On sweat and tears in fulfilling the promise of big data: lessons from meteorological data assimilation

Sylvain Lenfle & Akin Kazakci

SIG Design theory

Mines-ParisTech – 27 january 2015

Context

- **Big data as the « next frontier for innovation, competition and productivity » (McKinsey, 2011),**
- **Major challenge for data centric (Google, Facebook..) and non data-centric companies**
  - Jeff Immelt, GE CEO states that "every industrial company will become a software company"

• Little is known about the extraction of value and the role of design

• What remains unclear is **what kind of capabilities are required for companies to fulfill the promise of big data**.

➔In this work, we argue that, **even for organizations whose core capabilities are based on the manipulation of data, accessing the value of data may require significant efforts** and **radical breakthrough in data analysis**

Big data

- **Underlying hypothesis in the literature :**

Data → Value

- **The problem is obviously more complex.**

Algo (data) → Value

or

$\mathbf{Algo_i\ (data_j) \rightarrow Value}$

Our argument is that the role of algorithms and the complexity of their design is vastly underestimated.

=> **Illustrative case study** : the **emergence of a new dominant design in meteorological data assimilation**

Big data

- **Underlying hypoth**
- **The problem is obv**

Our argument is that their design is vastly u

=> **Illustrative case stuc meteorological data assim**

Methodology : 2 types of sources

- ***The literature***:
  - History / sociology of science on use of space data (from Courrain, 1991 to Edwards, 2010)
  - National Research Council reports on use of space data (2000, 2003 & 2007)
  - Scientific literature on data assimilation.
- ***Interviews*** :
  - O. Talagrand (LMD), le 26/5 & 5/9 2014, ENS, Paris
  - J. Pailleux (MF et ECMWF), le 4/6/2014, Paris
  - Ph. Courtier (MF et ECMWF), le 13/6 & 15/7 2014 (Champs / marne et Paris)
  - JN. Thépaut (ECMWF et MF), le 8/9/ 2014, ECMWF, Reading
  - Fl. Rabier (ECMWF et MF), le 8/9/2014, ECMWF, Reading
  - J. Derber (NOAA, email & interview 8/9/ 2014, ECMWF, Reading)
  - FX Le Dimet (UJF-INRIA) le 11/9/2014 à Grenoble.

Numerical Weather Prediction : a (very) brief overview

- The *Meteorological Research Project* launched by J. Von Neumann and directed by J. Charney in 1946 at Princeton as a part of the Electronic Computer Project (IAS Computer)

- First operational NWP in 1954 in Sweden and in 1955 in the US.

- A central tool of all weather services worldwide

- Over the last 30 years : a steady improvement in the accuracy of forecasts which are increasingly valuable for a wide array of uses (air transportation, agriculture, outside activities, industries, storm warnings, etc)

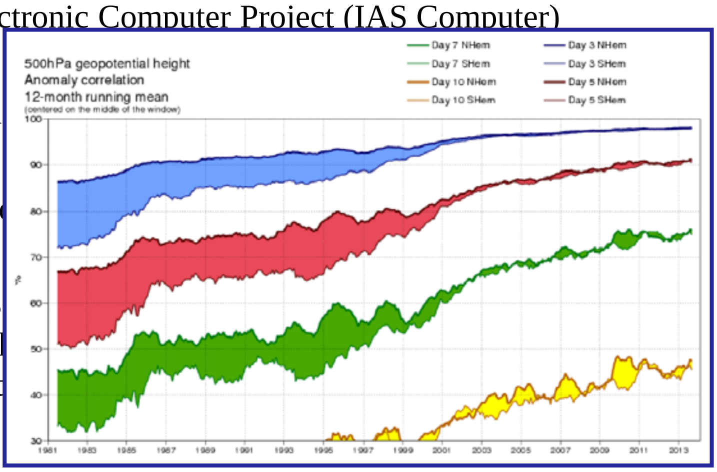
Numerical Weather Prediction : a (very) brief overview

- The *Meteorological Research Project* launched by J. Von Neumann and directed by J. Charney in 1946 at Princeton as a part of the Electronic Computer Project (IAS Computer)

