Assimilation of observations. The case of meteorology and oceanography. Inverse problems. Bayesian estimation.

Olivier Talagrand

School *Data Assimilation*

Nordic Institute for Theoretical Physics (NORDITA)

Stockholm, Sweden

26 April 2011

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:

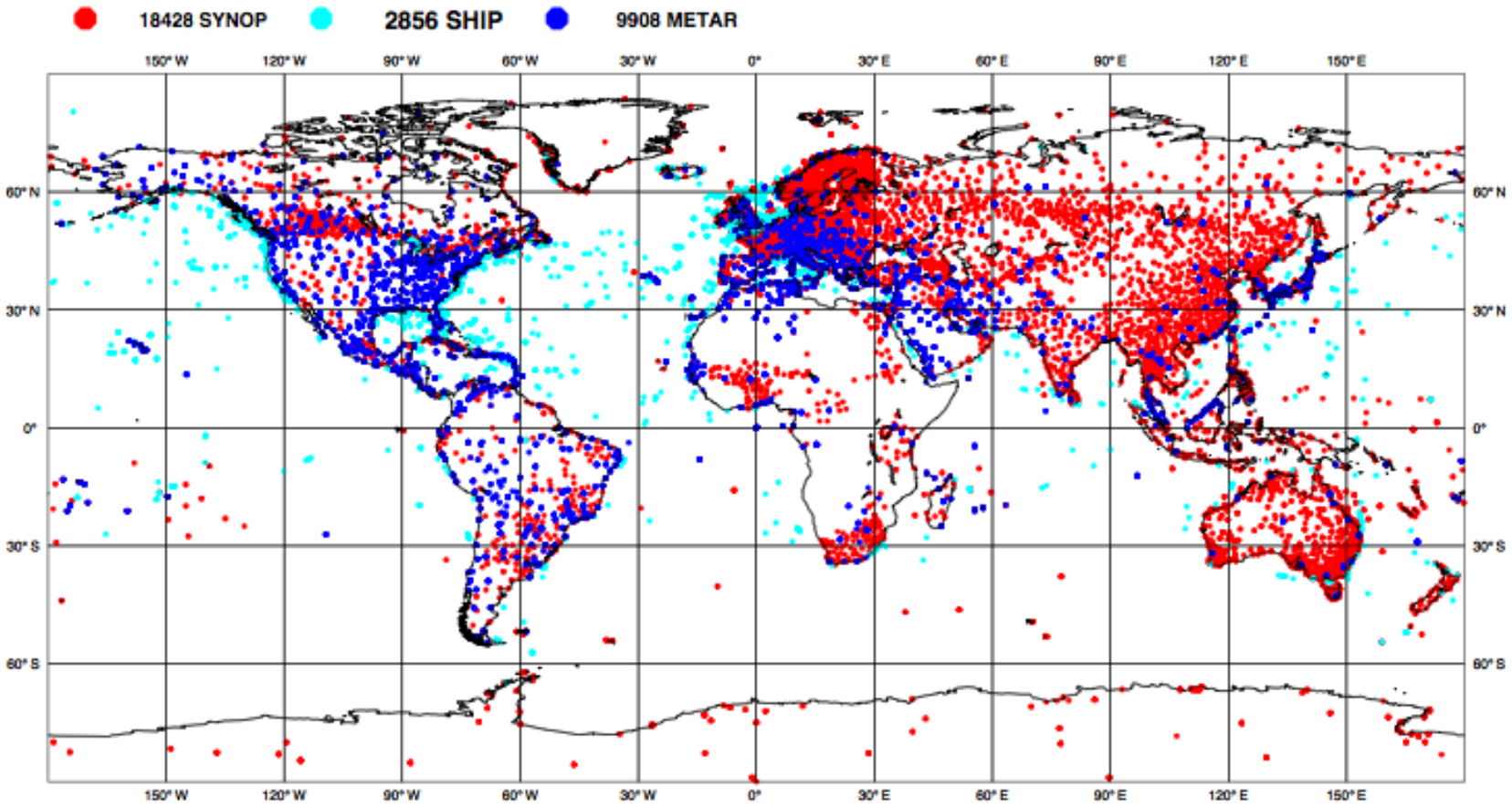
1st row: 1st panel: MSLP analysis for 12 UTC of 29 Aug
2nd panel: MSLP t+84h T_L511L60 forecast started at 00 UTC of 26 Aug
3rd panel: MSLP t+84h EPS-control T_L255L40 forecast started at 00 UTC of 26 Aug
Other rows: 50 EPS-perturbed T_L255L40 forecast started at 00 UTC of 26 Aug.

