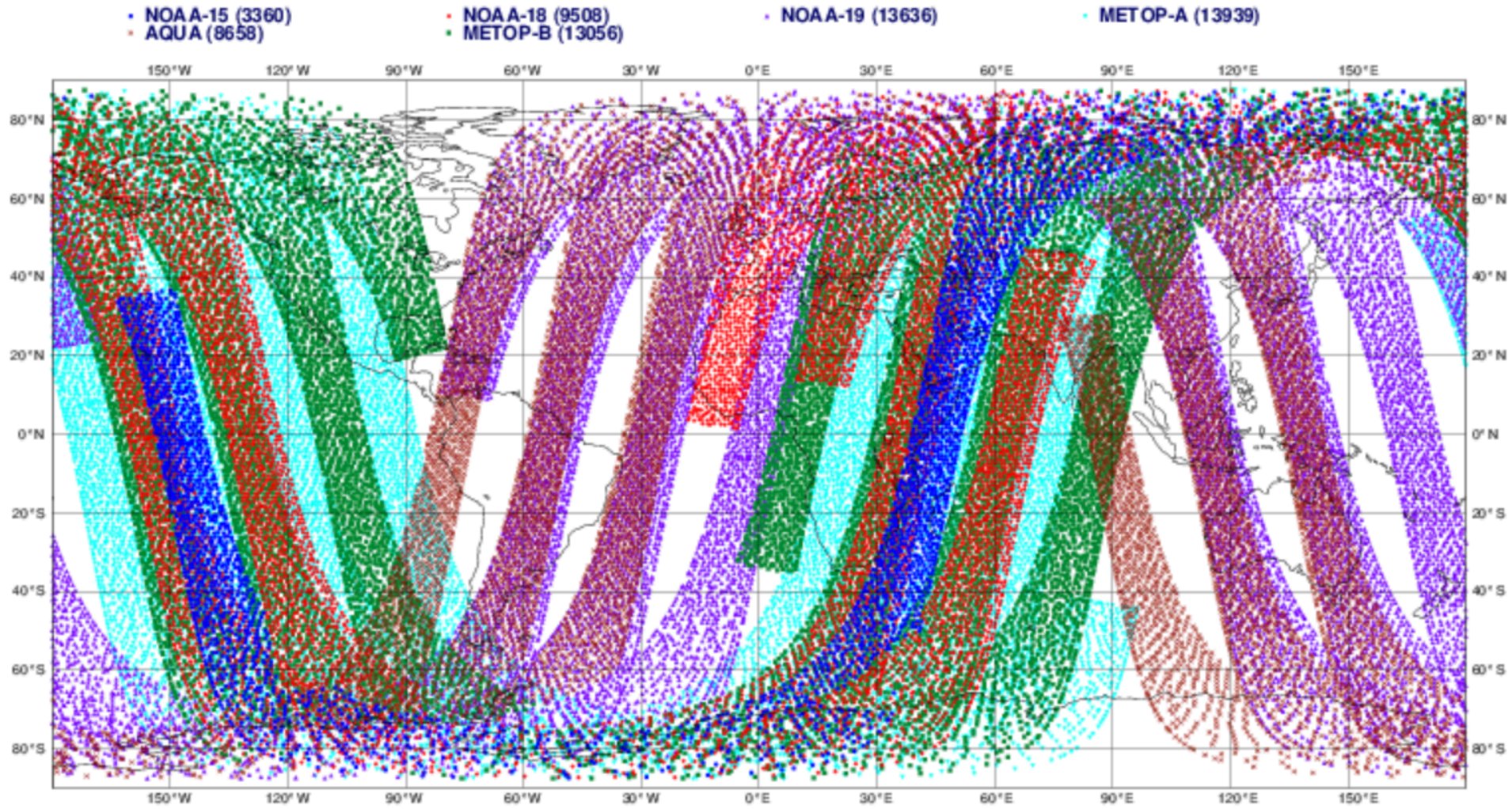
ECMWF data coverage (used observations) - AIRCRAFT
05/01/2018 00
Total number of obs = 152019



ECMWF data coverage (used observations) - AMSUA

05/01/2018 06

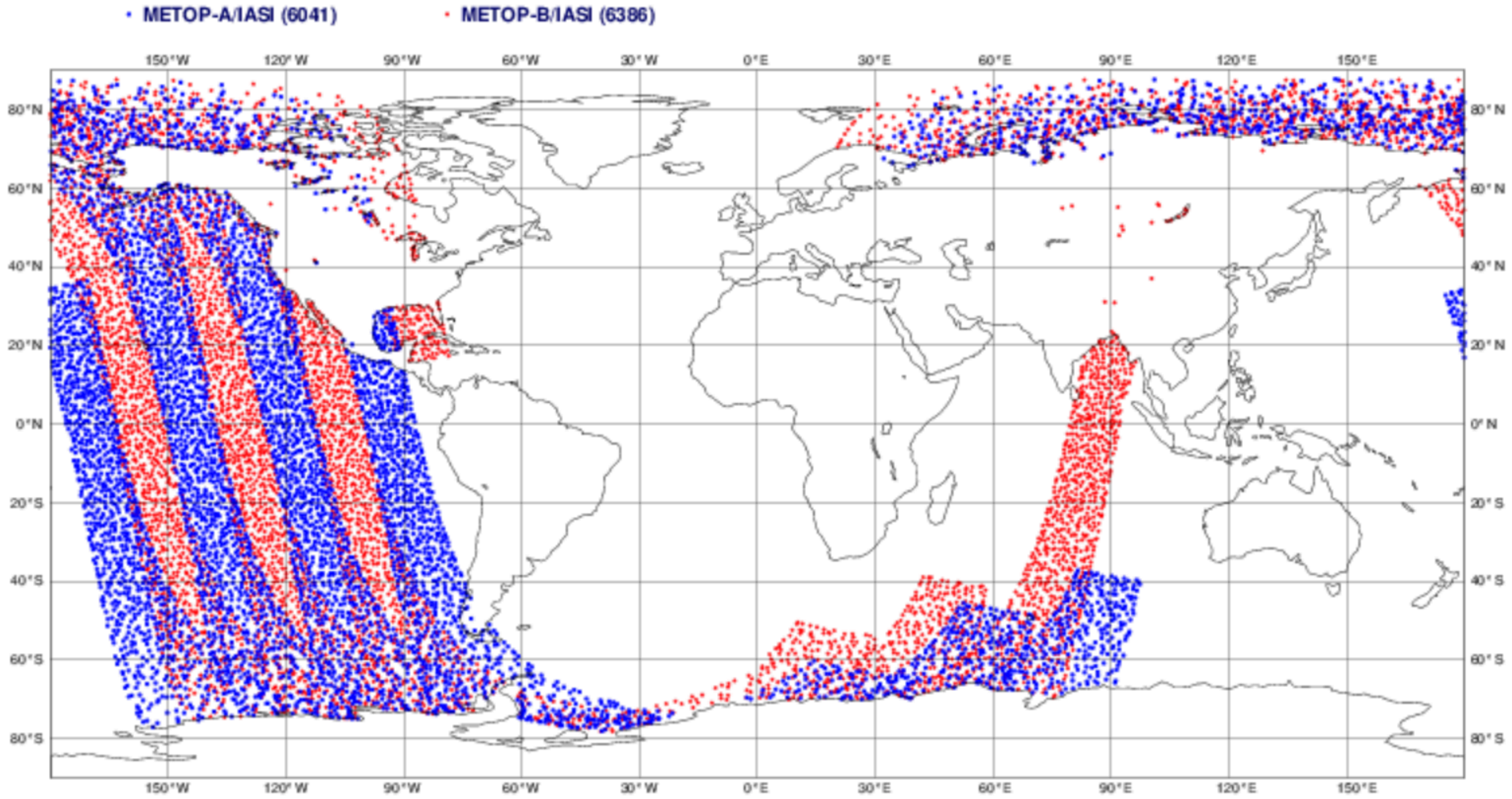
Total number of obs = 62157



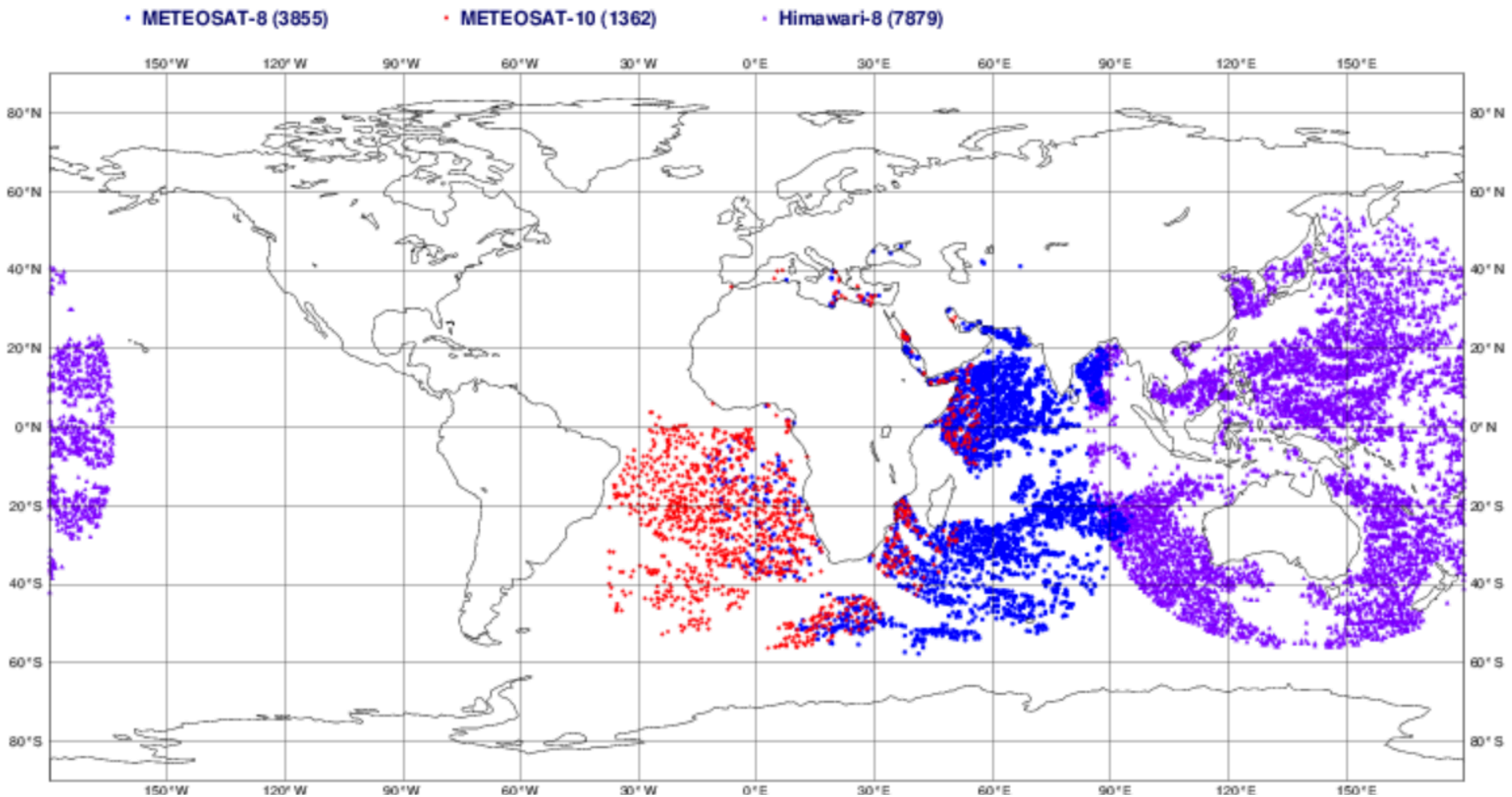
ECMWF data coverage (used observations) - IASI

05/01/2018 06

Total number of obs = 12427



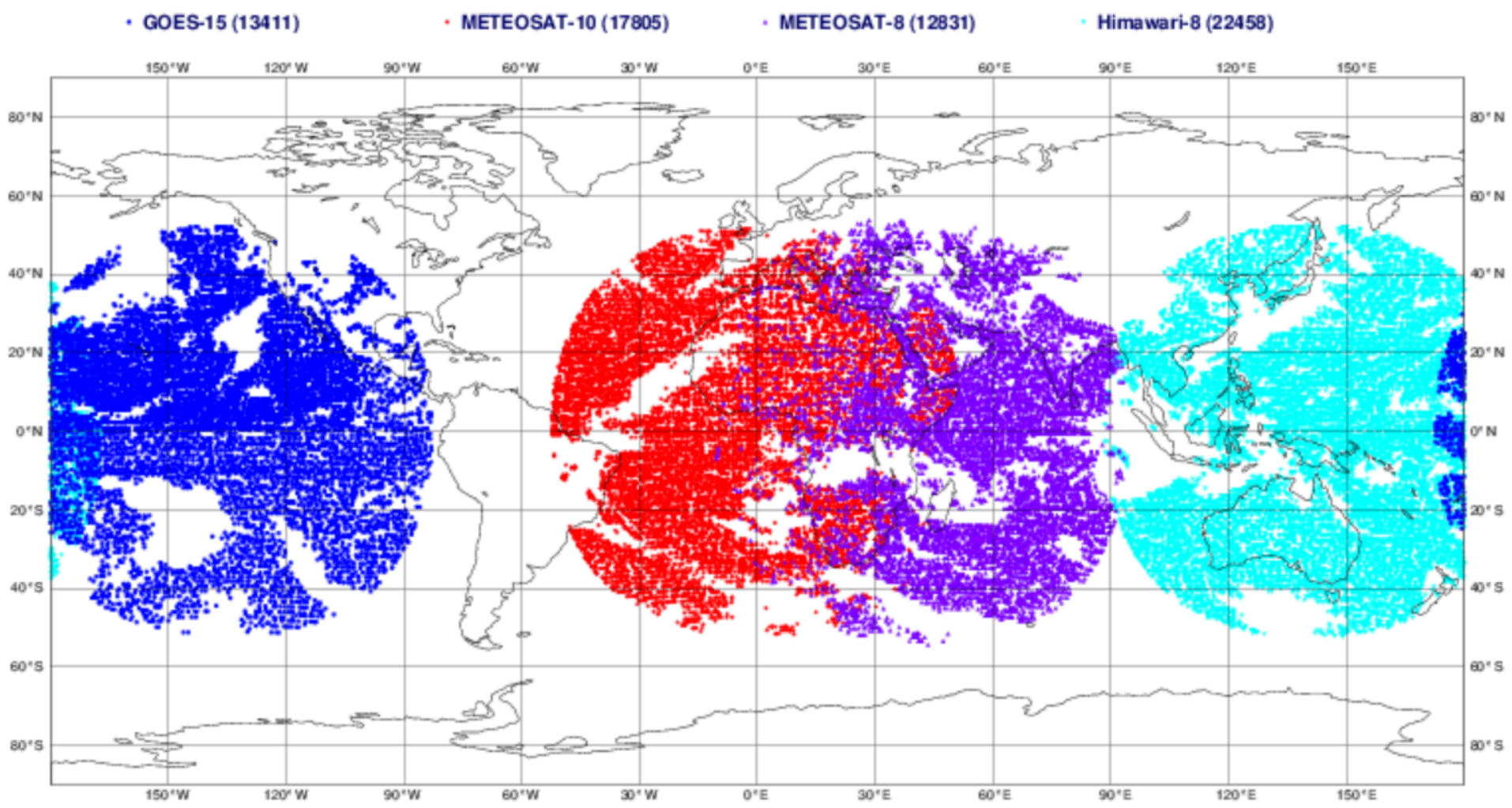
ECMWF data coverage (used observations) - AMV VIS
05/01/2018 06
Total number of obs = 13096



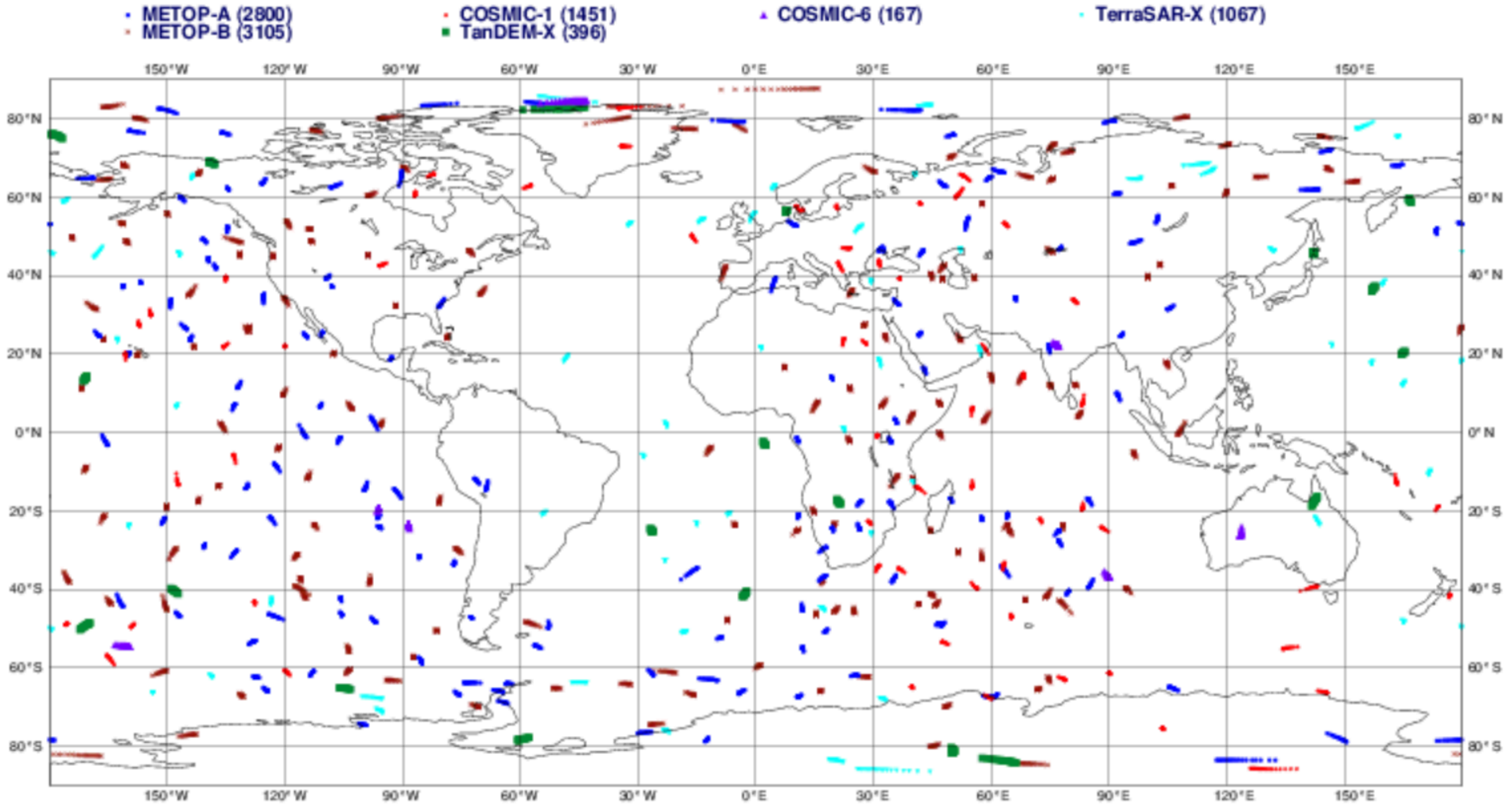
ECMWF data coverage (used observations) - GEOSTATIONARY RADIANCES