- First operation

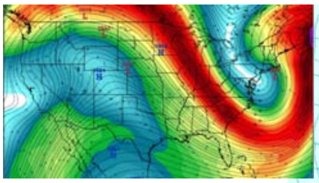
- A central tool

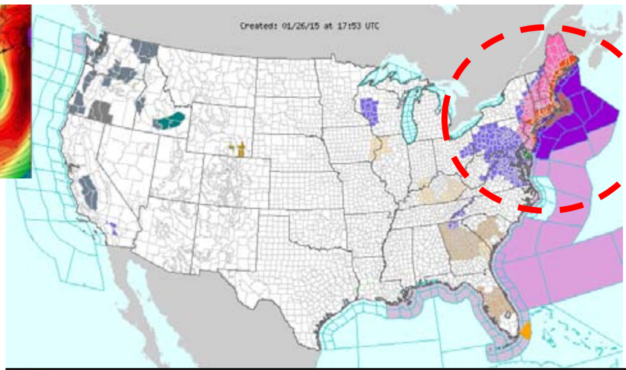
- Over the last 3 forecasts whic (air transportat warnings, etc)

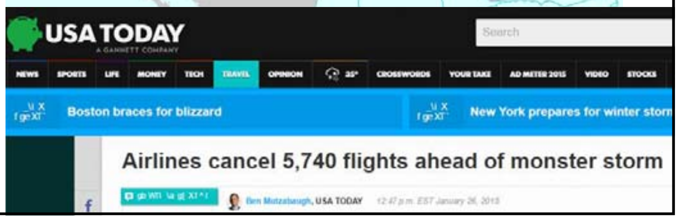


Blizzard alert over New-York

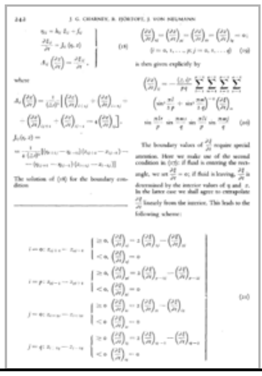
January 26-27, 2015

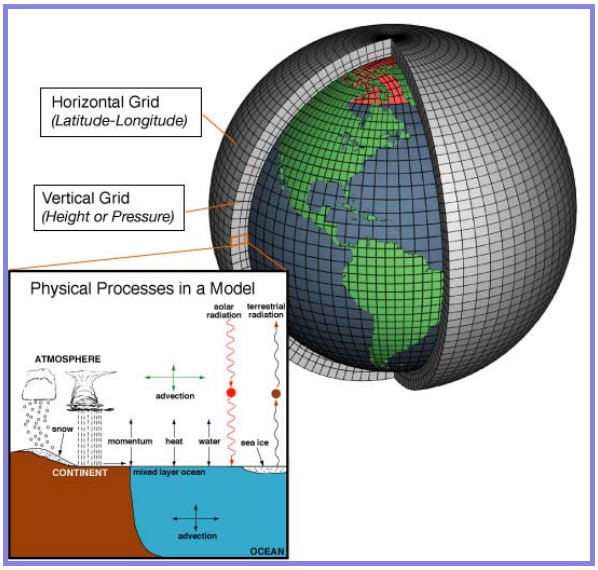




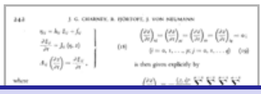


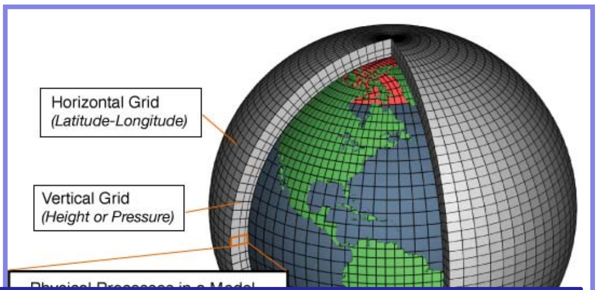
How does it works (1) ?





How does it works (1) ?





=> A NWP model uses known physical laws and the corresponding variables : temperature, pressure, humidity, etc.

« *Computer models demanded a degree of standardization never previously needed in meteorology. NWP required that values be entered at every gridpoint, both horizontal and vertical, even where no observations existed. Missing gridpoint values had to be interpolated from observations, or even (if necessary) filled in with climatological norms* » (Edwards, 2010, p. 252)

How does it works (2) ?