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

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.*

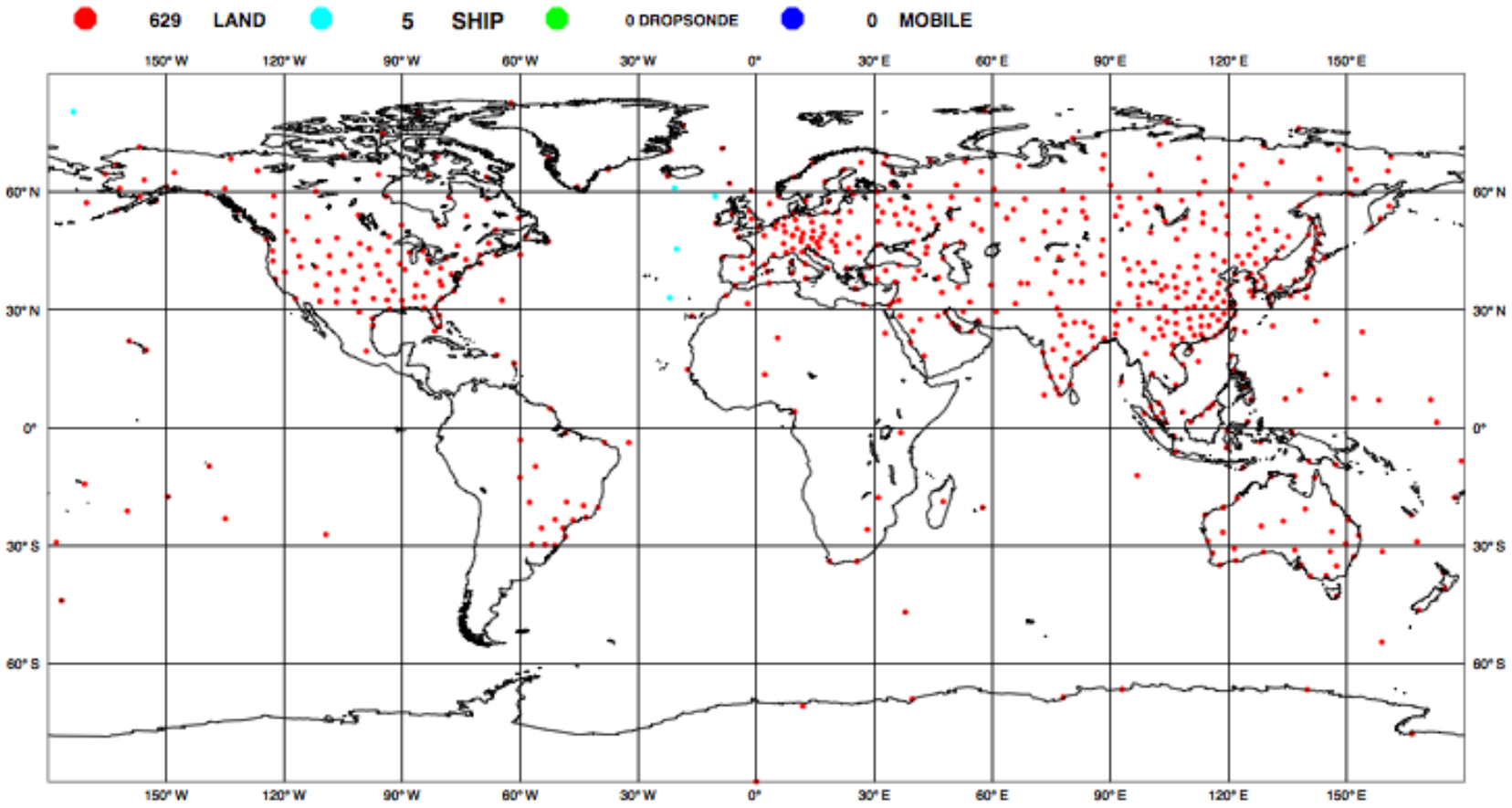
H. Poincaré, *Science et Méthode*, Paris, 1908
(translated Dover Publ., 1952)

ECMWF Data Coverage (All obs DA) - SYNOP/SHIP
23/APR/2011; 00 UTC
Total number of obs = 31192