05/01/2018 06

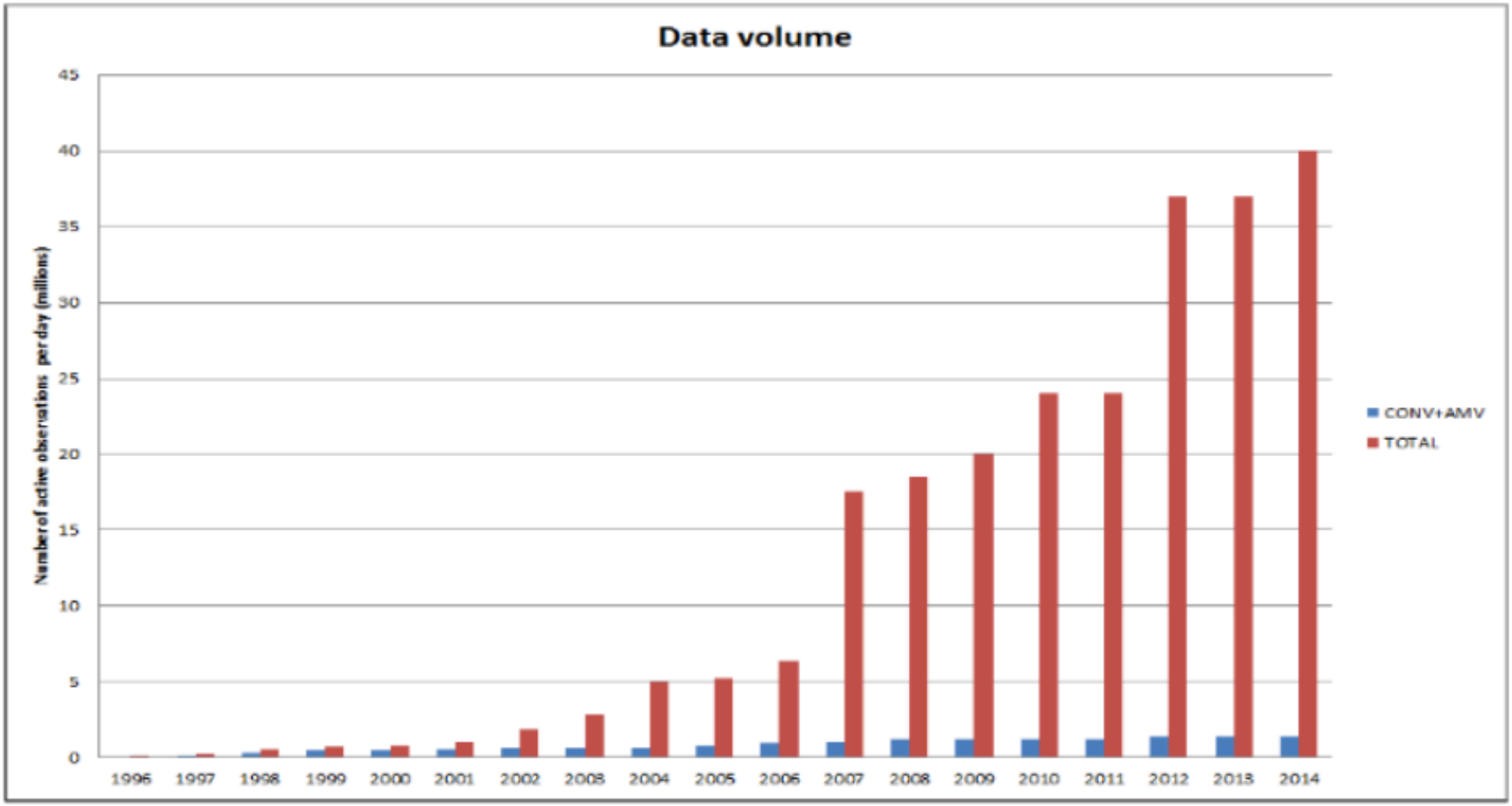
Total number of obs = 66505



ECMWF data coverage (used observations) - GPSRO
05/01/2018 06
Total number of obs = 8986



ECMWF


Data volume
Number of active observations per day (millions)
45
40
35
30
25
20
15
10
5
0
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
CONV+AMV
TOTAL

- *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.

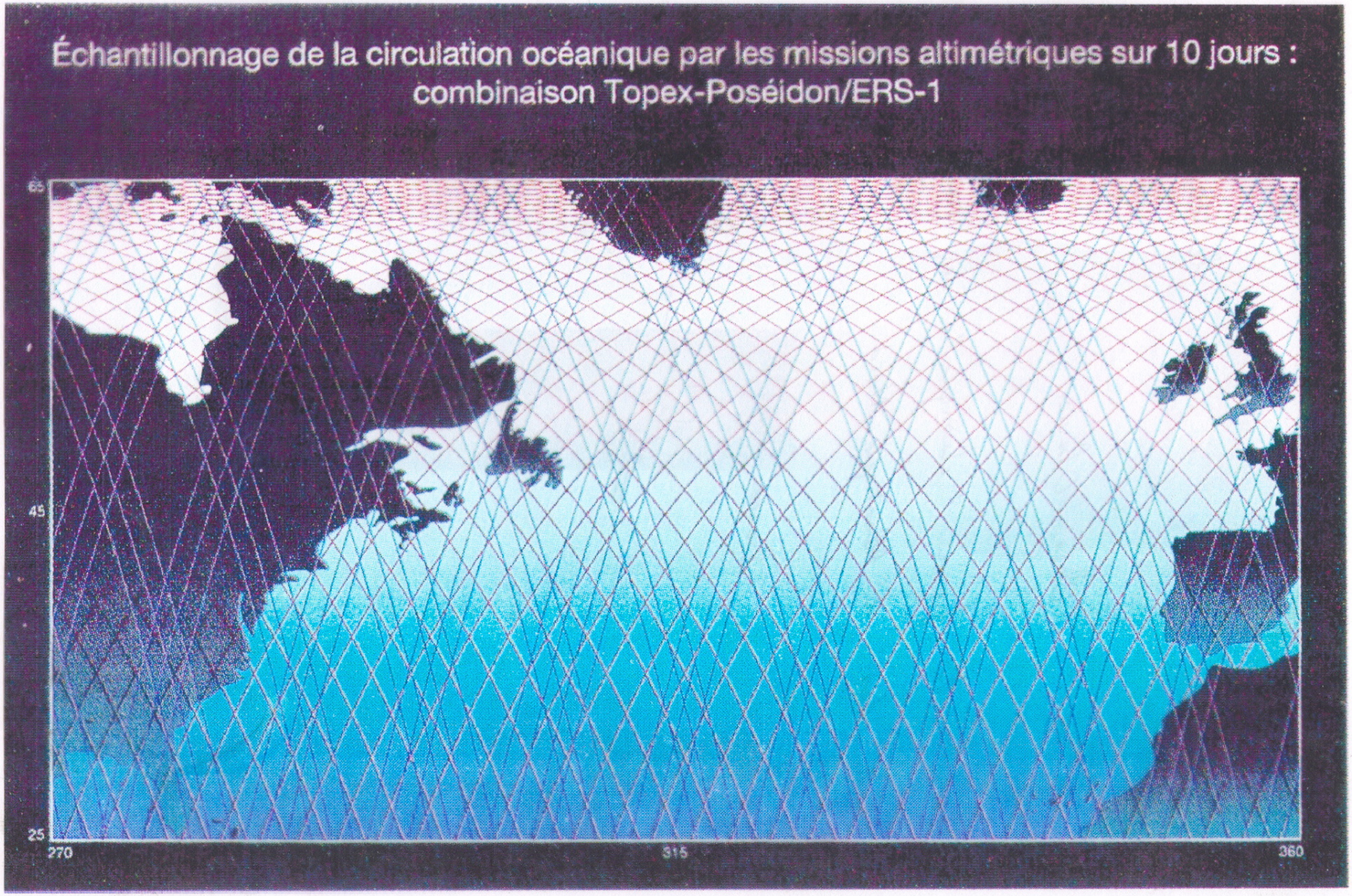
- *Asynoptic* observations (satellites, aircraft), performed more or less continuously in time.

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



S. Louvel, Doctoral Dissertation, 1999

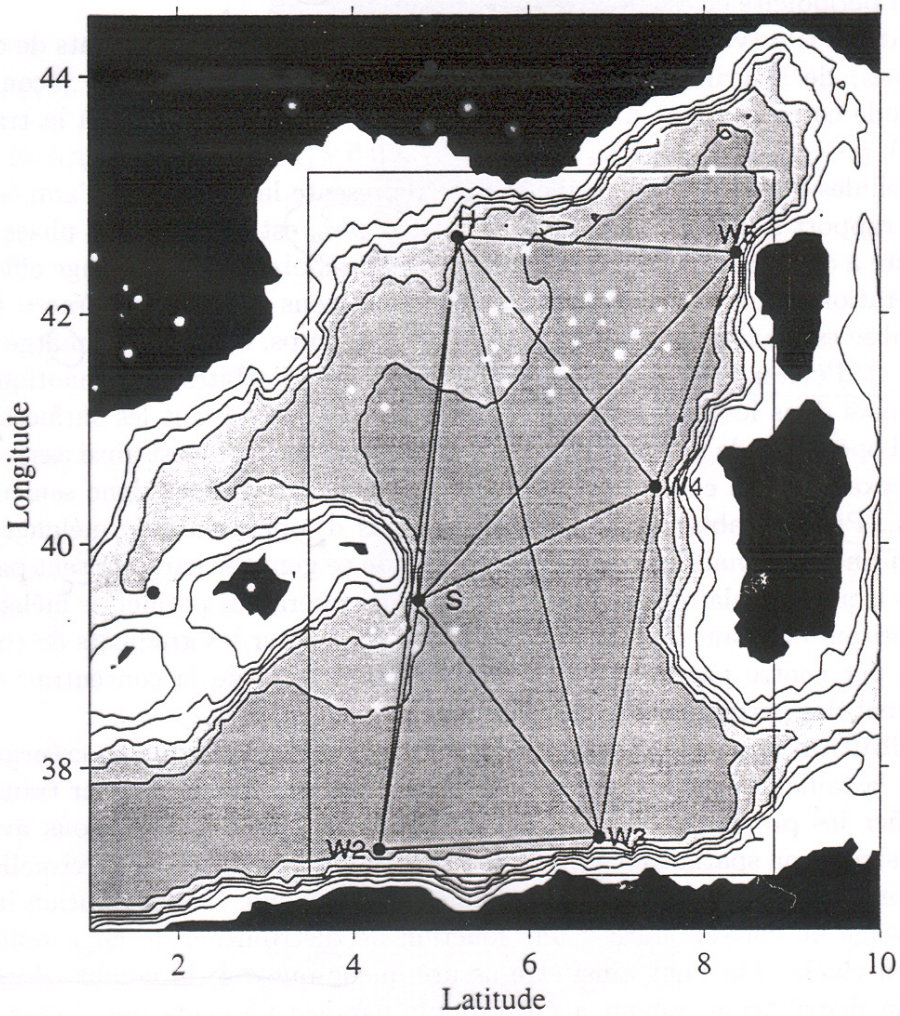


FIG. 1 – *Bassin méditerranéen occidental: réseau d'observation tomographique de l'expérience Thétis 2 et limites du domaine spatial utilisé pour les expériences numériques d'assimilation.*

E. Rémy, Doctoral Dissertation, 1999

Physical laws governing the flow

- 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 numerical models are built on 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). Most of them are non-hydrostatic, and therefore allow description of convection.

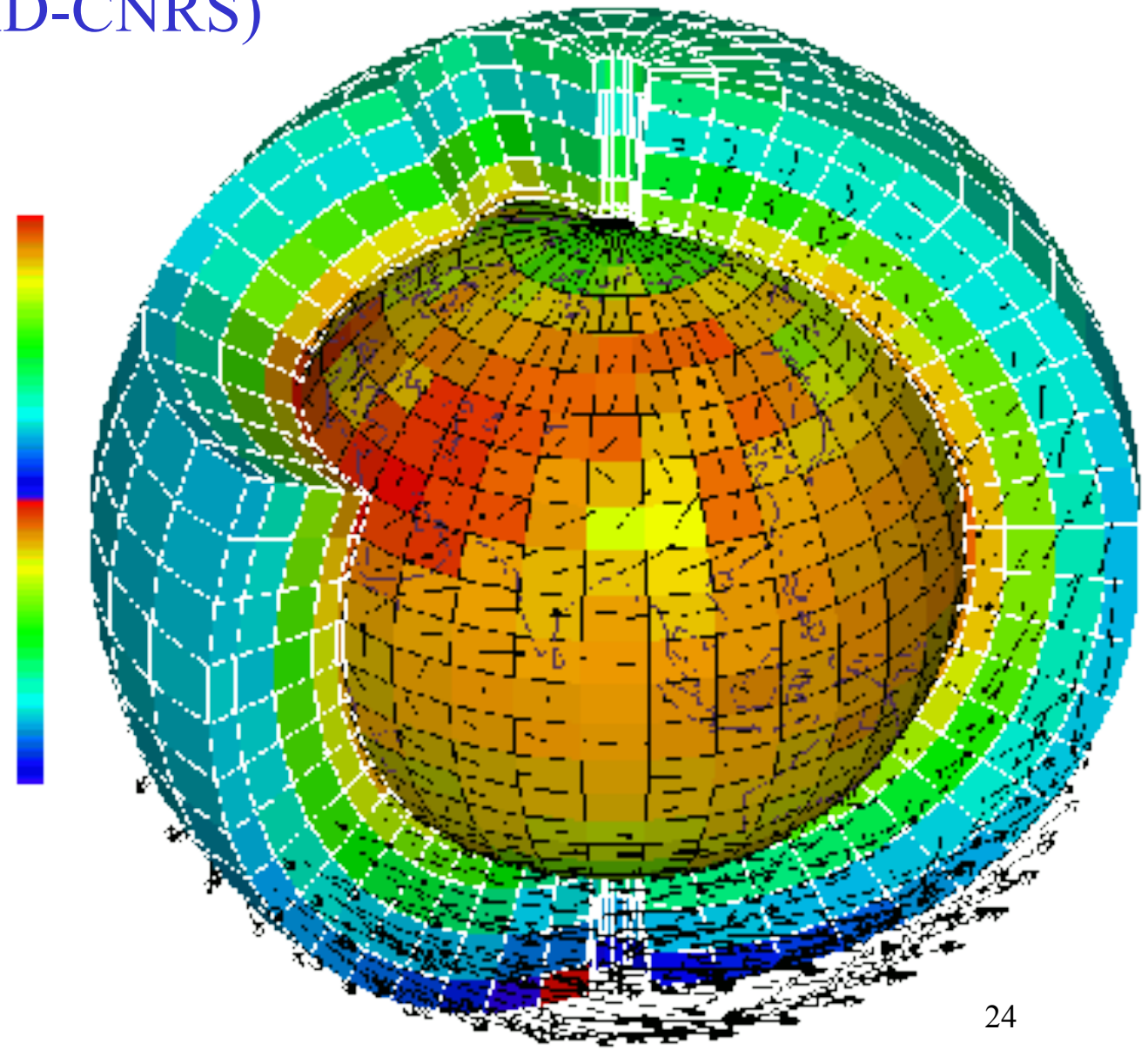
There exist at present two forms of discretization

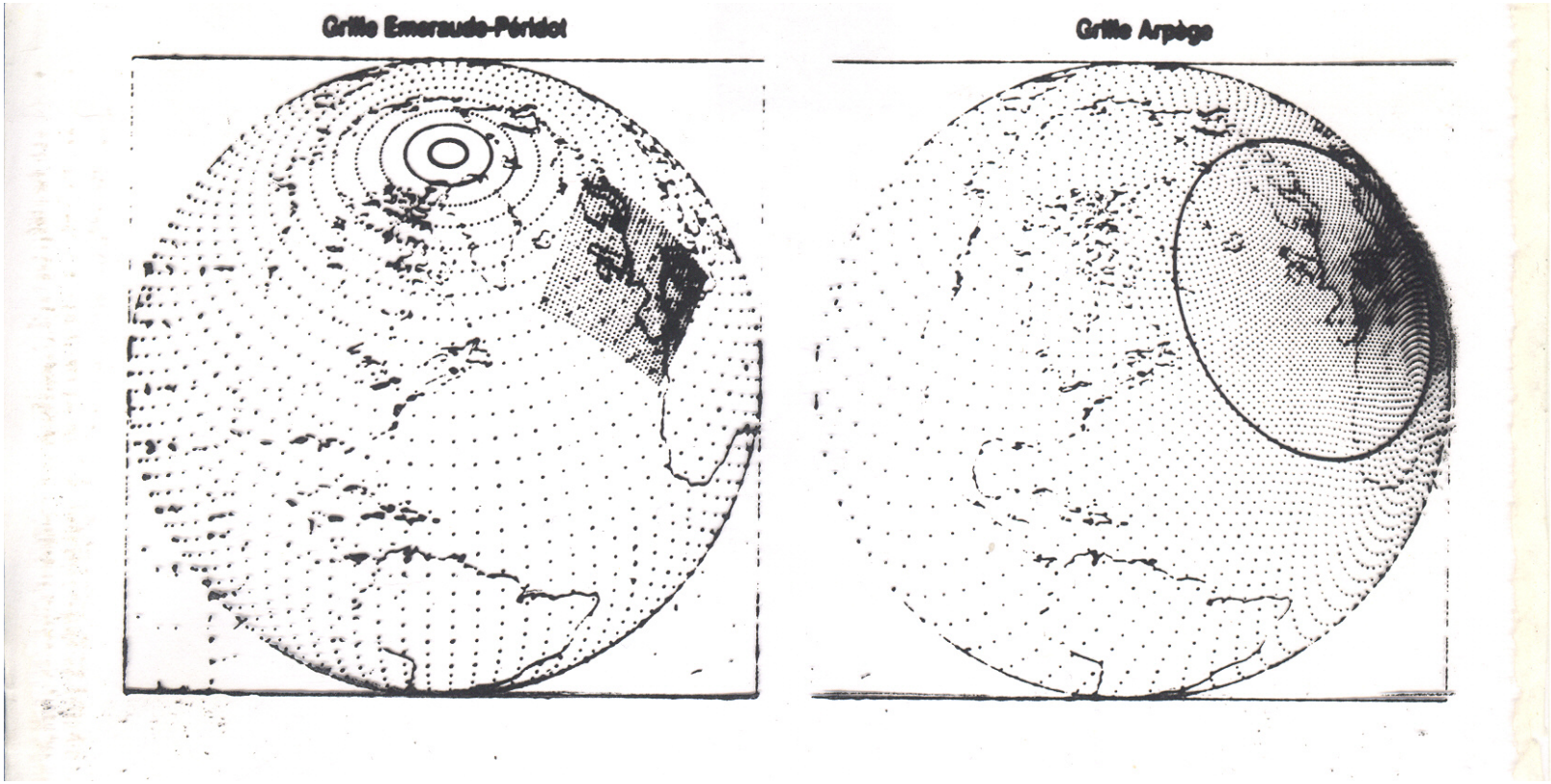
- Gridpoint discretization

- (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 into account the complicated geometry of coast-lines.