Data assimilation

Prediction

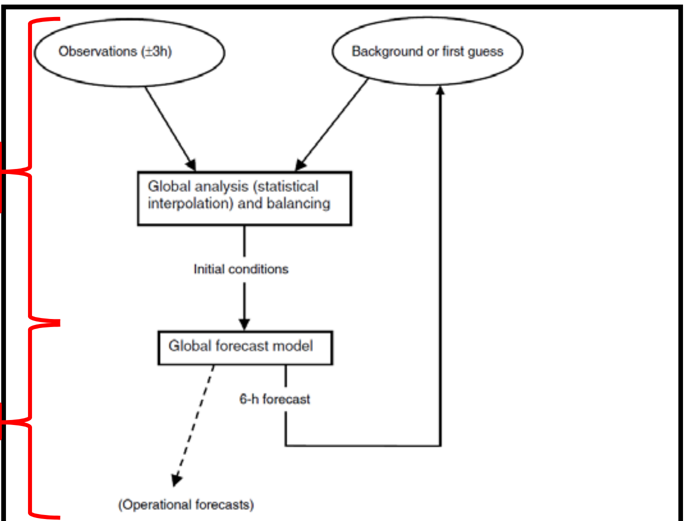


Figure 1.4.2: Flow diagram of a typical intermittent (6-h) data assimilation cycle.

How does it works (2) ?

Data assimilation

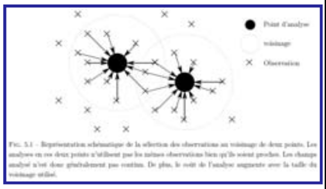
Prediction

- « ***NWP is an initial/boundary value problem*** **»,** (Kalnay, 2003), since initial conditions can radically alter the final result (Lorenz, 1963 & 1971 « *Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?* »)
- **Data assimilation** : « *using all the available information to determine as accurately as possible the state of atmospheric (or oceanic) flow* » (Talagrand, 1997).
- **Challenges of meteorological data assimilation**
  - Very large numerical dimensions (***10^9 data*** to initialize ECMWF model)…
  - … aggravated in NWP by the need for the forecast to be ready in time
  - Non-trivial, actually chaotic, underlying dynamics

Data assimilation : the Dominant design in the 80-90's

- **First step : Algo$_1$ (data$_1$) → Value**
  - Data$_1$= synoptic and direct measurements (typically balloons)
  - Algo$_1$= optimal interpolation (OI)

    « *OI became the operational analysis scheme of choice during the 1980's and early 1990's* » (Kalnay, 2003, p. 150)



- **Known limits in the 80's**
  - Of the data : limited spatial coverage
  - Of the algorithms : 1) unability to treat uncertainty dynamically (background error covariance matrix frozen), 2) difficulty to determine initial conditions precisely with more sophisticated models

A breakthrough in instruments : satellite soundings

- First atmospheric sounder launched in 1969
- Considered operational by meteorologist since the launch of TIROS-N in 1978

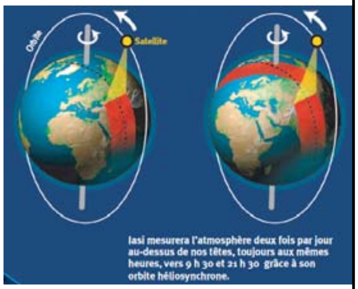
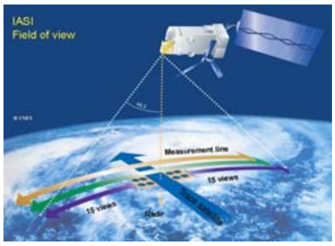
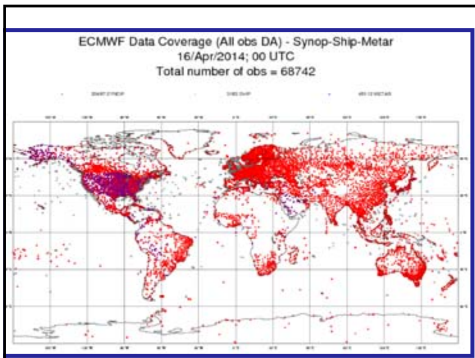


TABLE B.1 Chronology of Early Satellite Sounders Flown on NASA and NOAA Satellites

Instrument	Satellite	Primary Period of Operation
SIRS-A	Nimbus-3	1969-1971
SIRS-B	Nimbus-4	1970-1972
ITPR, NEMS, SCR	Nimbus-5	1972-1975
VTPR	ITOS series	1972-1979
HIRS, SCAMS, PMR	Nimbus-6	1975-1979
HIRS-2, MSU, PMR	NOAA series	1978-1998
VAS	GOES series	1980-1996
HIRS-3, AMSU	NOAA series	1998-present

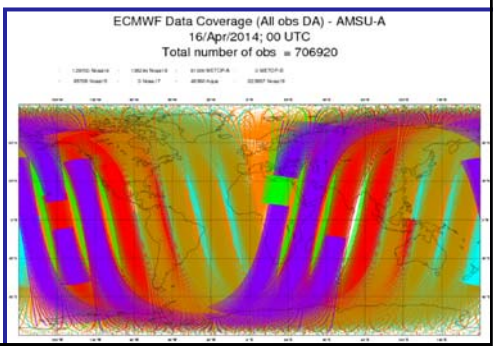
NOTE: The acronyms are spelled out in Appendix E.
SOURCE: Kalnay et al. (1996).