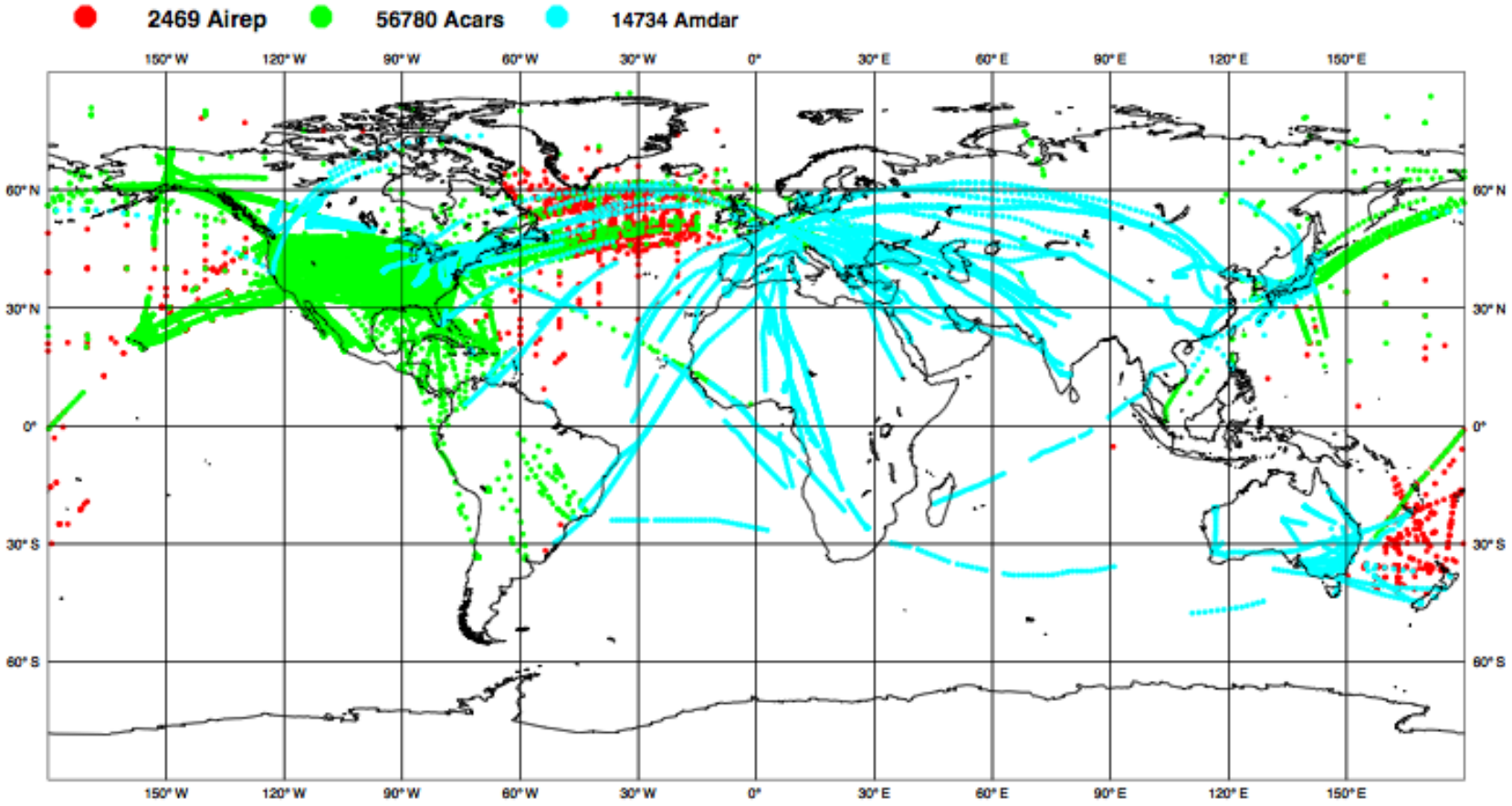
ECMWF

ECMWF Data Coverage (All obs DA) - TEMP
23/APR/2011; 00 UTC
Total number of obs = 634



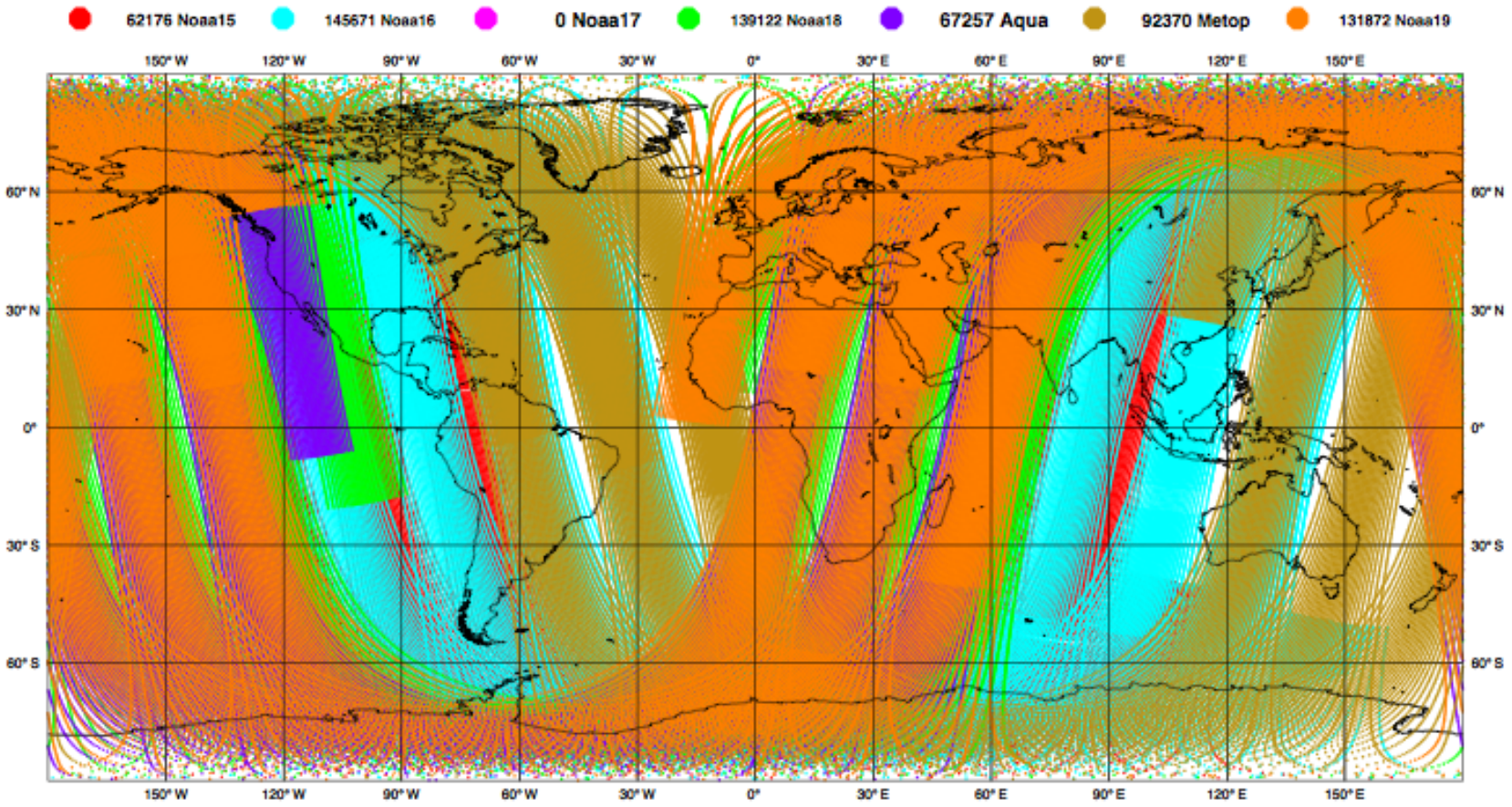
ECMWF

ECMWF Data Coverage (All obs DA) - AIRCRAFT
23/APR/2011; 00 UTC
Total number of obs = 73983