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





The grids of two of the models of Météo-France *(La Météorologie)*

In gridpoint models, meteorological fields are defined by values at the nodes of a 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*

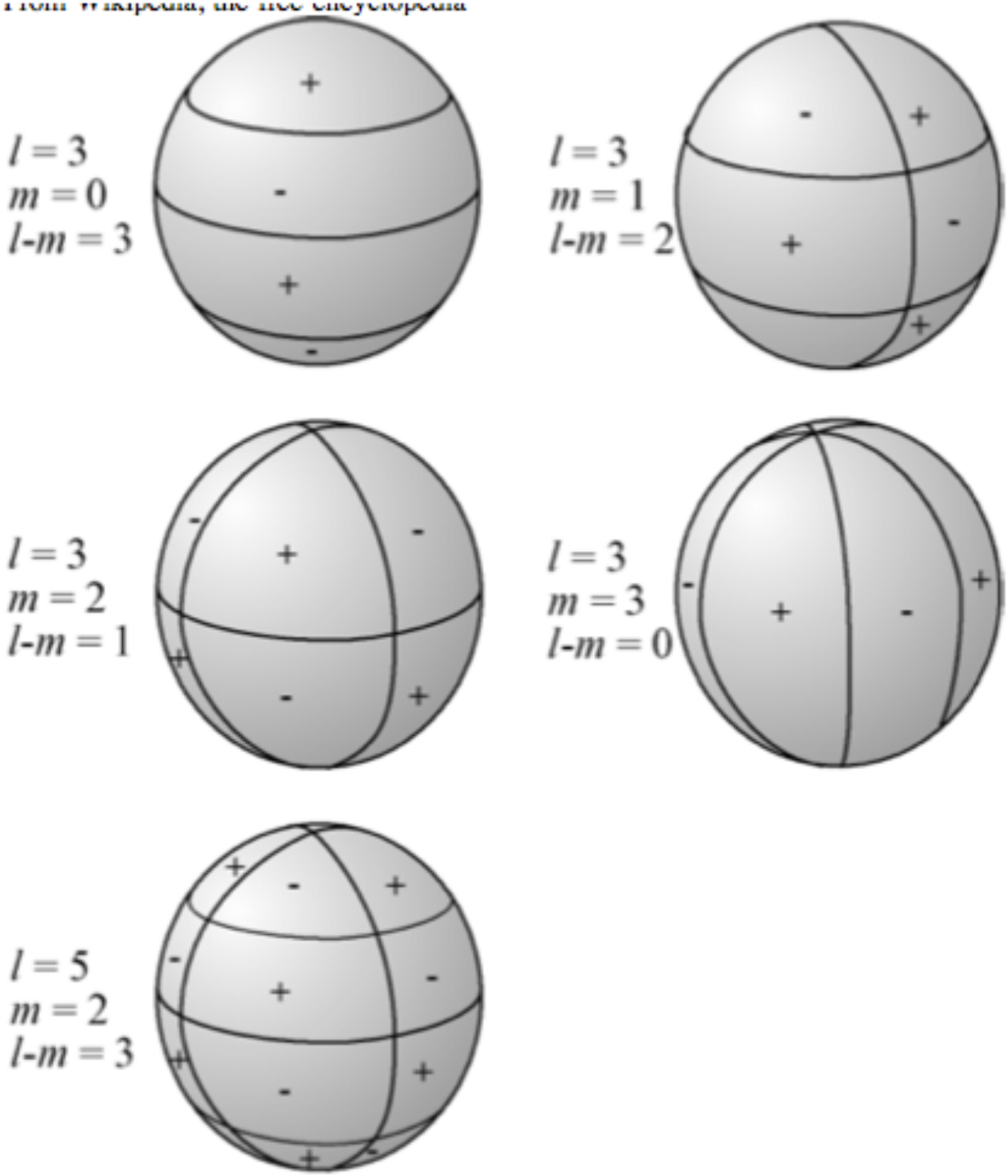
$$Y_n^m(\mu,\lambda) \propto P_n^m(\mu)\exp(im\lambda)$$

$P_n^m(\mu)$ 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 $Y_n^m(\mu,\lambda)$

From Wikipedia, the free encyclopedia



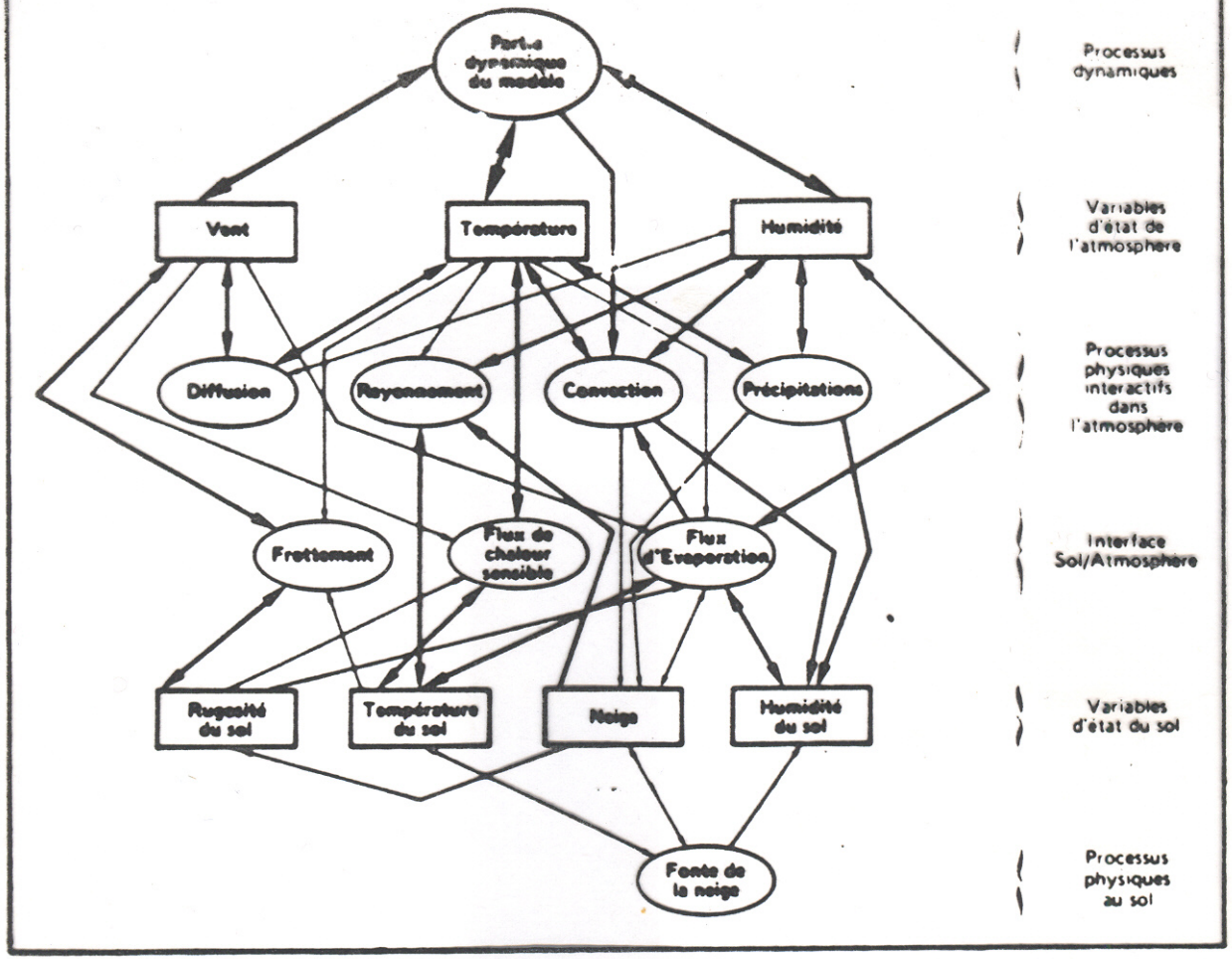
Linear operations, and in particular differentiation, are performed in spectral space, while nonlinear operations and ‘physical’ computations (advection, diabatic heating and cooling, …) are performed in gridpoint physical space. This requires constant transformations from one space to the other, which 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 progressively 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 subgrid scales parameterization, the water cycle and the associated exchanges of energy, or the exchanges between the atmosphere and the underlying medium). ‘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



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

(European Centre for Medium-range Weather Forecasts, ECMWF)

Depuis mars 2016 :

Troncature triangulaire TCO1279 / O1280 (résolution horizontale ≈ 9 kilomètres)

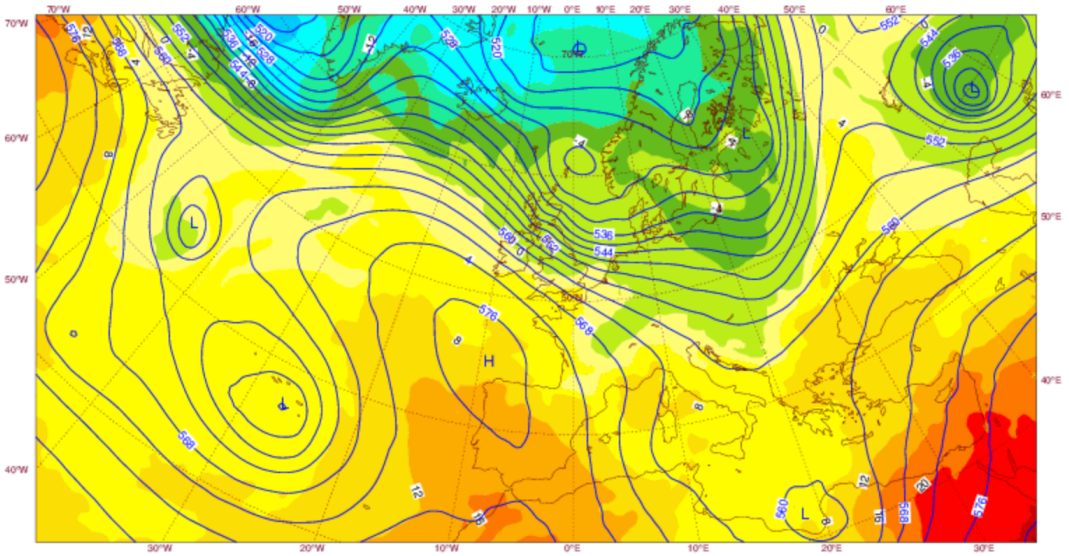
137 niveaux dans la direction verticale (0 - 80 km)

Discrétisation en éléments finis dans la direction verticale

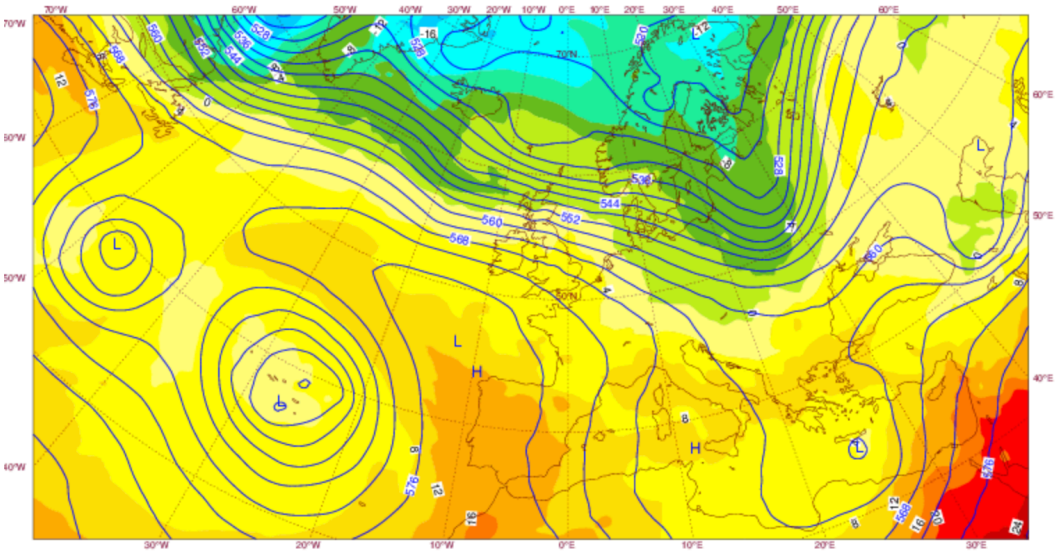
Dimension du vecteur d'état correspondant ≈ 4.10^9

Pas de discrétisation temporelle (schéma semi-Lagrangien semi-implicite): 450 secondes

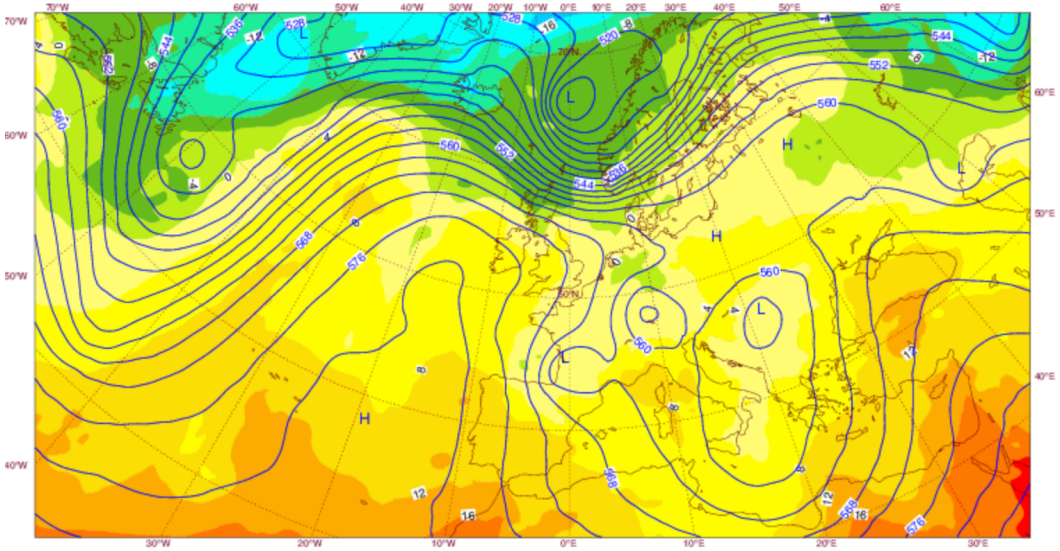
Wednesday 05 April 2017 0000 UTC ECMWF t+168 VT: Wednesday 12 April 2017 0000 UTC
850 hPa Temperature/500 hPa Geopotential



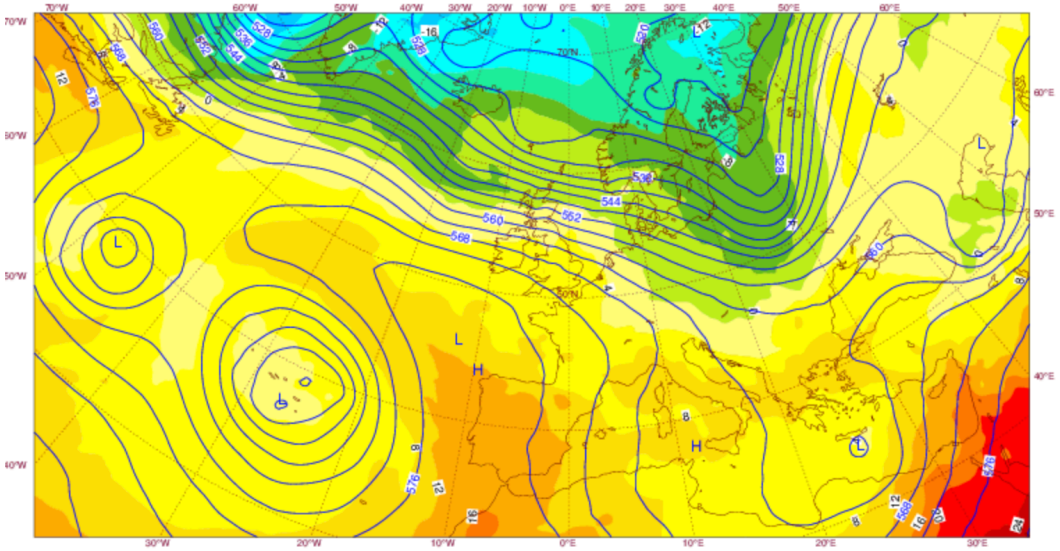
Wednesday 12 April 2017 0000 UTC ECMWF t+0 VT: Wednesday 12 April 2017 0000 UTC
850 hPa Temperature/500 hPa Geopotential



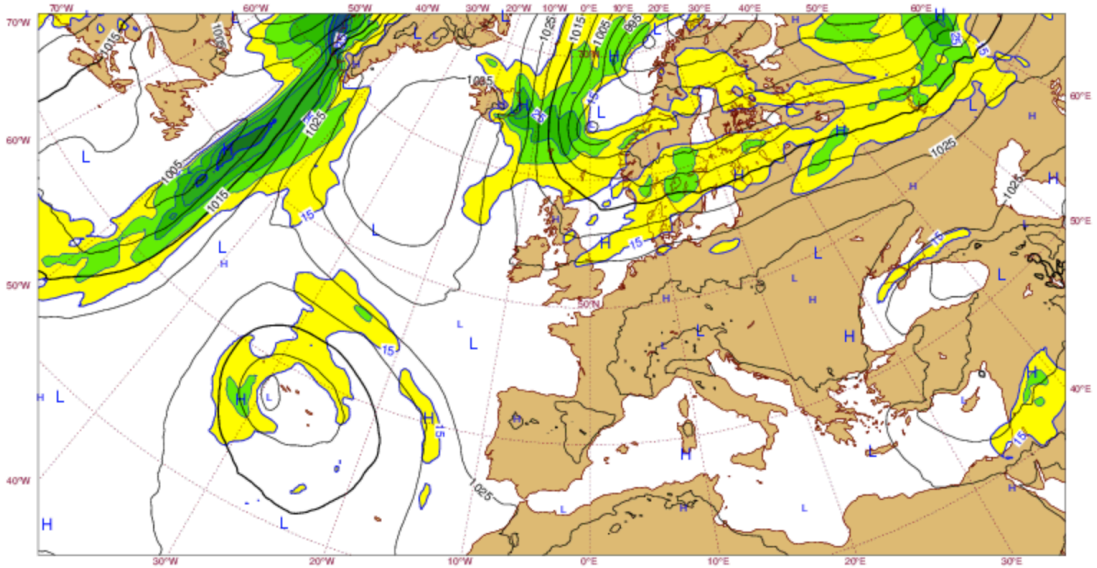
Wednesday 05 April 2017 0000 UTC ECMWF t+0 VT: Wednesday 05 April 2017 0000 UTC
850 hPa Temperature/500 hPa Geopotential