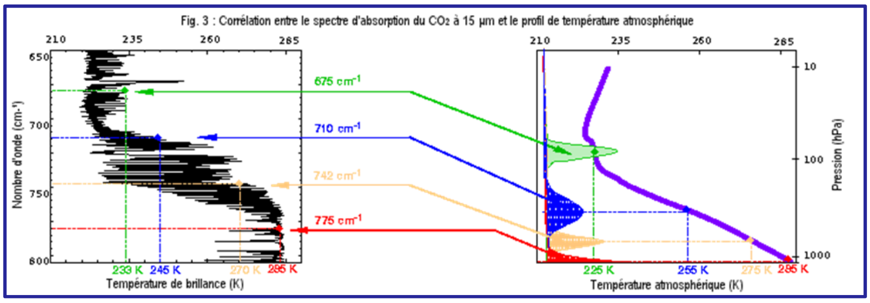
Traditional coverage

Satellite coverage



… but a breakthrough in data too.

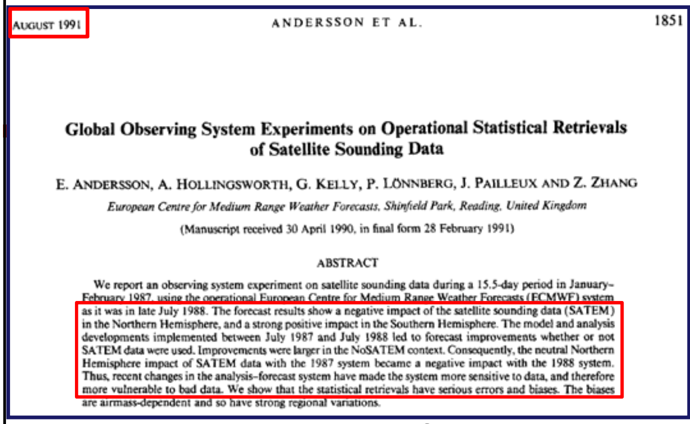
- **Satellite soudings are indirect observations : they measure radiances, not temperature.**
- **The link between radiances and temperature is physically very complex** (known as the ***radiative transfer equation***).



The problem

- **First approach to assimilate radiances = apply the dominant design (*satellite-to-model approach*)** :
  - "*If you can make them look like radiosonde data we can use them*" (NMC Director, 1969 in NRC, 2003)
  - "*Profiles of temperatures soundings were displayed to make them look like radiosondes because radiance could not be handled by numerical models*"
- **Problem : temperature profile provided by NOAA are of poor quality**. "*TOVS and MSU have only ~4-5 "pieces of information", the rest came from climatology*" (Eyre, 1993; Kalnay, 2009, etc)
- **Result : during 10 years (1981 – 1991) the impact of radiances on NWP is negligible or negative.**

=> $\text{Algo}_1\ (\text{data}_2) \rightarrow$ No Value


AUGUST 1991 ANDERSSON ET AL. 1851

Global Observing System Experiments on Operational Statistical Retrievals of Satellite Sounding Data

E. ANDERSSON, A. HOLLINGSWORTH, G. KELLY, P. LÖNNBERG, J. PAILLEUX AND Z. ZHANG

European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, United Kingdom

(Manuscript received 30 April 1990, in final form 28 February 1991)

ABSTRACT

We report an observing system experiment on satellite sounding data during a 15.5-day period in January–February 1987, using the operational European Centre for Medium Range Weather Forecasts (ECMWF) system as it was in late July 1988. The forecast results show a negative impact of the satellite sounding data (SATEM) in the Northern Hemisphere, and a strong positive impact in the Southern Hemisphere. The model and analysis developments implemented between July 1987 and July 1988 led to forecast improvements whether or not SATEM data were used. Improvements were larger in the NoSATEM context. Consequently, the neutral Northern Hemisphere impact of SATEM data with the 1987 system became a negative impact with the 1988 system. Thus, recent changes in the analysis–forecast system have made the system more sensitive to data, and therefore more vulnerable to bad data. We show that the statistical retrievals have serious errors and biases. The biases are airmass-dependent and so have strong regional variations.