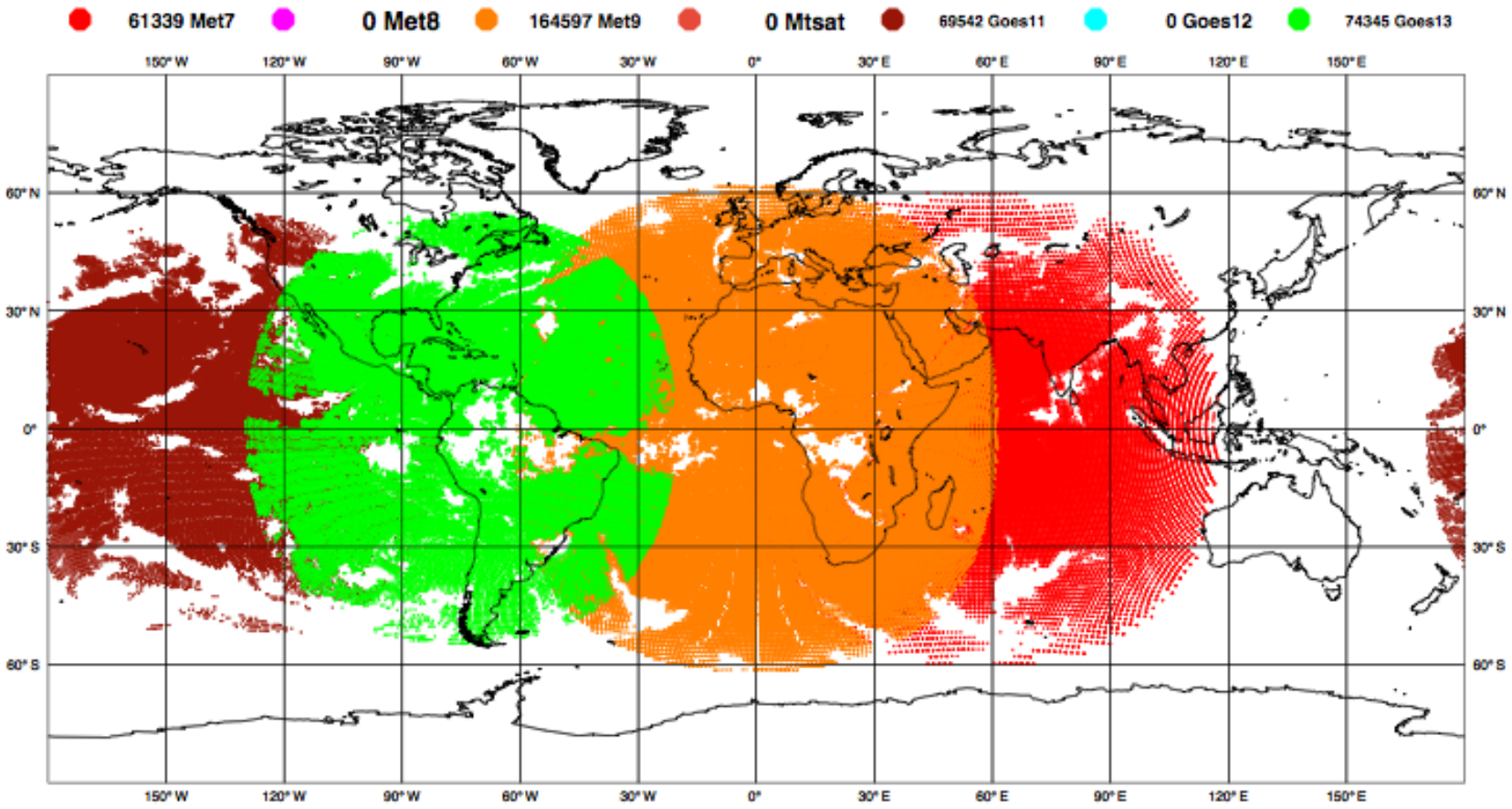


ECMWF

ECMWF Data Coverage (All obs DA) - AMSU-A
23/APR/2011; 00 UTC
Total number of obs = 638468