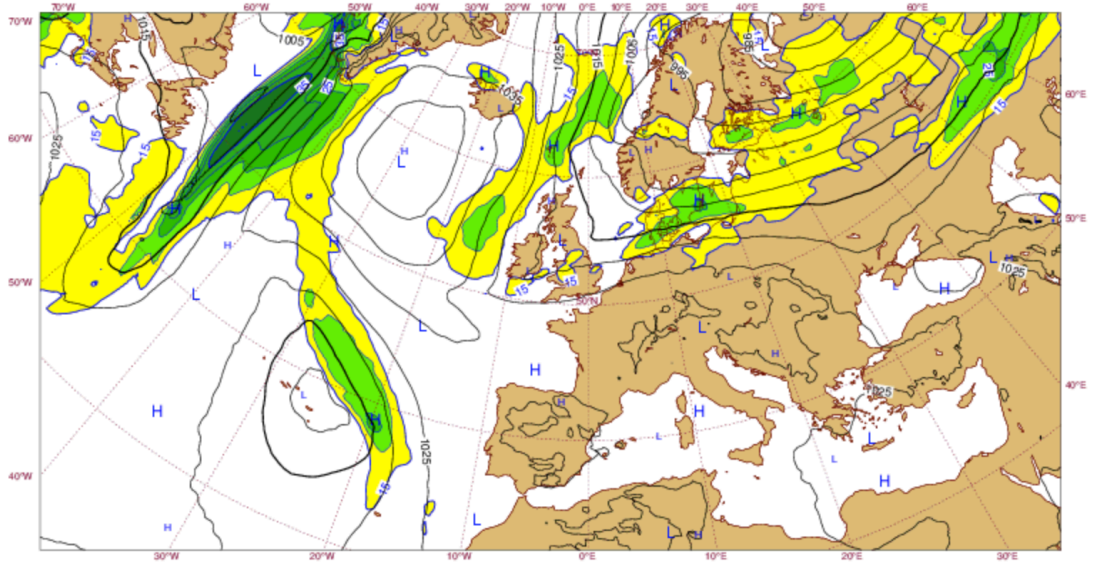
Wednesday 12 April 2017 0000 UTC ECMWF t+0 VT: Wednesday 12 April 2017 0000 UTC
850 hPa Temperature/500 hPa Geopotential



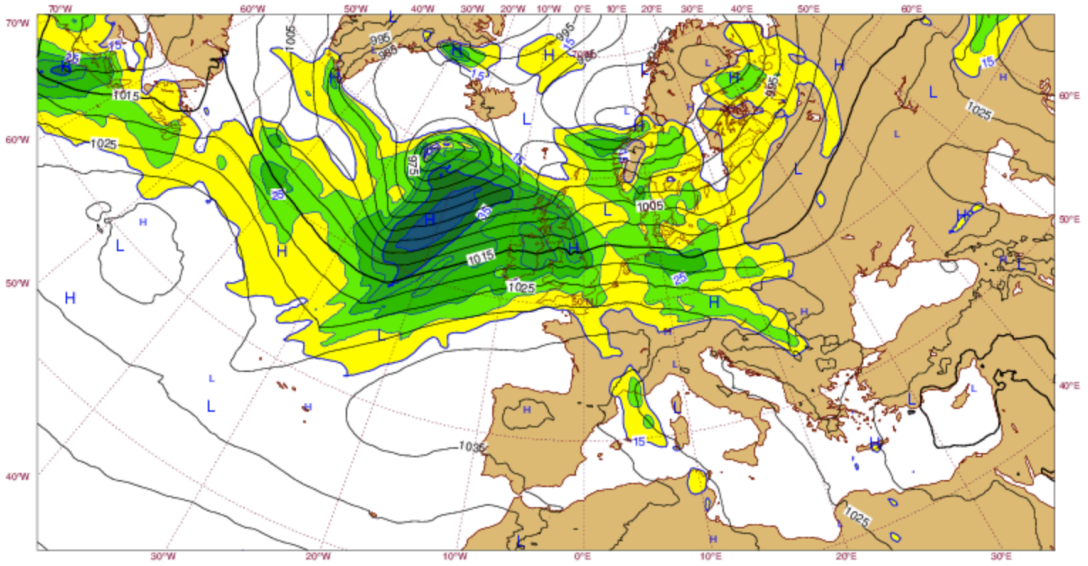
Sunday 25 December 2016 0000 UTC ECMWF t+168 VT: Sunday 01 January 2017 0000 UTC
Surface: Mean sea level pressure / 850hPa wind speed



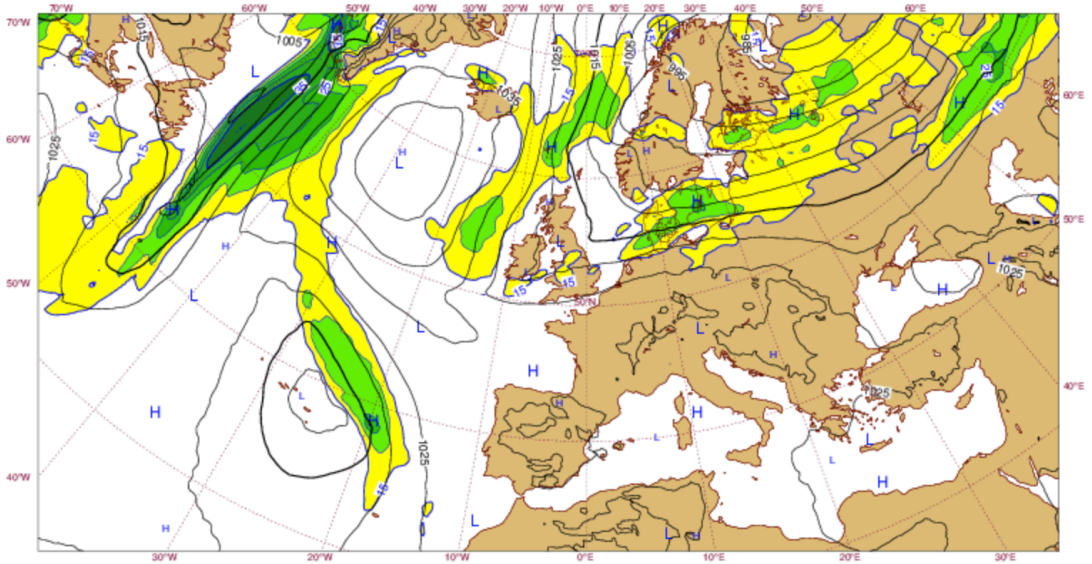
Sunday 01 January 2017 0000 UTC ECMWF t+0 VT: Sunday 01 January 2017 0000 UTC
Surface: Mean sea level pressure / 850hPa wind speed



Sunday 25 December 2016 0000 UTC ECMWF t+0 VT: Sunday 25 December 2016 0000 UTC
Surface: Mean sea level pressure / 850hPa wind speed



Sunday 01 January 2017 0000 UTC ECMWF t+0 VT: Sunday 01 January 2017 0000 UTC
Surface: Mean sea level pressure / 850hPa wind speed



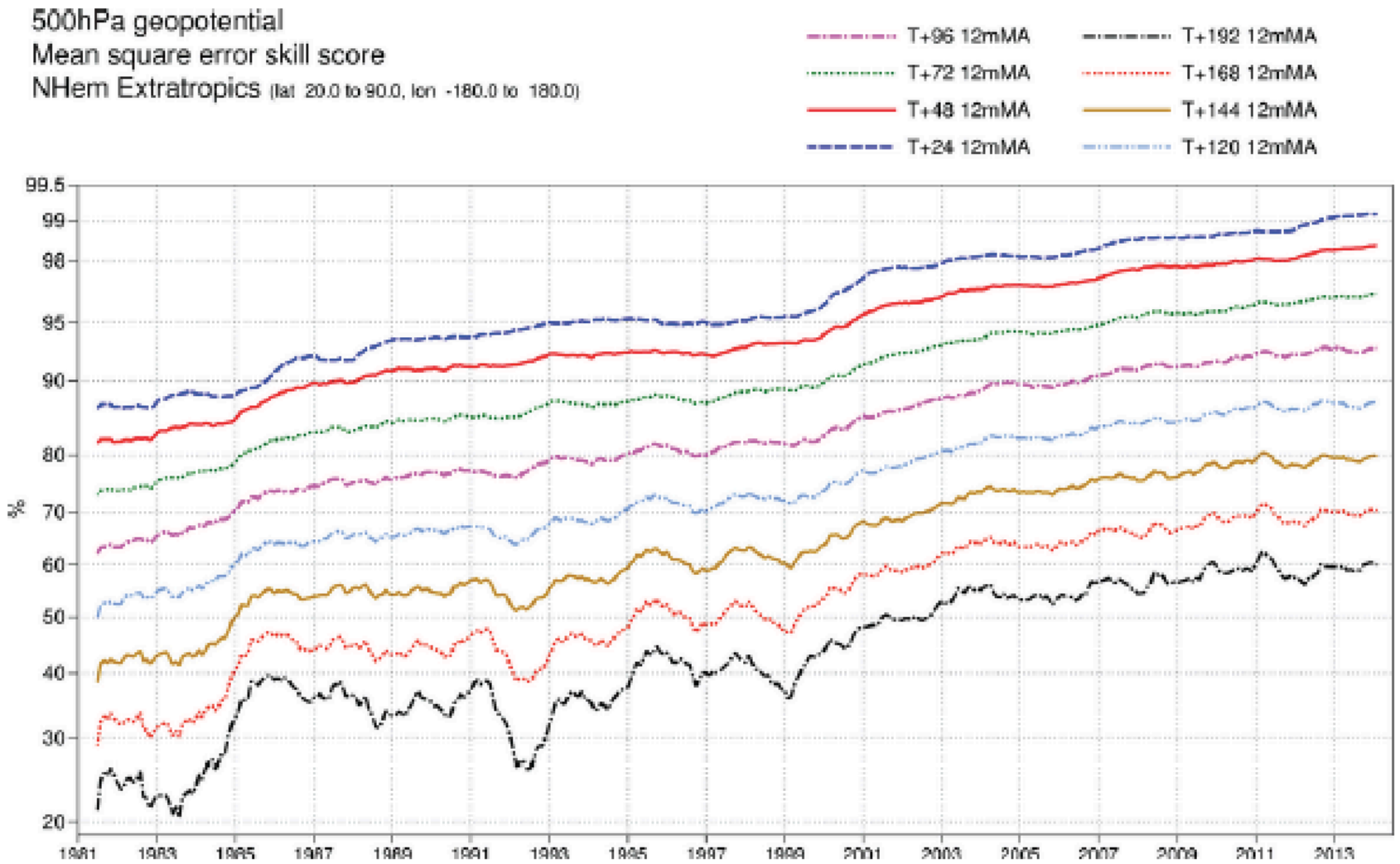
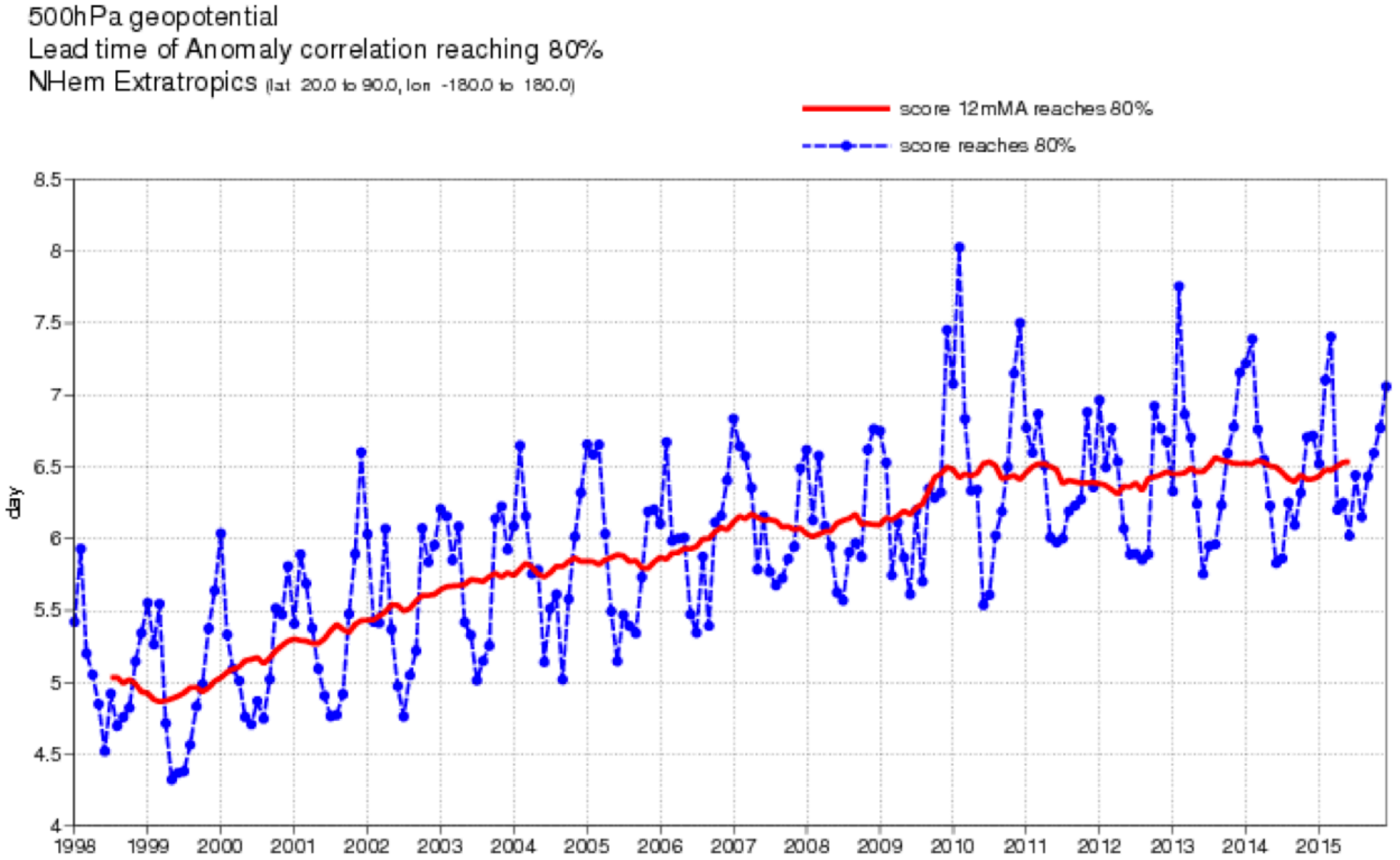


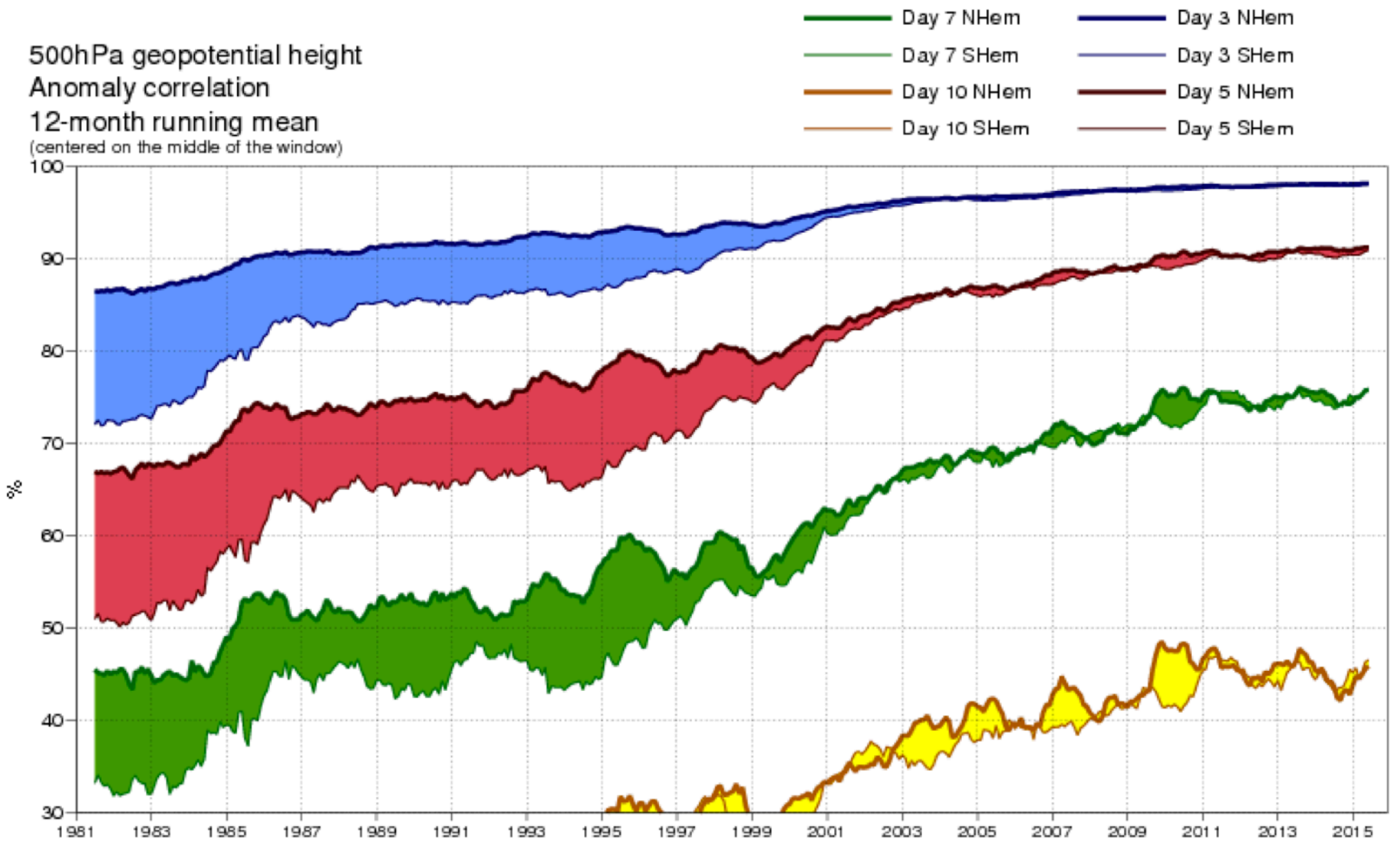
Figure 3: 500 hPa geopotential height mean square error skill score for Europe (top) and the northern hemisphere extratropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2013–July 2014.