he dominant design

n use them" (NMC

make them look like
numerical models"

- **Problem : temperature profile provided by NOAA are of poor quality**. "*TOVS and MSU have only ~4-5 "pieces of information", the rest came from climatology*" (Eyre, 1993; Kalnay, 2009, etc)

- **Result : during**
NWP is neglig

=>

<u>Consequence</u> : **a NASA / NOAA conflict** that leads to « *a two decade long hiatus in new instruments for the polar orbiters. (…) and the instrument generation of 1978, with only minor updates, continued to fly through the end of the century*" (Conway, 2008, p. 91 92)

The emergence of a new dominant design : variational assimilation (Algo$_2$)

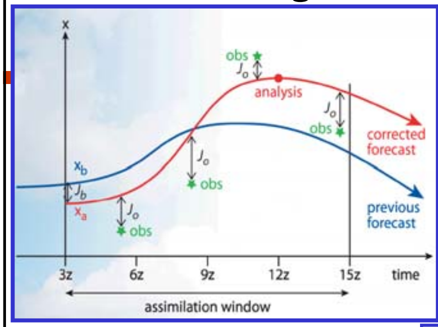
- To overcome the limits of optimal interpolation, meteorologists explore new data assimilation methods.

  => **<u>Variational assimilation</u>** becomes a major research topic in the late 80's and 90's. Originally to improve uncertainty integration in assimilation, not for satellite data.

- The logic of this method is to minimize a cost function « *which measures the misfit between the estimate and the information weighted by its statistical quality*" (Courtier, 1997) over a given time windows (thus its name : 4D-VAR)

- The breakthrough comes from the **<u>interactions of meteorologists and mathematicians of optimal control</u>**.

The emergence of a new dominant design :



Variational equations: for 1D-Var, 3D-Var, 4D-Var

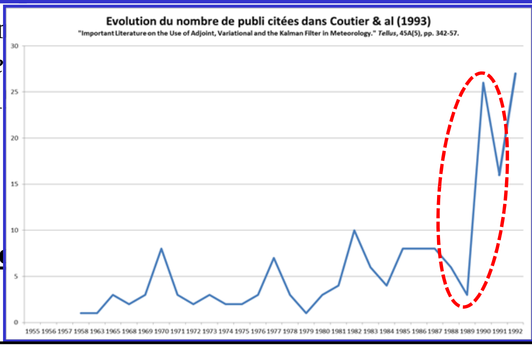
Minimize:

$$J[x] = \tfrac{1}{2}(x-x^b)^T B^{-1}(x-x^b) + \tfrac{1}{2}(y^o-H[x])^T (E+F)^{-1}(y^o-H[x])$$

where x contains the NWP model state
x^b is background estimate of x (short-range forecast)
B is its error covariance,
y^o is vector of measurements
$H[\ldots]$ is "observation operator" or "forward model", mapping state x into "measurement space"
E is error covariance of measurements,
F is error covariance of forward model.

- The logic of this method is to r *the misfit between the estimate statistical quality*" (Courtier, 1 name : 4D-VAR)

- The breakthrough comes from **<u>mathematicians of optimal c</u>**



When meteorology meets optimal control

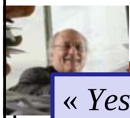
<u>Olivier Talagrand</u>

- Emeritus Research Director at the Laboratoire de Météorologie Dynamique
- Silver Medal of the CNRS in 2004 & L.F. Richardson Medal of the European Geoscience Union, 2014 for pioneering work in data assimilation.
- Ph . D Supervisor of Ph. Courtier, JN Thépaut et Fl. Rabier
- Director of the Scienfic Board of the ECMWF during IFS/ARPEGE project.

<u>François-Xavier Le Dimet</u>

- Emeritus Professor in mathematics at Grenoble University / INRIA.
- Fellow of the American Meteorological Society for his pioneering work in data assimilation
- Applied mathematician, expert in optimal control methods (JL Lions, 1971).
- The only one who does not work in a meteorological lab.