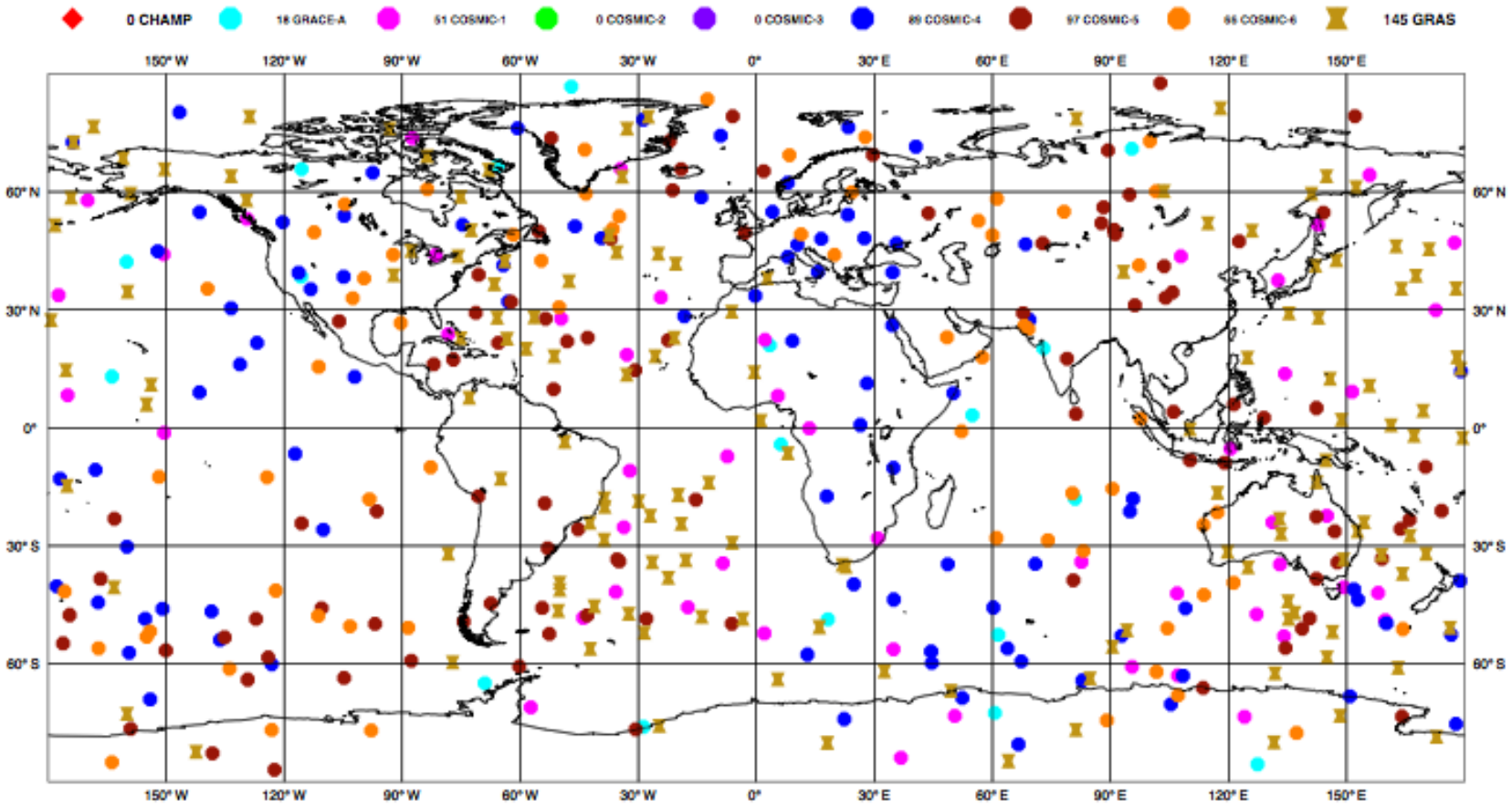


ECMWF Data Coverage (All obs DA) - GRAD
23/APR/2011; 00 UTC
Total number of obs = 369823



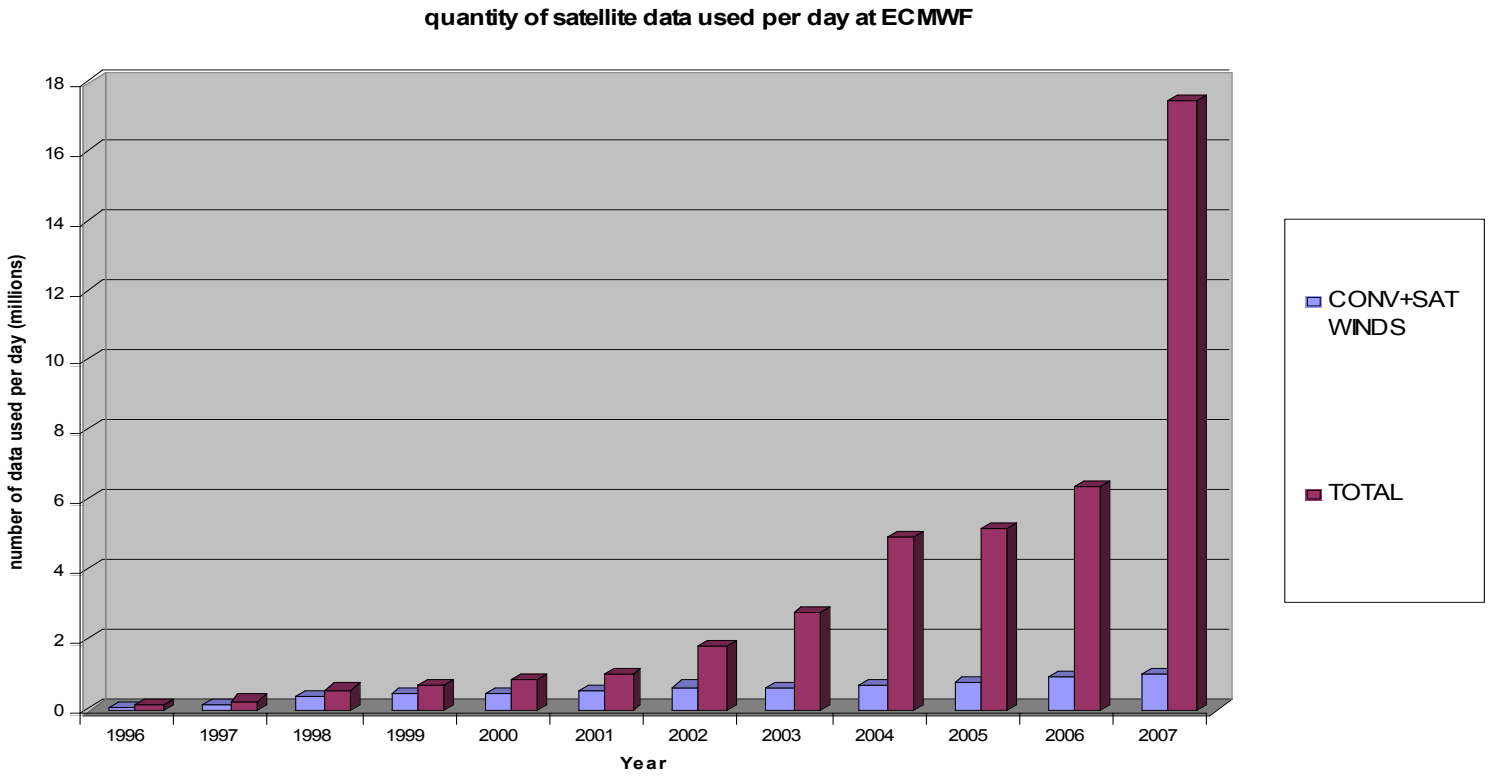
ECMWF

ECMWF Data Coverage (All obs DA) - GPSRO
23/APR/2011; 00 UTC
Total number of obs = 466