Persistence = 0 ; climatology = 50 at long range

http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800tm742.pdf



The plot shows for each month the range at which the month mean (blue line) or 12-month mean centred on that month (red line) of forecast anomaly correlation dropped below 80% (or 60%). The


500hPa geopotential height
Anomaly correlation
12-month running mean
(centered on the middle of the window)
Day 7 NHem
Day 7 SHem
Day 10 NHem
Day 10 SHem
Day 3 NHem
Day 3 SHem
Day 5 NHem
Day 5 SHem
%
100
90
80
70
60
50
40
30
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015

ECMWF Technical Memorandum 792

http://www.ecmwf.int/sites/default/files/elibrary/2016/16924-evaluation-ecmwf-forecasts-including-2016-resolution-upgrade.pdf

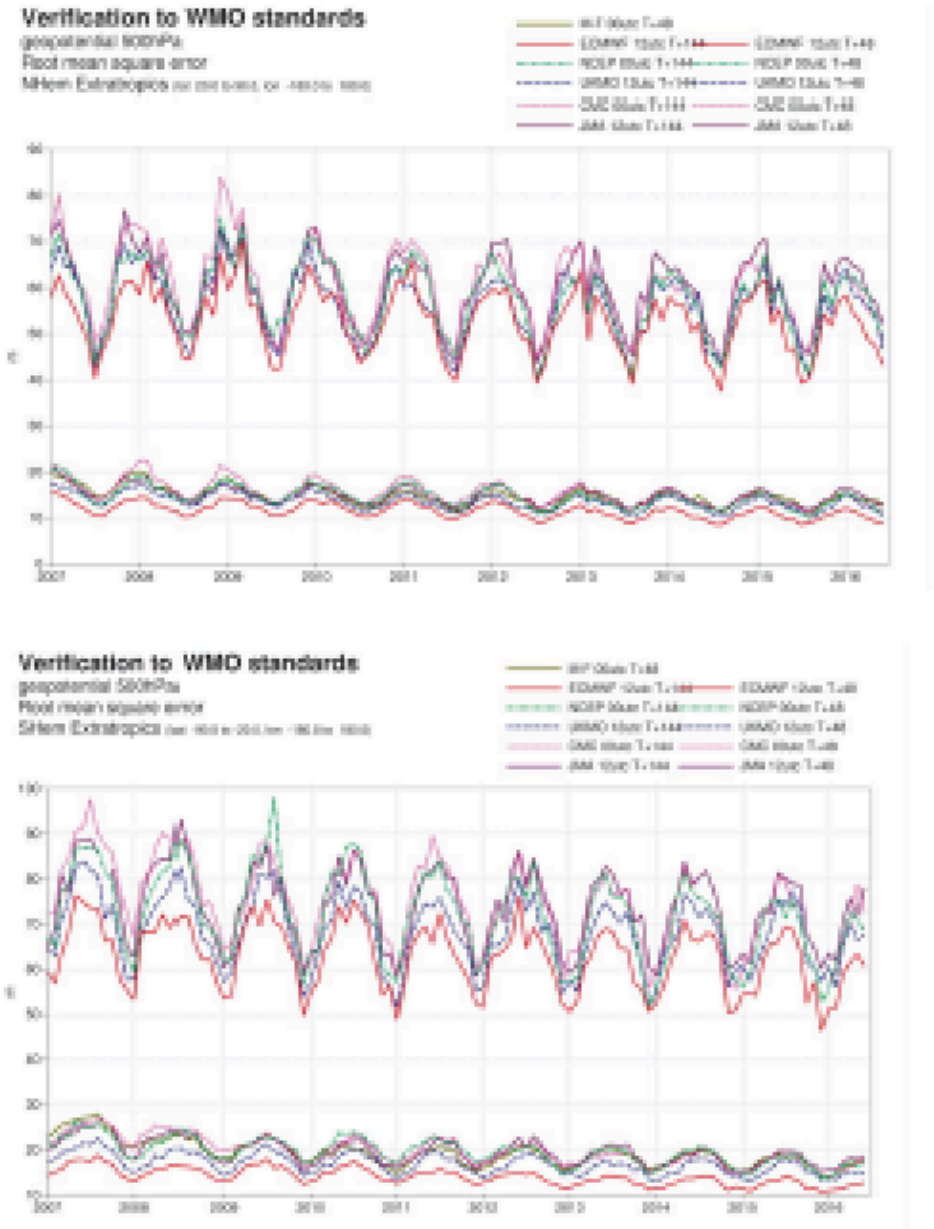


Figure 12: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top) and southern (bottom) extratropics. In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model is verified against its own analysis. JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, NCEP = U.S. National Centers for Environmental Prediction, M-F = Météo France.

ECMWF Technical Memorandum 792

http://www.ecmwf.int/sites/default/files/elibrary/2016/16924-evaluation-ecmwf-forecasts-including-2016-resolution-upgrade.pdf

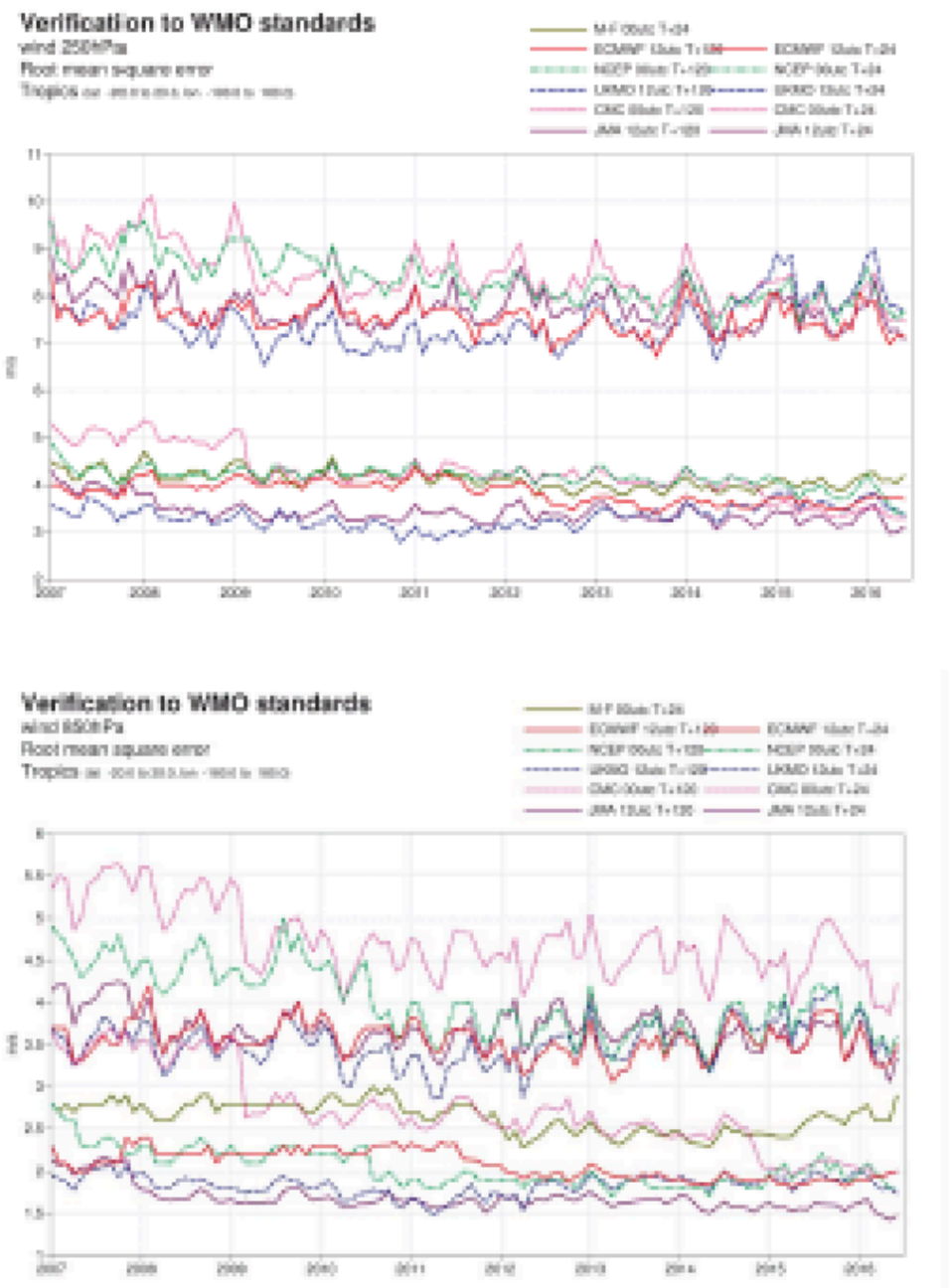


Figure 15: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropical 250 hPa (top) and 850 hPa (bottom). In each panel the upper curve shows the five-day forecast error and the lower curve shows the one-day forecast error. Each model is verified against its own analysis.

ECMWF Technical Memorandum 792

http://www.ecmwf.int/sites/default/files/elibrary/2016/16924-evaluation-ecmwf-forecasts-including-2016-resolution-upgrade.pdf

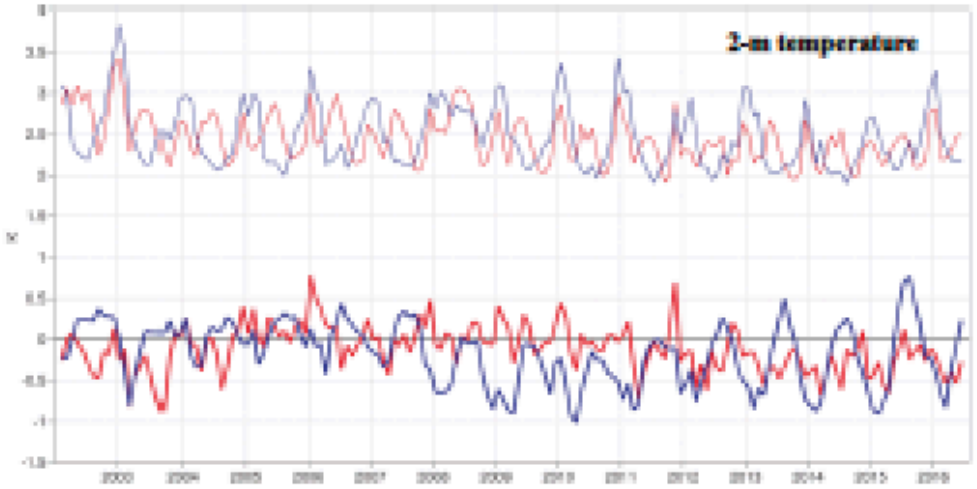


Figure 19: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

Figure 20: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

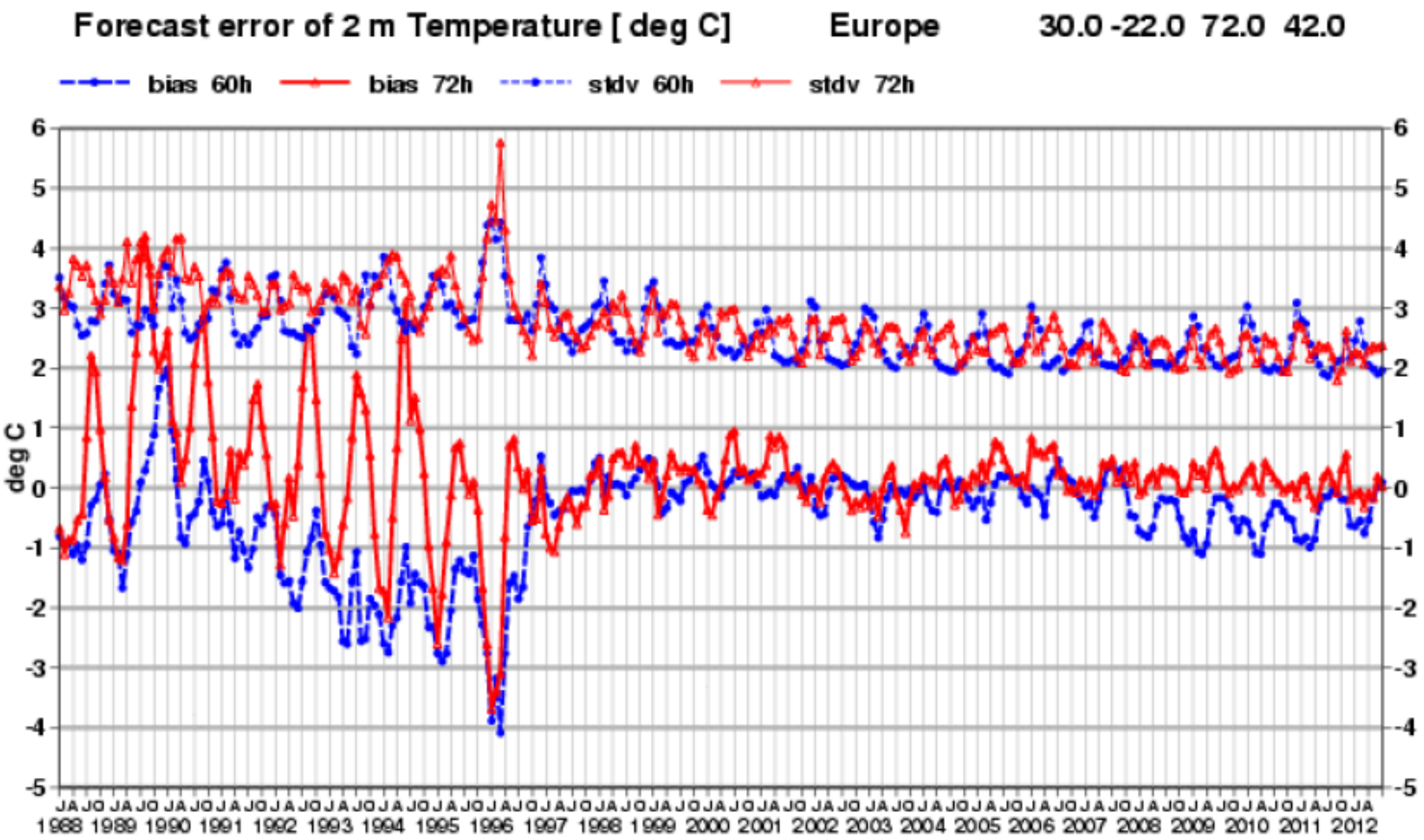


Figure 19: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm688.pdf

ECMWF Technical Memorandum 792

http://www.ecmwf.int/sites/default/files/elibrary/2016/16924-evaluation-ecmwf-forecasts-including-2016-resolution-upgrade.pdf

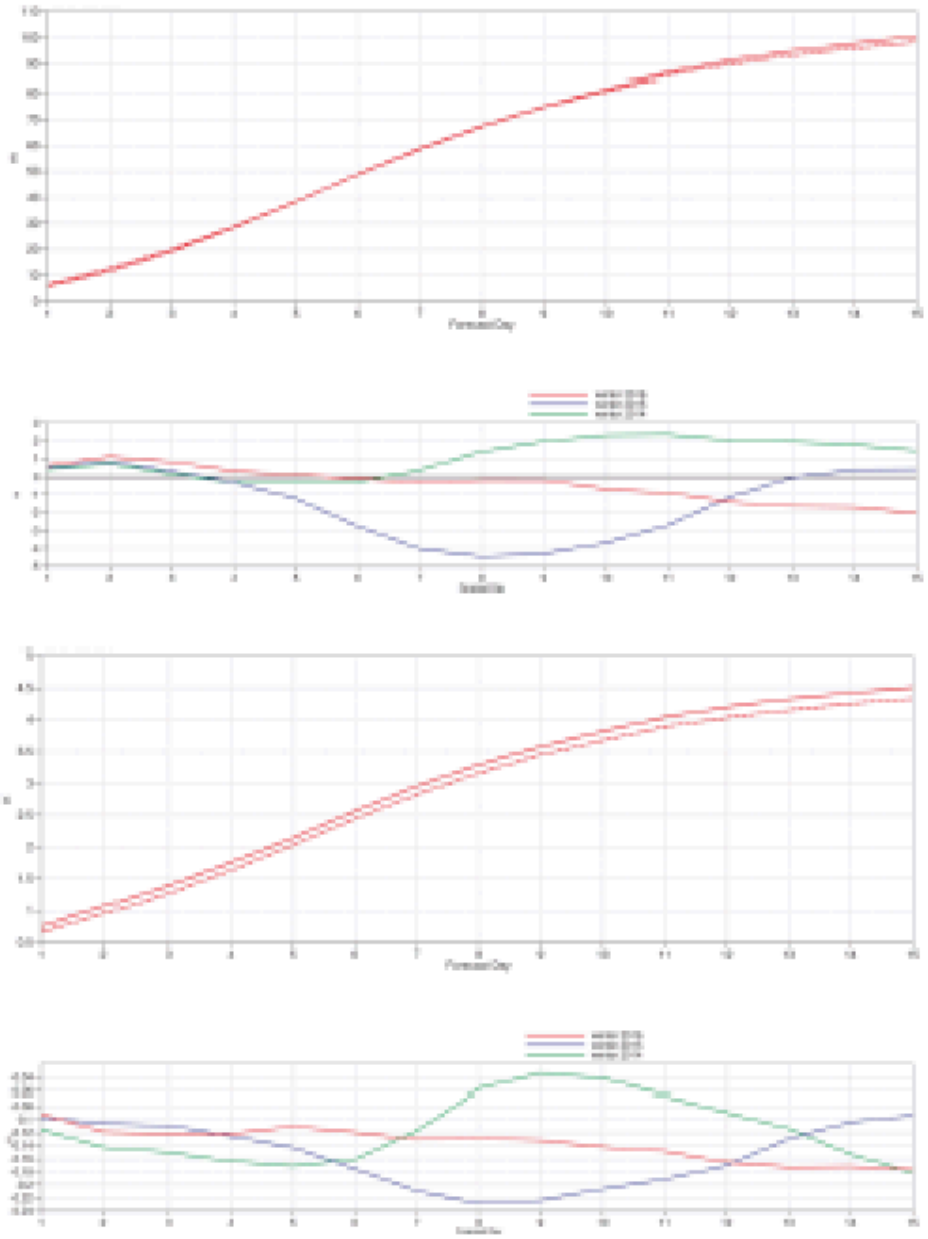


Figure 7: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2015–2016 (upper figure in each panel), and difference of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.

ECMWF

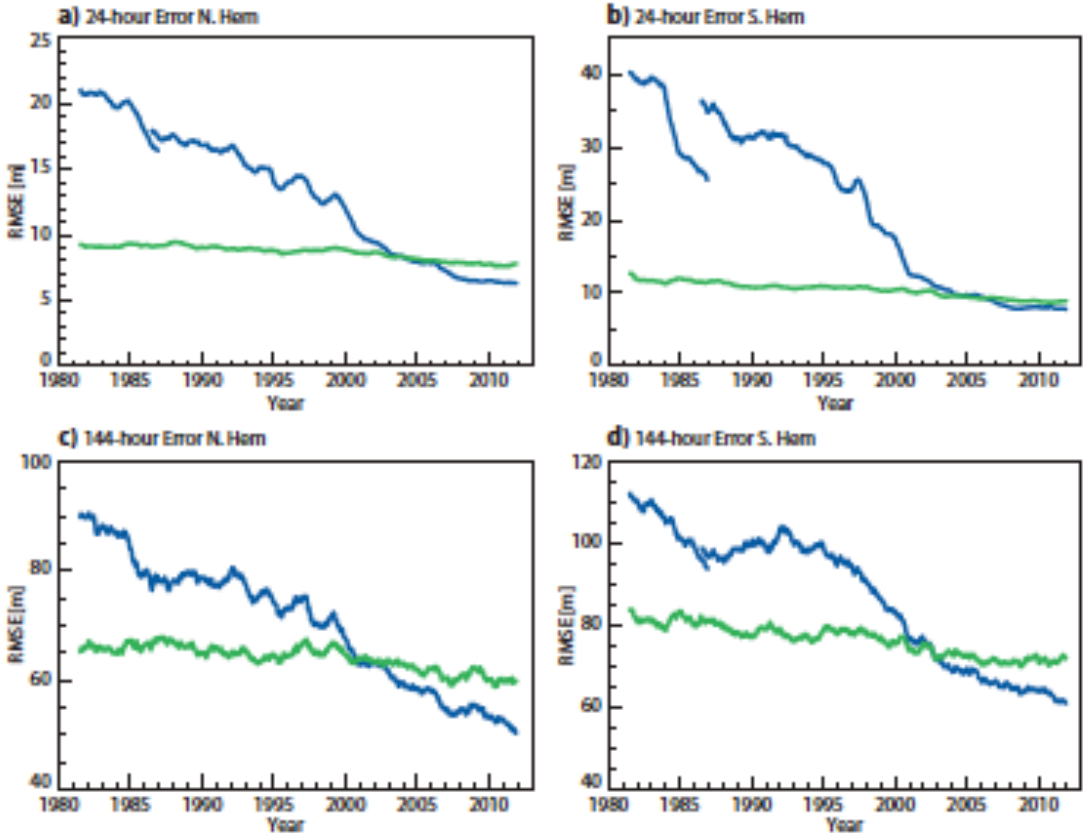


FIG. 3. Evolution of forecast errors from 1981 to 2012 for N.Hem (a and c) and S.Hem (b and d). Operational forecasts (blue) and ERA Interim (green). Note that before 1986 the operational analysis is used to verify the operational forecasts, after 1986 ERA Interim is used for the verification (with an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

Magnusson and Källén, *Mon. Wea. Rev.*, in press

Remaining Problems

Mostly in the 'physics' of models (Q and 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, …)

- …

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.

- The physical laws governing the evolution of the flow, available in practice in the form of a discretized, and necessarily approximate, numerical model.

- '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 ⇒ 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, $p \approx 4.10^7$ 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

Relative cost of the various components of the operational prediction suite at ECMWF (september 2015, J.-N. Thépaut) :

- 4DVAR: 9.5%
- Ensemble Data Assimilation (EDA) : 30%

EDA produces both the background error covariances for 4D-Var and the initial perturbations (in addition to Singular Vectors) for EPS.

- High resolution deterministic model : 4.5%
- Ensemble Prediction System (EPS) : 22%
- Ensemble hindcasts : 14%
- Others : 20% (among which 17% for computation of boundary conditions of a number of limited-area models ; those 17% include both assimilation and forecast)

Assimilation takes more than 40% of the computing power devoted to operational prediction

$z_1 = x + \zeta_1$ 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

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

$z_1 = x + \zeta_1$ 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

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

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

- $P(x = \xi \mid z_1, z_2) \propto p_1(z_1 - \xi)\, p_2(z_2 - \xi)$

$\propto \exp[- (\xi - x^a)^2/2p^a]$

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

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

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

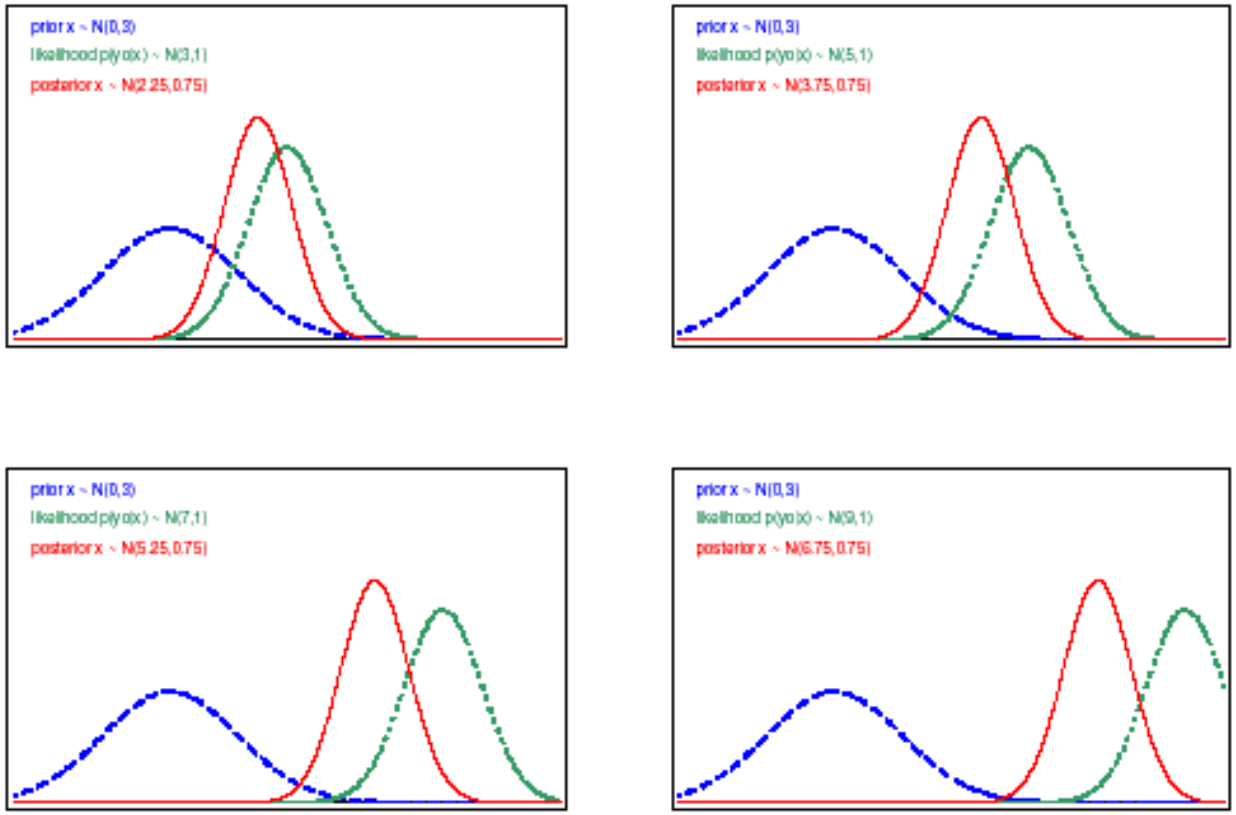


Fig. 1.1: Prior pdf $p(x)$ (dashed line), posterior pdf $p(x|y^o)$ (solid line), and Gaussian likelihood of observation $p(y^o|x)$ (dotted line), plotted against x for various values of y^o. (Adapted from Lorenc and Hammon 1988.)

$z_1 = x + \zeta_1$
$z_2 = x + \zeta_2$

Same as before, but ζ_1 and ζ_2 are now distributed according to exponential law with parameter a, *i. e.*

$$p(\zeta) \propto \exp[-|\zeta|/a] \quad ; \quad \mathrm{Var}(\zeta) = 2a^2$$

Conditional probability density function is now uniform over interval $[z_1, z_2]$, exponential with parameter $a/2$ outside that interval

$$E(x \mid z_1, z_2) = (z_1+z_2)/2$$

$\mathrm{Var}(x \mid z_1, z_2) = a^2\,(2\delta^3/3 + \delta^2 + \delta + 1/2) / (1 + 2\delta)$, with $\delta = |z_1 - z_2|/(2a)$
Increases from $a^2/2$ to ∞ as δ increases from 0 to ∞. Can be larger than variance $2a^2$ of original errors (probability 0.08)

Bayesian estimation

State vector x, belonging to *state space* $\mathcal{S}$ ($\dim \mathcal{S} = n$), to be estimated.

Data vector z, belonging to *data space* $\mathcal{D}$ ($\dim \mathcal{D} = m$), available.

$$z = F(x, \zeta) \qquad (1)$$

where ζ is a random element representing the uncertainty on the data (or, more precisely, on the link between the data and the unknown state vector).

For example

$$z = \Gamma x + \zeta$$

Bayesian estimation (continued)

Probability that $x = \xi$ for given ξ ?

$$x = \xi \quad \Rightarrow \quad z = F(\xi, \zeta)$$

$$P(x = \xi \mid z) = P[z = F(\xi, \zeta)] \,/\, \int_{\xi'} P[z = F(\xi', \zeta)]$$

Unambiguously defined iff, for any ζ, there is at most one x such that $z = F(x, \zeta)$.

$\Leftrightarrow$ data contain information, either directly or indirectly, on any component of x. *Determinacy* condition.

Bayesian estimation is however impossible in its general theoretical form in meteorological or oceanographical practice because

- It is impossible to explicitly describe a probability distribution in a space with dimension even as low as $n \approx 10^3$, not to speak of the dimension $n \approx 10^{6\text{-}9}$ of present Numerical Weather Prediction models ('*curse of dimensionality*').

- Probability distribution of errors on data very poorly known (model errors in particular).

One has to restrict oneself to a much more modest goal. Two approaches exist at present

- Obtain some 'central' estimate of the conditional probability distribution (expectation, mode, …), plus some estimate of the corresponding spread (standard deviations and a number of correlations).

- Produce an ensemble of estimates which are meant to sample the conditional probability distribution (dimension $N \approx O(10\text{-}100)$).

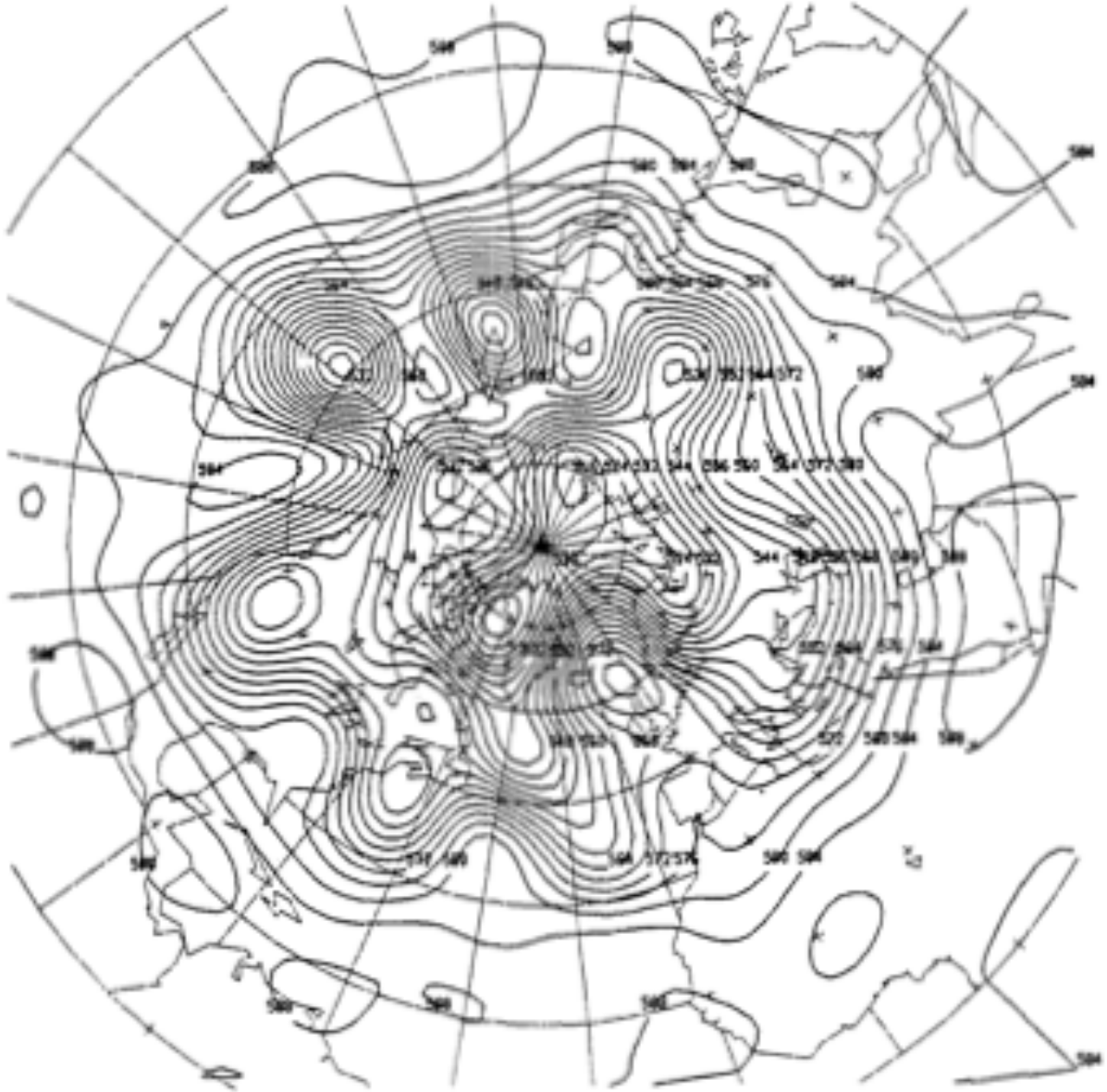


Figure 2. 500 mb height field produced by the operational analysis procedure of Direction de la Météorologie for 00 GMT, 26 April 1984. Units: dam, contour interval: 4 dam. The field has been truncated to the truncation of the model used for the experiments described in the article.

Courtier and Talagrand, *QJRMS*, 1987

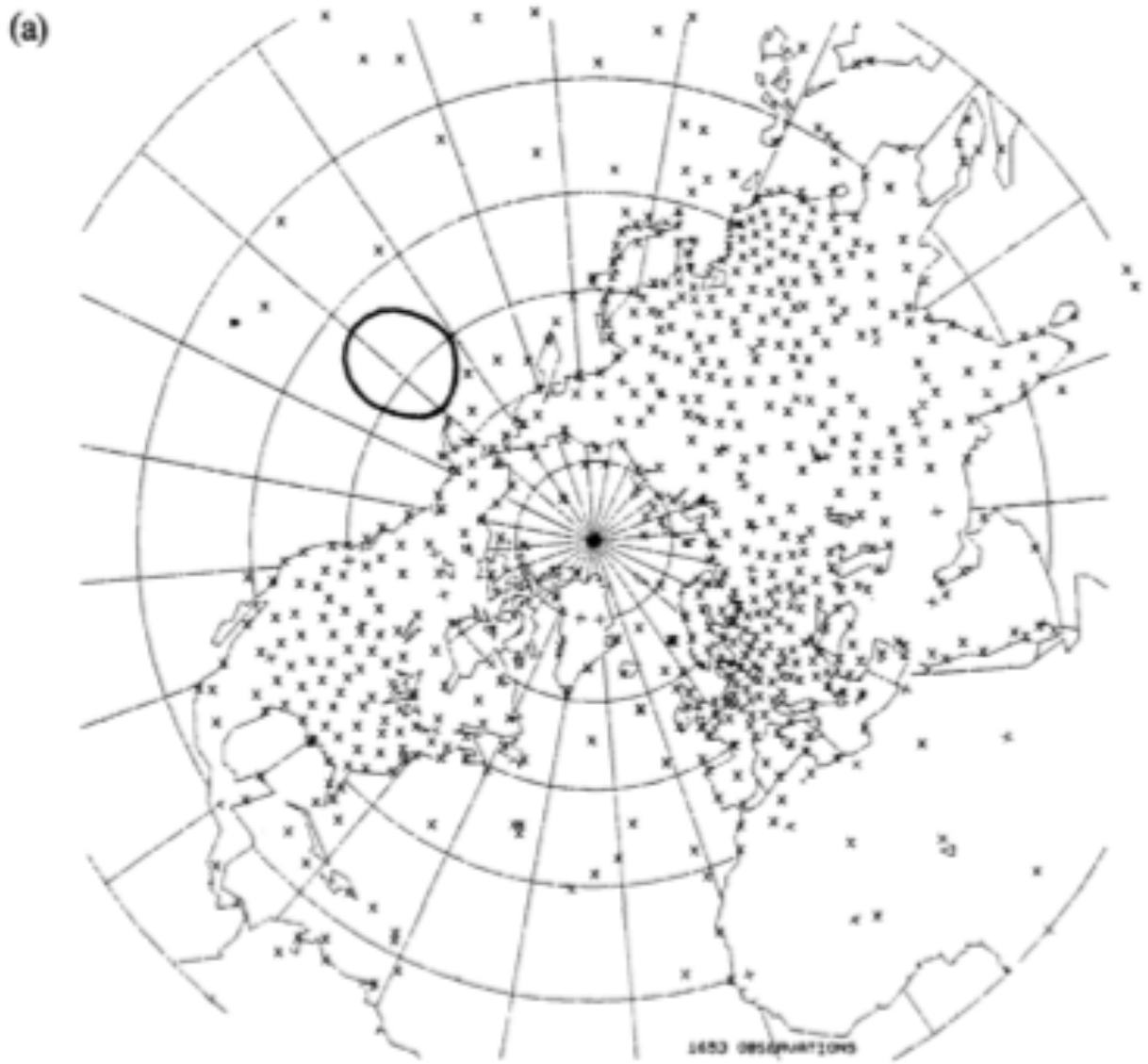


Figure 1. Geographical distribution of the observations used for the assimilation experiments. (a): geopotential observations; (b): wind observations. At most of the points plotted, several observations were made at successive synoptic hours. On each of the two charts, the heavy line delineates the Aleutian depression (see Figure 2).

Courtier and Talagrand, *QJRMS*, 1987

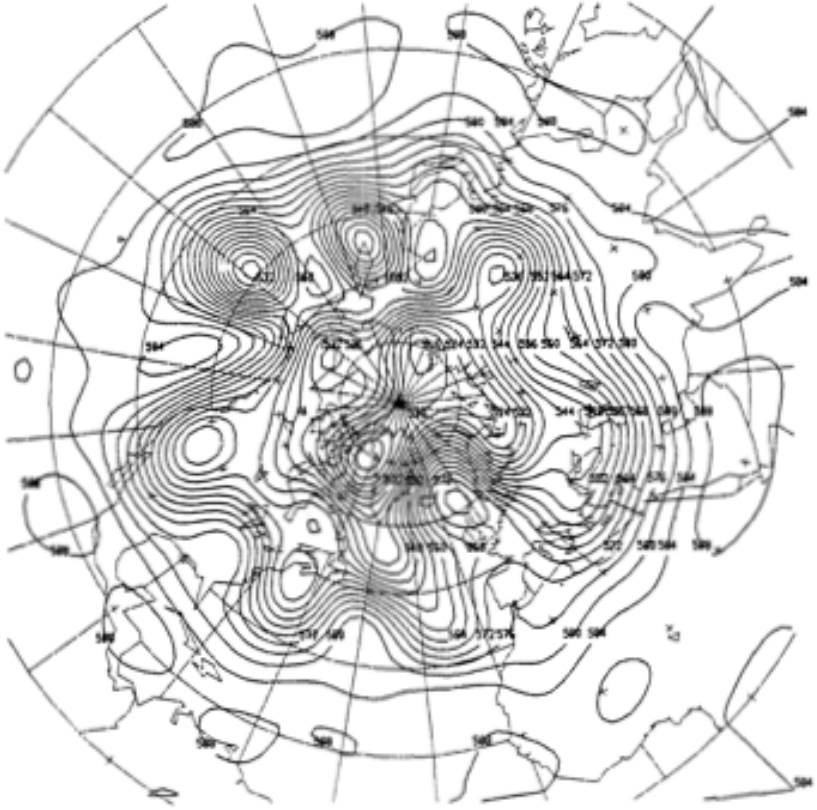


Figure 2. 500 mb height field produced by the operational analysis procedure of Direction de la Météorologie for 00 GMT, 26 April 1984. Units: dam, contour interval: 4 dam. The field has been truncated to the truncation of the model used for the experiments described in the article.