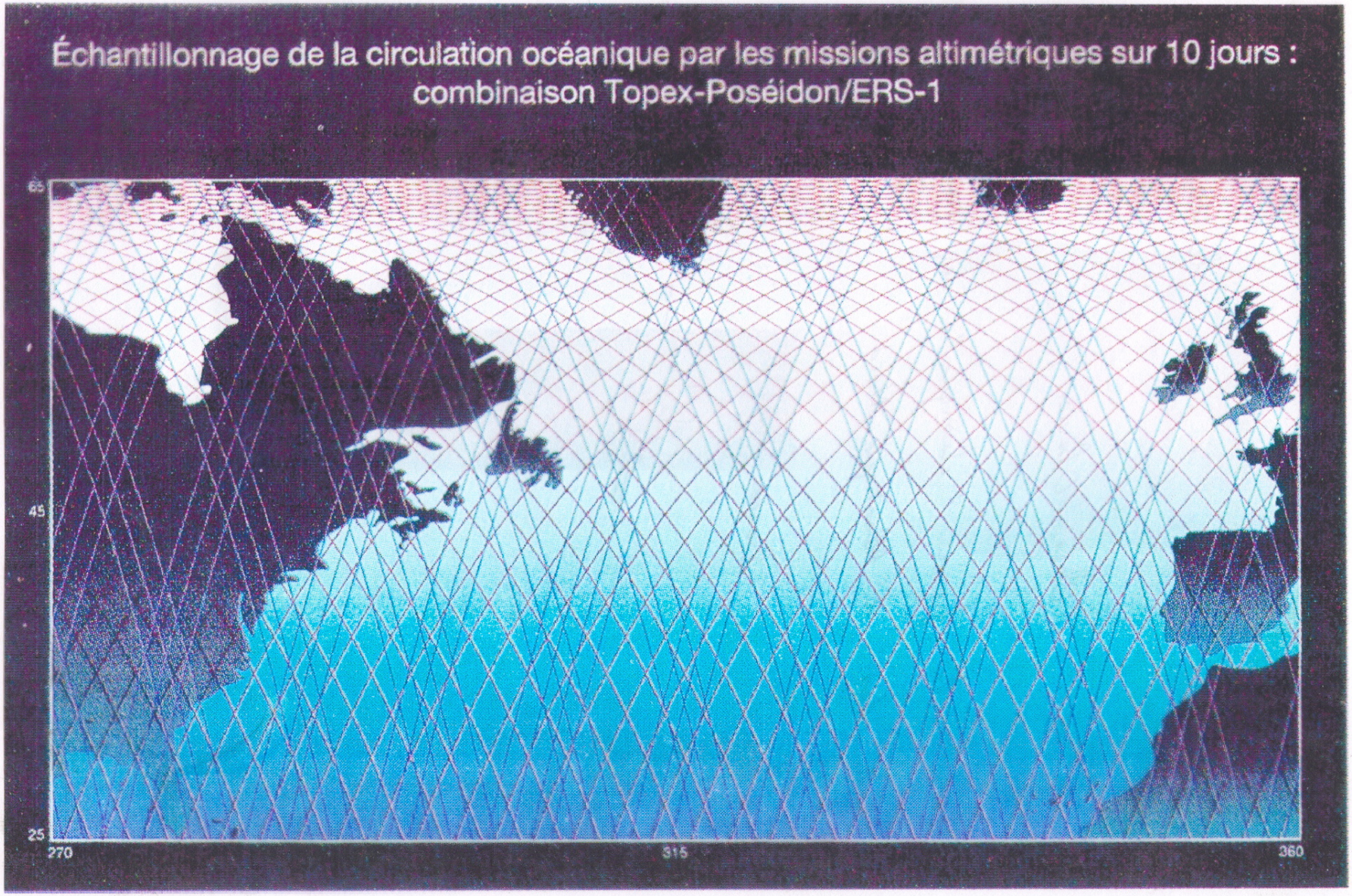


ECMWF

December 2007: Satellite data volumes used: around 18 millions per day



Value as of March 2010 : 25 millions per day



S. Louvel, Doctoral Dissertation, 1999

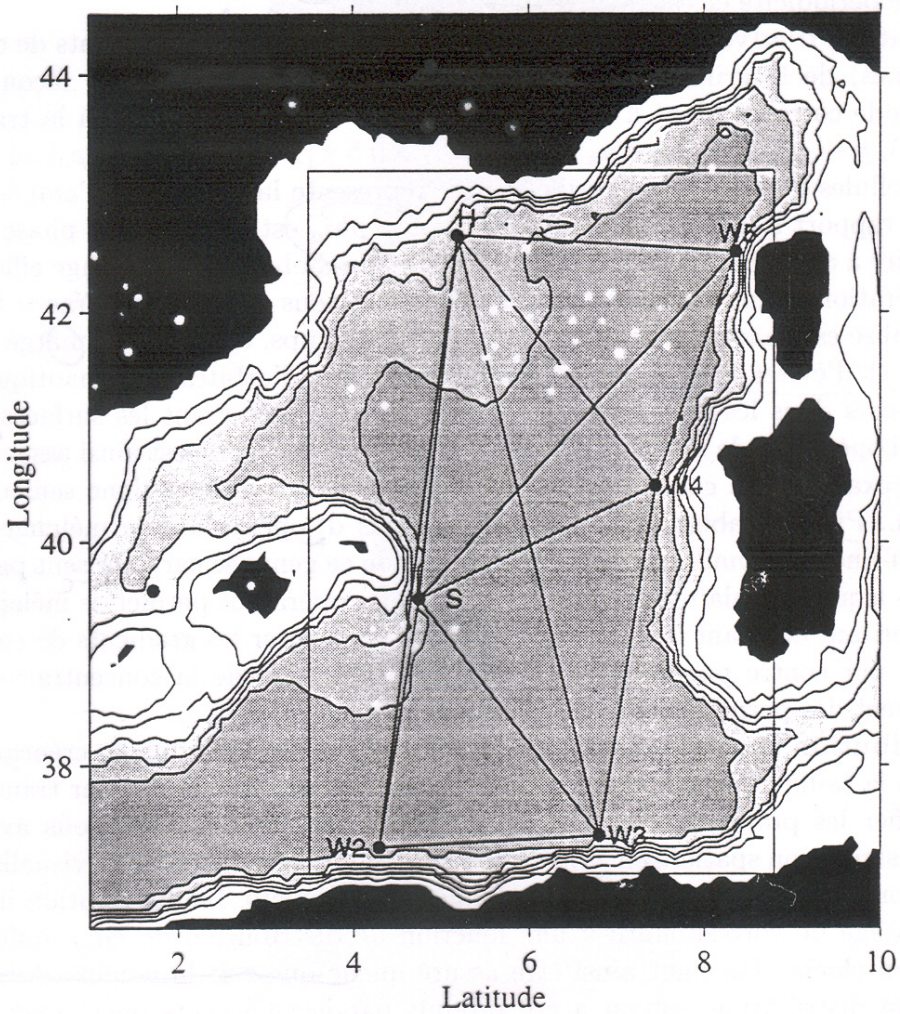


FIG. 1 – *Bassin méditerranéen occidental: réseau d'observation tomographique de l'expérience Thétis 2 et limites du domaine spatial utilisé pour les expériences numériques d'assimilation.*