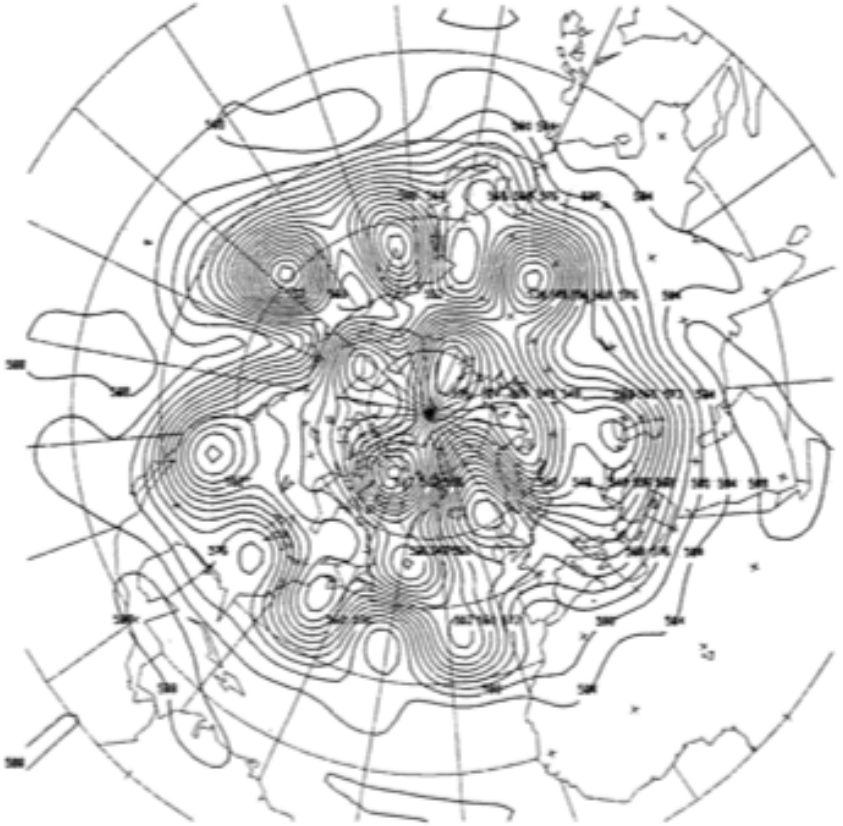


Figure 3. 500 mb height field produced for 00 GMT, 26 April 1984, by the variational analysis minimizing the distance function defined by Eqs. (1)–(2) over a 24-hour period. Units: dam; contour interval: 4 dam.

500-hPa geopotential field as determined by : (left) operational assimilation system of French Weather Service (3D, primitive equation) and (right) experimental variational system (2D, vorticity equation)

Courtier and Talagrand, *QJRMS*, 1987

Random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T = (x_i)$ (*e. g.* pressure, temperature, abundance of given chemical compound at n grid-points of a numerical model)

- Expectation $E(\boldsymbol{x}) \equiv [E(x_i)]$; centred vector $\boldsymbol{x}' \equiv \boldsymbol{x} - E(\boldsymbol{x})$

- Covariance matrix

$$E(\boldsymbol{x}'\boldsymbol{x}'^T) = [E(x_i'x_j')]$$

dimension nxn, symmetric non-negative (strictly definite positive except if linear relationship holds between the x_i's with probability 1).

- Two random vectors

$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$

$\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^T$

$$E(\boldsymbol{x}'\boldsymbol{y}'^T) = E(x_i'y_j')$$

dimension nxp

Covariance matrices will be denoted

$$C_{xx} \equiv E(\boldsymbol{x}'\boldsymbol{x}'^{\mathrm{T}})$$

$$C_{xy} \equiv E(\boldsymbol{x}'\boldsymbol{y}'^{\mathrm{T}})$$

Random function $\Phi(\xi)$ (field of pressure, temperature, abundance of given chemical compound, … ; ξ is now spatial and/or temporal coordinate)

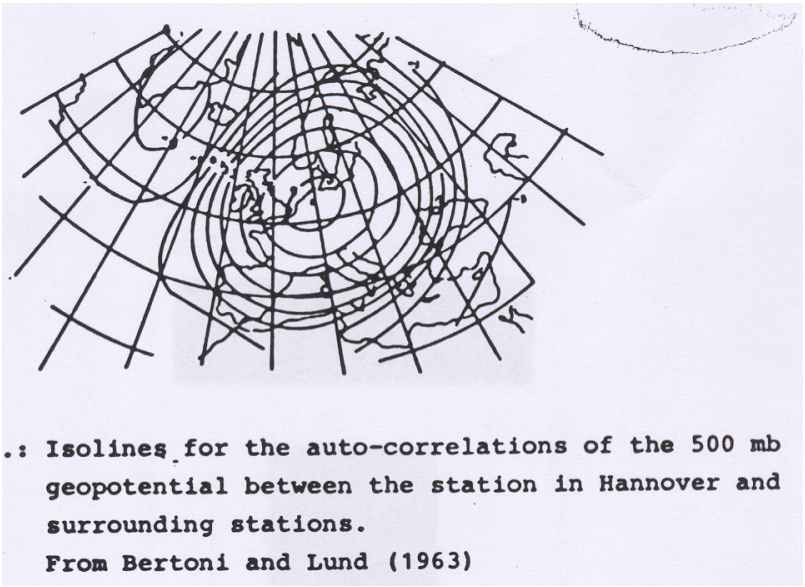
- Expectation $E[\Phi(\xi)]$; $\Phi'(\xi) \equiv \Phi(\xi) - E[\Phi(\xi)]$
- Variance $Var[\Phi(\xi)] = E\{[\Phi'(\xi)]^2\}$

- Covariance function

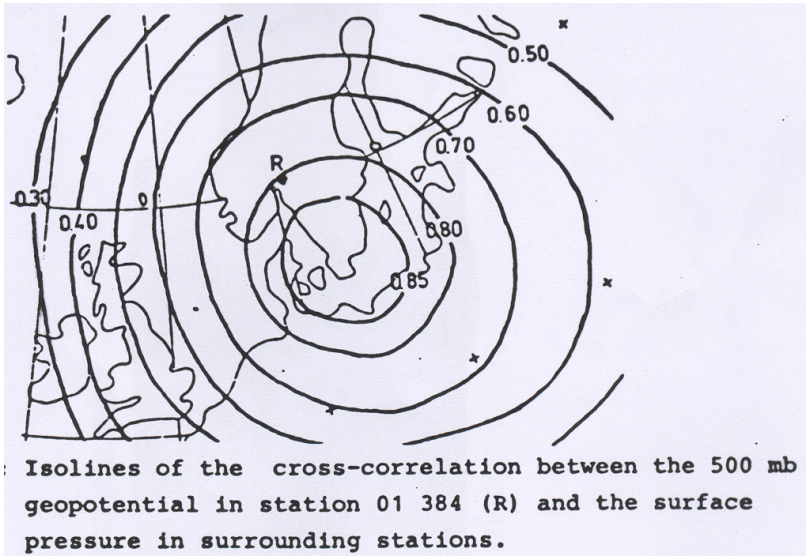
$$(\xi_1, \xi_2) \rightarrow C_\Phi(\xi_1, \xi_2) \equiv E[\Phi'(\xi_1)\, \Phi'(\xi_2)]$$

- Correlation function

$$Cor_\varphi(\xi_1, \xi_2) \equiv E[\Phi'(\xi_1)\, \Phi'(\xi_2)] \,/\, \{Var[\Phi(\xi_1)]\, Var[\Phi(\xi_2)]\}^{1/2}$$



.: Isolines for the auto-correlations of the 500 mb geopotential between the station in Hannover and surrounding stations.
From Bertoni and Lund (1963)



Isolines of the cross-correlation between the 500 mb geopotential in station 01 384 (R) and the surface pressure in surrounding stations.

After N. Gustafsson

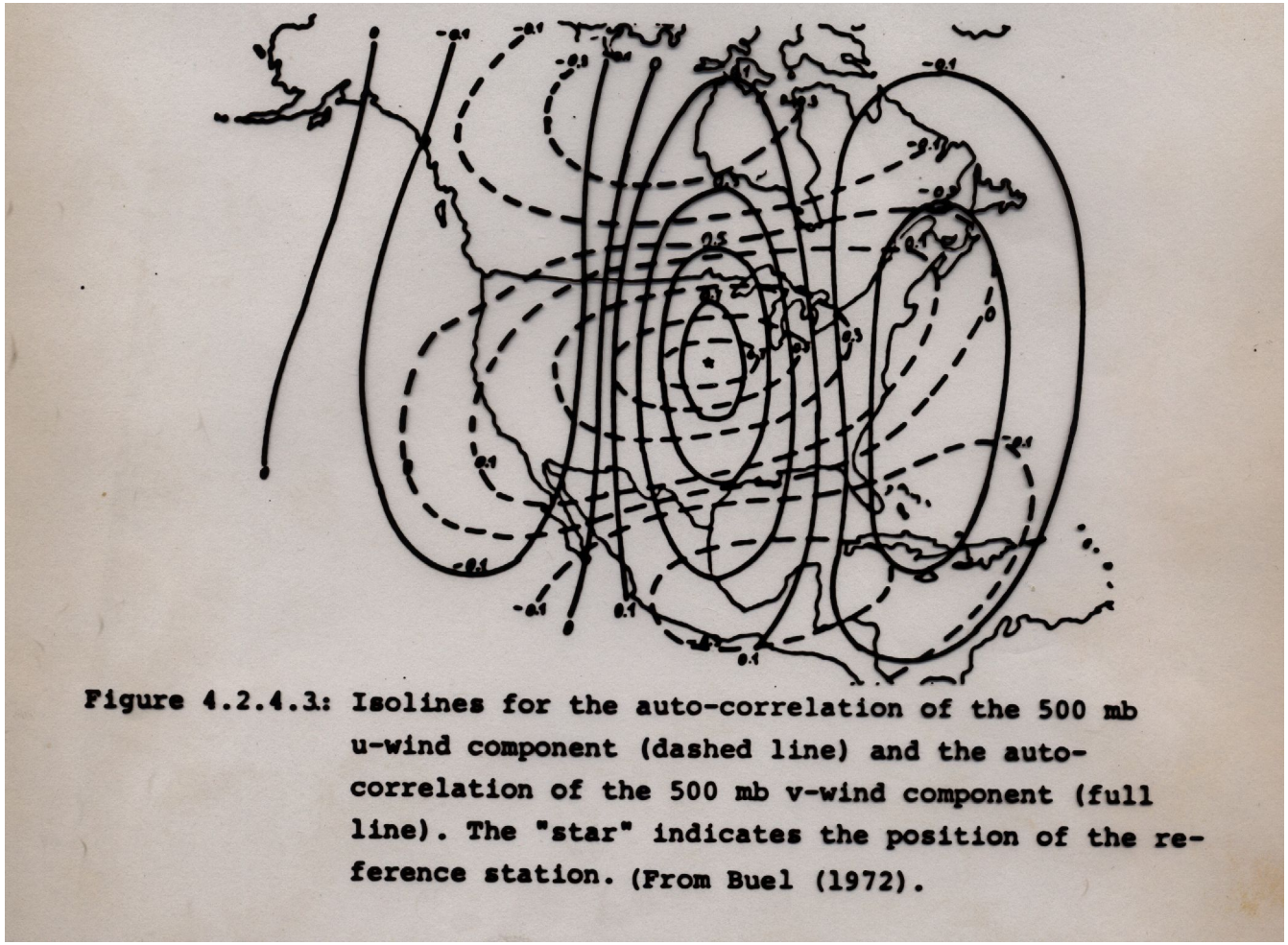


Figure 4.2.4.3: Isolines for the auto-correlation of the 500 mb u-wind component (dashed line) and the auto-correlation of the 500 mb v-wind component (full line). The "star" indicates the position of the reference station. (From Buel (1972).

After N. Gustafsson

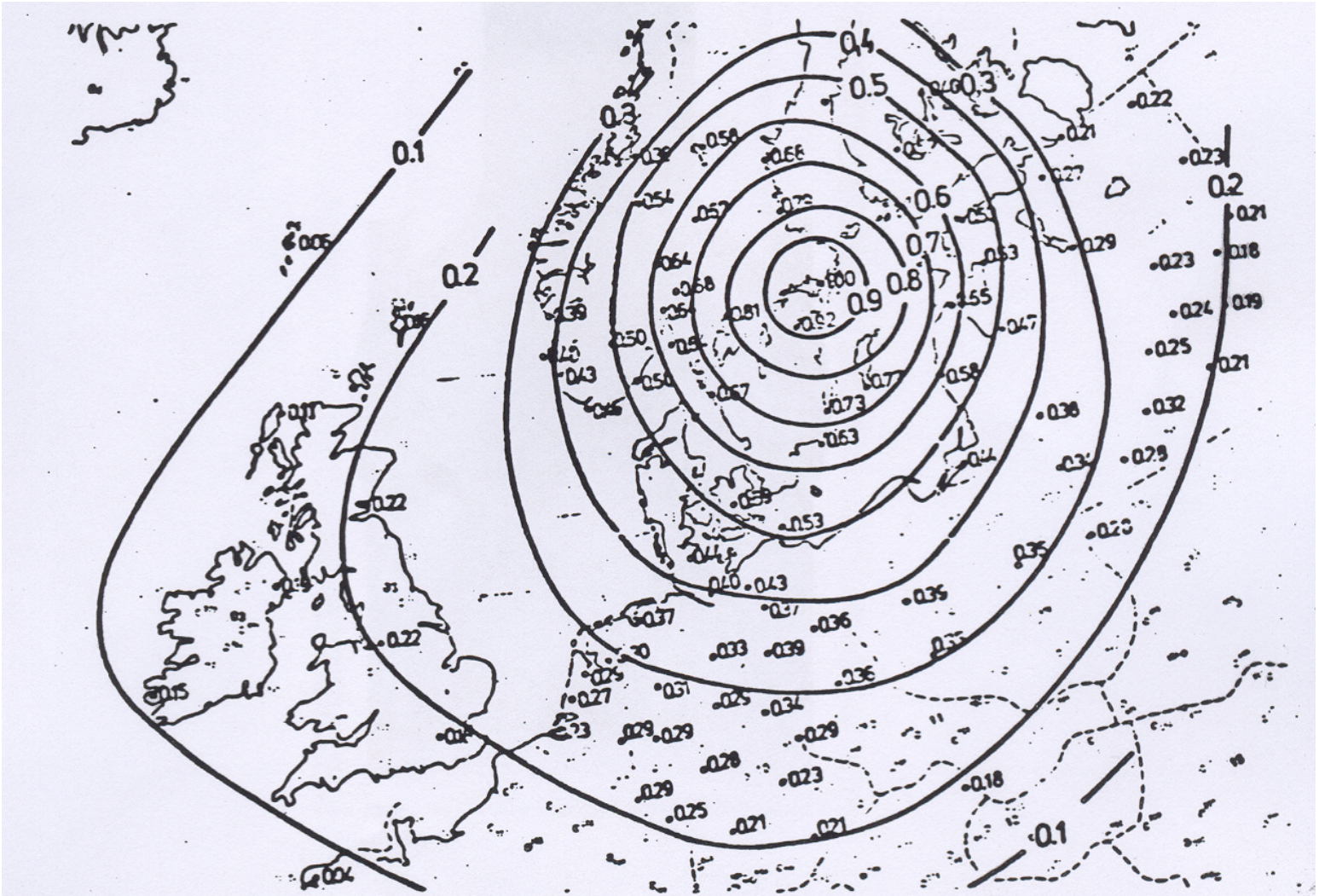


Figure 5.1.1.4.1 Auto-correlation of errors in 12h numerical forecasts of surface pressure in a reference station (Stockholm) and other stations.

After N. Gustafsson

Optimal Interpolation

Random field $\Phi(\boldsymbol{\xi})$

Observation network $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_p$

For one particular realization of the field, observations

$y_j = \Phi(\boldsymbol{\xi}_j) + \varepsilon_j \quad , \; j = 1, \ldots, p \qquad ,$ making up vector $\boldsymbol{y} = (y_j)$

Estimate $x = \Phi(\boldsymbol{\xi})$ at given point $\boldsymbol{\xi}$, in the form

$$x^a = \alpha + \Sigma_j \, \beta_j \, y_j \; = \alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{y} \qquad , \qquad \text{where } \boldsymbol{\beta} = (\beta_j)$$

α and the β_j's being determined so as to minimize the expected quadratic estimation error $E[(x\text{-}x^a)^2]$

Optimal Interpolation (continued 1)

Solution

$$x^a = E(x) + E(x'\boldsymbol{y}'^{\mathrm{T}})\,[E(\boldsymbol{y}'\boldsymbol{y}'^{\mathrm{T}})]^{-1}\,[\boldsymbol{y} - E(\boldsymbol{y})]$$
$$= E(x) + \boldsymbol{C}_{xy}\,[\boldsymbol{C}_{yy}]^{-1}\,[\boldsymbol{y} - E(\boldsymbol{y})]$$

i. e., $\boldsymbol{\beta}^{\mathrm{T}} = C_{xy}\,[C_{yy}]^{-1}$

$\alpha = E(x) - \boldsymbol{\beta}^{\mathrm{T}} E(\boldsymbol{y})$

Estimate is unbiased $E(x\text{-}x^a) = 0$

Minimized quadratic estimation error

$$E[(x\text{-}x^a)^2] = E(x'^2) - E[(x'^a)^2])$$
$$= \boldsymbol{C}_{xx} - \boldsymbol{C}_{xy}\,[\boldsymbol{C}_{yy}]^{-1}\,\boldsymbol{C}_{yx}$$

Estimation made in terms of deviations x' and $\boldsymbol{y}'$ from expectations $E(x)$ and $E(\boldsymbol{y})$.