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

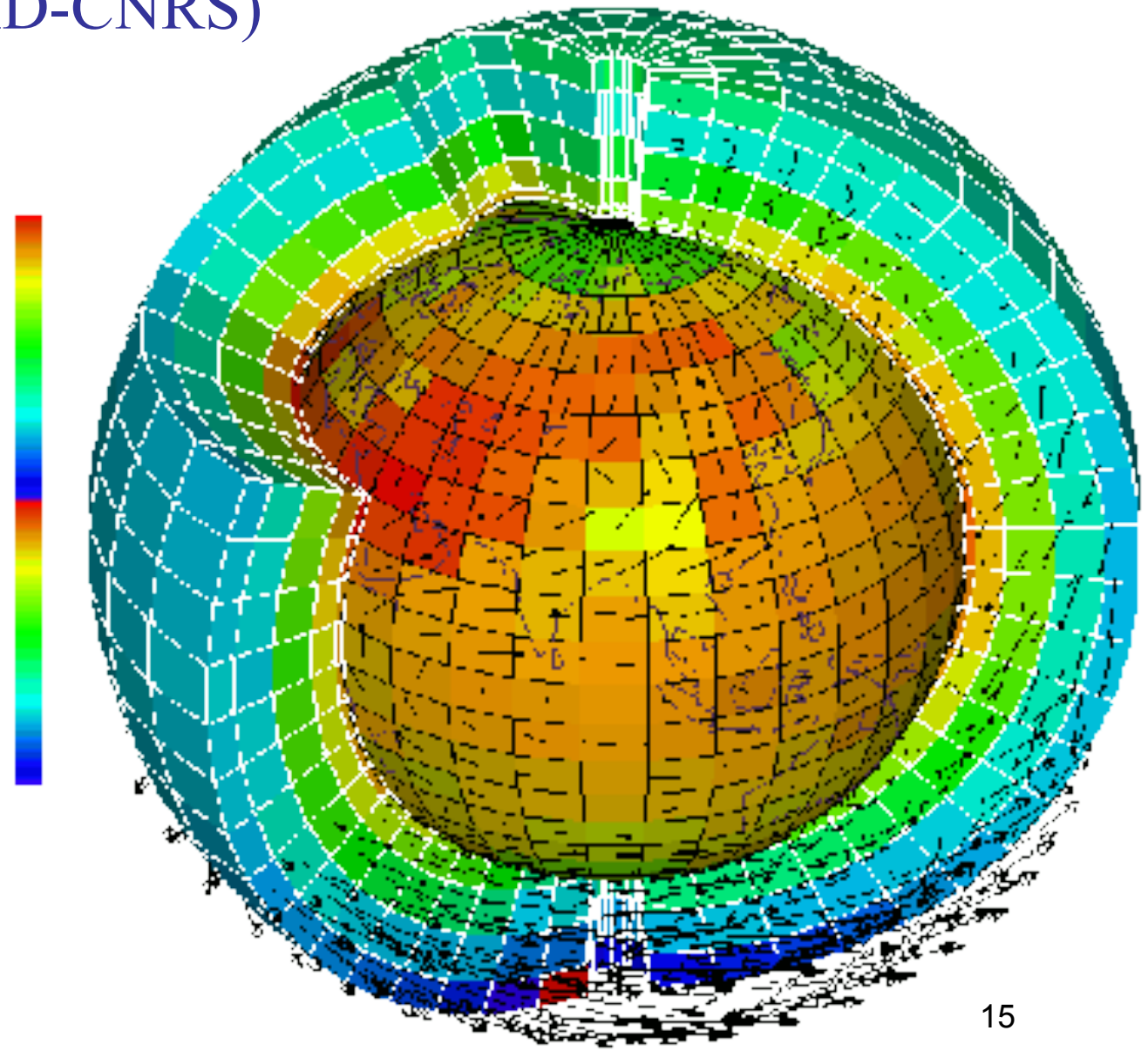
  $f(p, \rho, e) = 0$ $\qquad\qquad$ $(p/\rho = rT, e = C_v T)$

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

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

Physical laws available in practice in the form of a discretized (and necessarily imperfect) numerical model

Schéma de principe d'un modèle atmosphérique (L. Fairhead /LMD-CNRS)



European Centre for Medium-range Weather Forecasts

(ECMWF, Reading, UK)

Horizontal spherical harmonics triangular truncation T1279 (horizontal resolution ≈ 16 kilometres)

91 levels on the vertical (0 - 80 km)

Dimension of state vector $n \approx 1.5\ 10^9$

Timestep = 10 minutes

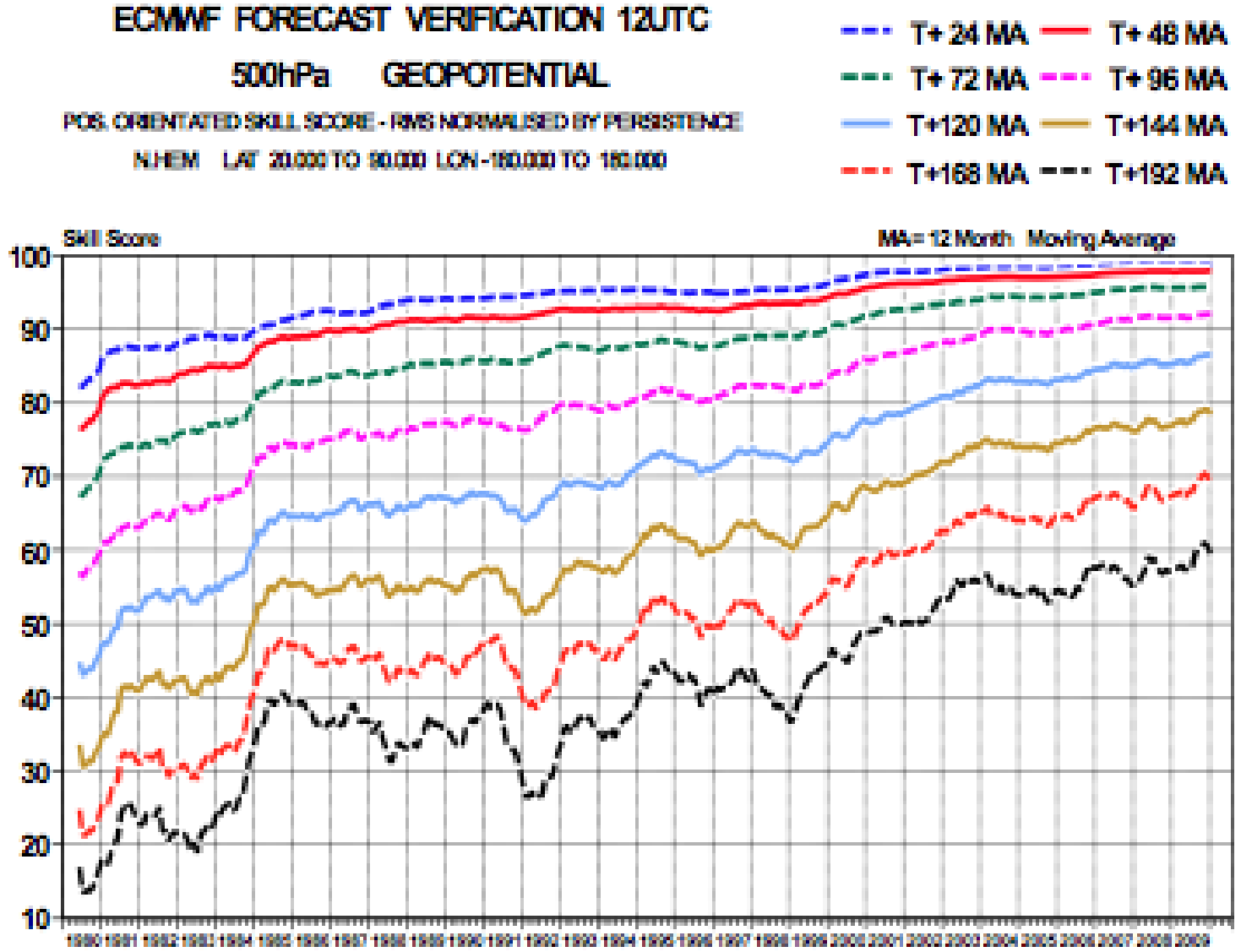


Figure 1: 500 hPa geopotential height skill score for Europe (top) and the northern hemisphere extra-tropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2009 - July 2010.

Extracted from Richardson *et al.*, 2010, ECMWF Technical Memorandum 635

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.

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 and oceanographical observations :

- Very large numerical dimensions ($n \approx 10^7$-10^9 parameters to be estimated, $p \approx 2.10^7$ observations per 24-hour period). Difficulty aggravated in Numerical Weather Prediction by the need for the forecast to be ready in time.

- Non-trivial underlying dynamics.

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).

Assimilation is a problem in bayesian estimation.

Determine the conditional probability distribution for the state of the system, knowing everything we know (unambiguously defined if a prior probability distribution is defined; see Tarantola, 2005).

Bayesian Estimation

Determine conditional probability distribution of the state of the system, given the probability distribution of the uncertainty on the data

$z_1 = x + \zeta_1$ $\qquad$ $\zeta_1 = \mathcal{N}[0, s_1]$

density function $p_1(\zeta) \propto \exp[- (\zeta^2)/2s_1]$

$z_2 = x + \zeta_2$ $\qquad$ $\zeta_2 = \mathcal{N}[0, s_2]$

density function $p_2(\zeta) \propto \exp[- (\zeta^2)/2s_2]$

ζ_1 and ζ_2 mutually independent

What is the conditional probability $P(x = \xi \mid z_1, z_2)$ that x be equal to some value ξ ?

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

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

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

$$P(x = \xi \mid z_1, z_2) \propto p_1(z_1 - \xi)\, p_2(z_2 - \xi)$$
$$\propto \exp[- (\xi - x^a)^2/2s]$$

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

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

$s < (s_1, s_2)$ independent of z_1 and z_2

$z_1 = x + \xi_1$
$z_2 = x + \xi_2$

Same as before, but ξ_1 and ξ_2 are now distributed according to exponential law with parameter a, *i. e.*

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

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)

(Entropy $-\int p \ln p$ always decreases in bayesian estimation)

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

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

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 (1) is verified.

$\Leftrightarrow$ 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 $n \approx 10^{7\text{-}9}$ of present Numerical Weather Prediction models.

- 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 $N \approx O(10\text{-}100)$).

Proportion of resources devoted to assimilation in Numerical Weather Prediction has steadily increased over time.

At present at ECMWF, the cost of 24 hours of assimilation is half the global cost of the 10-day forecast *(i. e.*, including the ensemble forecast).

Sequential Assimilation

- Assimilating model is integrated over period of time over which observations are available. Whenever model time reaches an instant at which observations are available, state predicted by the model is updated with new observations.

Variational Assimilation

- Assimilating model is globally adjusted to observations distributed over observation period. Achieved by minimization of an appropriate *objective function* measuring misfit between data and sequence of model states to be estimated.