Optimal Interpolation (continued 2)

$$x^a = E(x) + E(x'\mathbf{y}'^{\mathrm{T}})\,[E(\mathbf{y}'\mathbf{y}'^{\mathrm{T}})]^{-1}\,[\mathbf{y} - E(\mathbf{y})]$$

$$y_j = \Phi(\xi_j) + \varepsilon_j$$

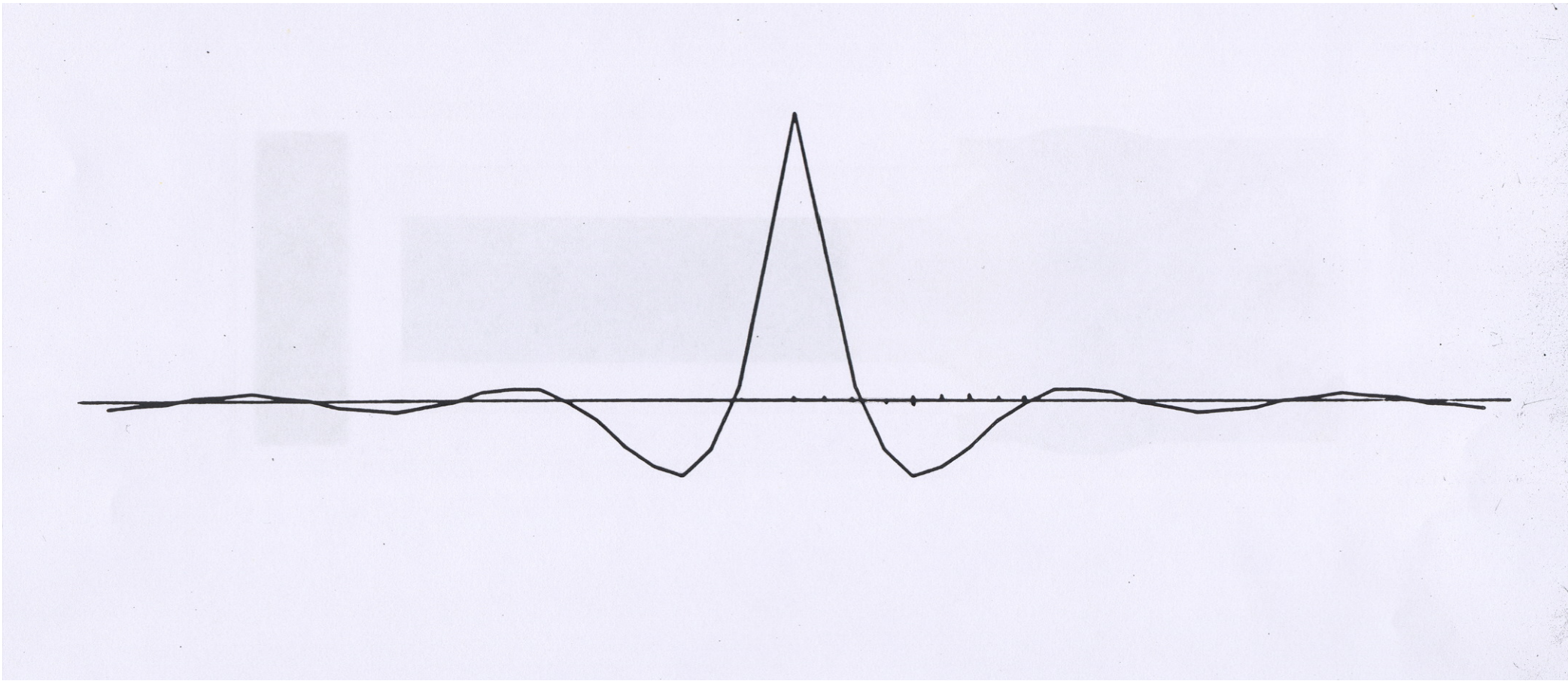
$$E(y_j' y_k') = E[\Phi'(\xi_j) + \varepsilon_j'][\Phi'(\xi_k) + \varepsilon_k']$$

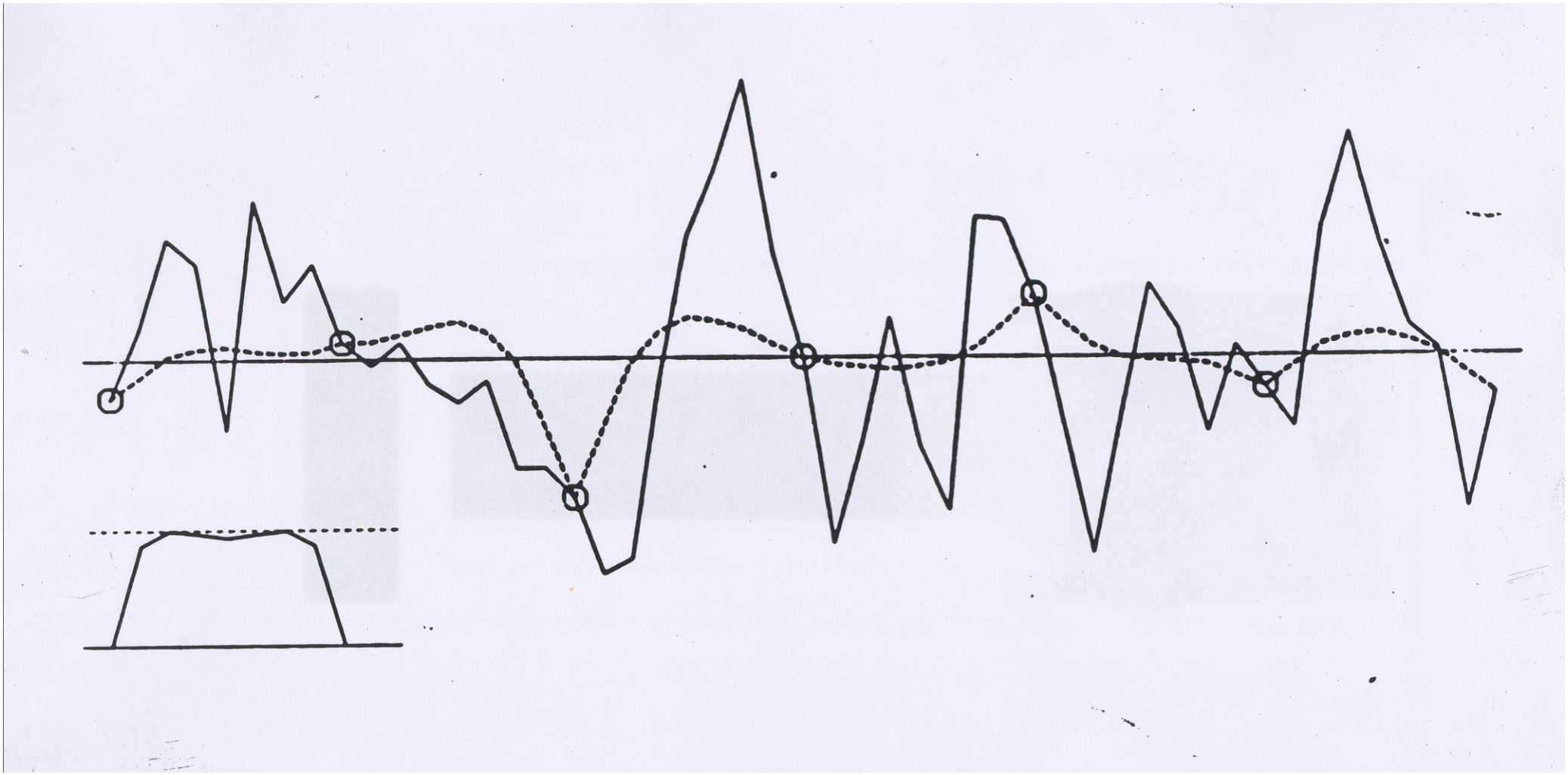
If observation errors ε_j are mutually uncorrelated, have common variance r, and are uncorrelated with field Φ, then

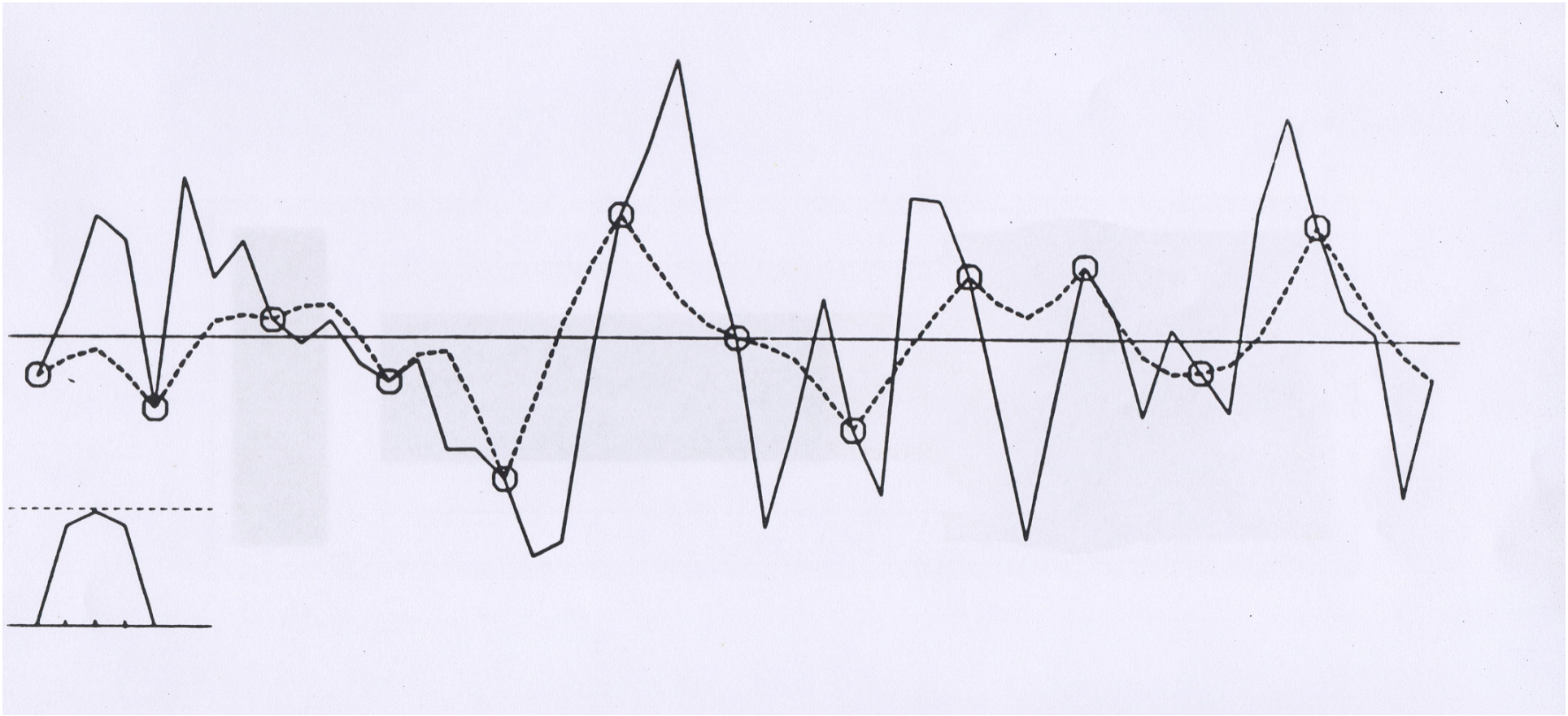
$$E(y_j' y_k') = C_\Phi(\xi_j, \xi_k) + r\delta_{jk}$$

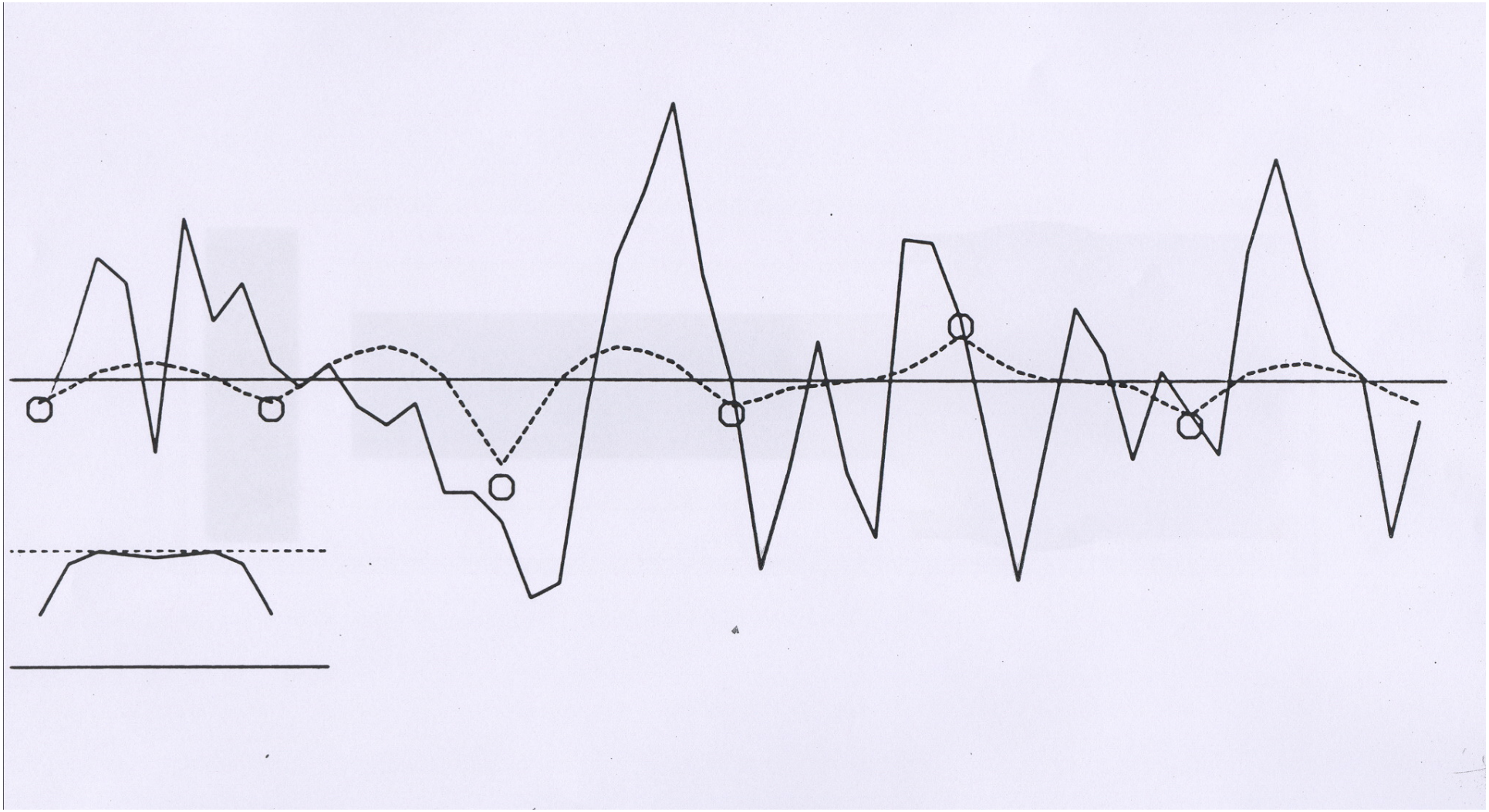
and

$$E(x' y_j') = C_\Phi(\xi, \xi_j)$$









Optimal Interpolation (continued 3)

$$x^a = E(x) + \boldsymbol{C}_{xy} [\boldsymbol{C}_{yy}]^{-1} [\boldsymbol{y} - E(\boldsymbol{y})]$$

Vector

$$\boldsymbol{\mu} = (\mu_j) \equiv [\boldsymbol{C}_{yy}]^{-1} [\boldsymbol{y} - E(\boldsymbol{y})]$$

is independent of variable to be estimated

$$x^a = E(x) + \Sigma_j \mu_j E(x' y_j')$$

$$\begin{aligned} \Phi^a(\xi) &= E[\Phi(\xi)] + \Sigma_j \mu_j E[\Phi'(\xi) y_j'] \\ &= E[\Phi(\xi)] + \Sigma_j \mu_j C_\Phi(\xi, \xi_j) \end{aligned}$$

Correction made on background expectation is a linear combination of the p functions $C_\Phi(\xi, \xi_j)$

$C_\Phi(\xi, \xi_j)$, considered as a function of estimation position ξ, is the *representer* associated with observation y_j.

Optimal Interpolation (continued 4)

$$x^a = E(x) + \boldsymbol{C}_{xy} [\boldsymbol{C}_{yy}]^{-1} [\boldsymbol{y} - E(\boldsymbol{y})]$$

$$p^a \equiv E[(x-x^a)^2] = \boldsymbol{C}_{xx} - \boldsymbol{C}_{xy} [\boldsymbol{C}_{yy}]^{-1} \boldsymbol{C}_{yx}$$

If random variables x and $\boldsymbol{y}$ are *globally* gaussian, Optimal Interpolation achieves bayesian estimation, in the sense that $P(x \mid \boldsymbol{y}) = \mathcal{N}[x^a, p^a]$.

Optimal Interpolation (continued 5)

Univariate interpolation. Each physical field (*e*. *g*. temperature) determined from observations of that field only.

Multivariate interpolation. Observations of different physical fields are used simultaneously. Requires specification of cross-covariances between various fields.

Cross-covariances between mass and velocity fields can simply be modelled on the basis of geostrophic balance.

Cross-covariances between humidity and temperature (and other) fields still a problem.

Schlatter's (1975) multivariate covariances

Specified as multivariate 2-point functions.

Not easy to ensure that specified functions are actually valid covariances.

Used in OI and related observation-space methods.

Courtesy A. Lorenc

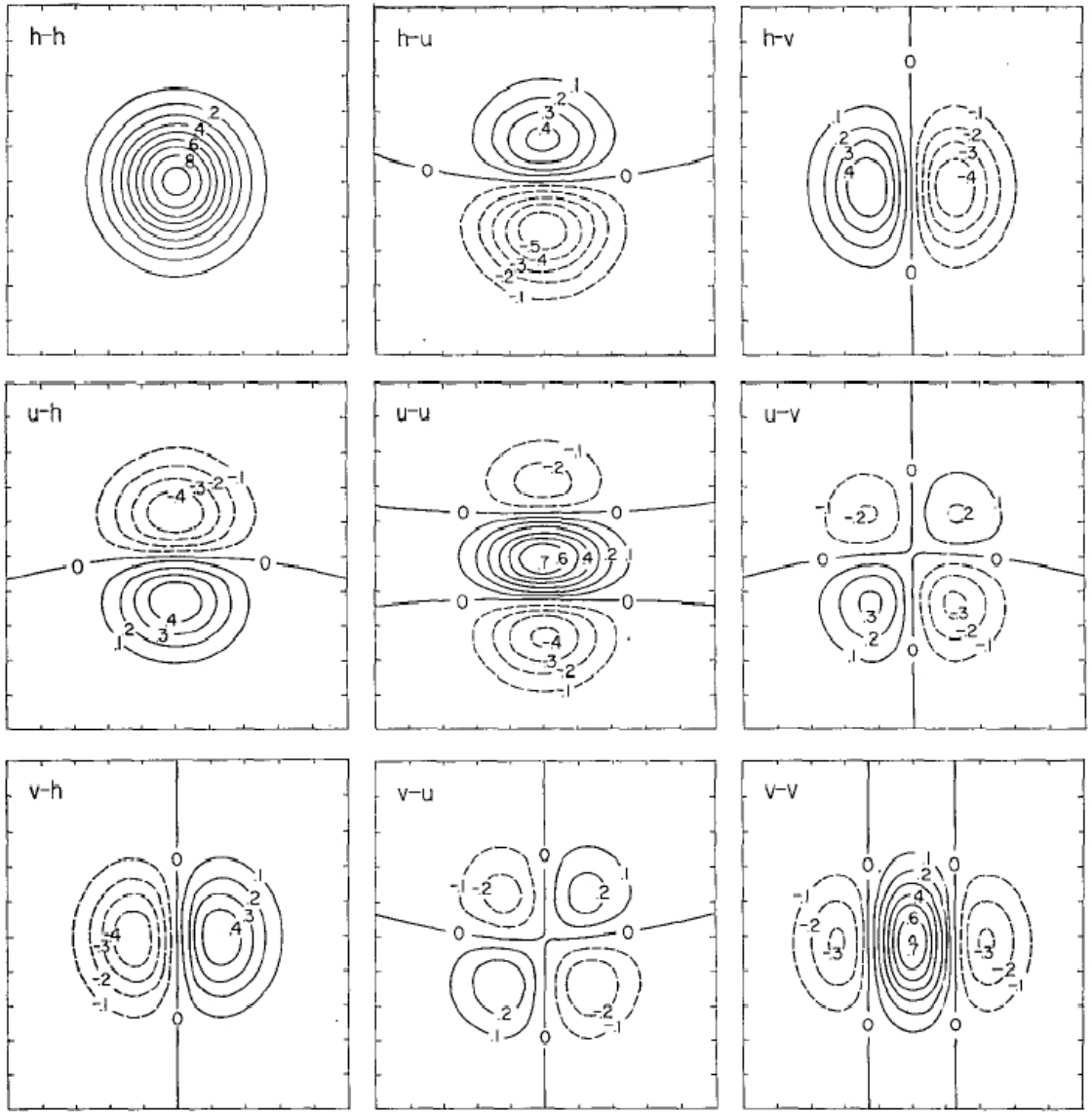


FIG. 3. Correlations among the variables h, u, and v based upon the expression $\mu = 0.95 \exp(-1.24s^2)$ for height-height correlation and the geostrophic relations. Diagrams centered at 110°W, 35°N. Tick marks 500 km apart.