When meteorology meets optimal control



•
•
•

« *Yes, I have studied in my Ph. D [1977] a simplistic assimilation method (but I didn't know how to do differently). It's only when I've met FX Le Dimet that I've understood how to do what I think desirable. The problem was not the variational idea, quite simple in itself, than its numerical implementation in large systems.* ***The technique of adjoint equations provides the solution to the problem****. That being so, this was very far from the question of satellite data assimilation.* »

Mail O. Talagrand, le 15 septembre 2014

- **Le Dimet, FX and O Talagrand. 1986**. "Variational Algorithms for Analysis and Assimilation of Meteorological Observations: Theoretical Aspects." *Tellus*, 36A, pp. 97-110. [**cité + de 1300 fois]**
- **Co-organizers of the 1st WMO international symposium on data assimilation in meteorology and oceanography** **(Clermont-Ferrand, july 1990).**

Implementation (1) : the challenge

(Courtier, 1997)

ues. In numerical weather prediction, the practical difficulty is that it is impossible to use Eqs. 2 and 3 directly. $\mathbf{B}$, for example, is a matrix of size $10^7 \times 10^7$ which is about 1000 times the total archiving capabilities of ECMWF and one million times the memory size of the current computers. The scientific difficulty of data assimilation is to find algorithms which simplify Eqs. 2 and 3 to an affordable amount of computer resources, while preserving some of the essential characteristics.

« *Had I know* [in 1982 – 1983] *what it cost* [in computing power] *I would have given up immediatly* [laugh]!! *We didn't suspect the difficulties* [of operational implementation].

FX Le Dimet, Grenoble, le 11/9/2014

Implementation (2) : the IFS / ARPEGE project (1987 – 1997)

- **Launched in 1987 by ECMWF and Meteo-France.**
- **Implementing 4D-VAR means designing** (Andersson & Thepaut, 2008)
  - A forecast model and its adjoint.
  - The observation operators linking the observed variables to the model quantities; code to compute the observation cost function *J*o and its gradient.
  - The first-guess operator, to incorporate information from recent analyses; code to compute the first-guess cost function *J*b and its gradient.
  - Balance operators to ensure the appropriate relationship between mass and wind.
  - General minimisation algorithm, to seek the analysis as the minimum of the cost function *J*o+*J*b.
  - A suitable solution algorithm that can take advantage of the computing power available on multi-processor computing platforms.

- **The driving force : Philippe Courtier**
- **Result : operational integration of radiances at ECMWF in 1996 for 3D-VAR, and 1997 for 4D-VAR.**

Algorithmic breakthrough in implementation too.

Q. J. R. Meteorol. Soc. (1994), **120**, pp. 1367–1387 551.509.313.22

A strategy for operational implementation of 4D-Var, using an incremental approach

By P. COURTIER*, J.-N. THÉPAUT and A. HOLLINGSWORTH
European Centre for Medium-range Weather Forecasts, UK

(Received 9 July 1993; revised 13 January 1994)

SUMMARY

An order of magnitude reduction in the cost of four-dimensional variational assimilation (4D-Var) is required before operational implementation is possible. Preconditioning is considered and, although it offers a significant reduction in cost, it seems that it is unlikely to provide a reduction as large as an order of magnitude. An approximation to 4D-Var, namely the incremental approach, is then considered and is shown to produce the same result at the end of the assimilation window as an extended Kalman filter in which no approximations are made in the assimilating model but in which instead a simplified evolution of the forecast error is introduced. This approach provides the flexibility for a cost–benefit trade-off of 4D-Var to be made.

BUT : developing and implementing Algo$_2$ is a 10 years effort which mobilize approx.100 FTE

… in a data-centric / research-based organization

=> Algo$_2$ (data$_2$) → ↑ Value (better forecasts)

Digging very deep in Algo_2 : the forecast error covariance *B*

- In 1993-94, despites years of hard work, variational assimilation still does not assimilate radiances satisfactorily (no improvements in forecasts).

- Ph. Courtier understood that **the problem comes from the fine-tuning of the background error covariance matrix**, which is excessively complex.

  « At this moment, in 1993-94, I remember that B has been adjusted in 1985-86 by Hollingsworth & Lundberg to make the best use of wind data collected by airplanes. And I realize that this leads to a bad use of temperature data. The correlation functions were filtering out the temperature information of radiances. This is why the impact of 3D-VAR remains marginal. We change this. And it works. » (interview with Ph. Courtier, june 13, 2014).

Digging very deep in Algo_2 : the forecast error covariance *B*

- In 1993-94, despites years of hard work, variational assimilation still does not assimilate radiances satisfactorily (no improvements in forecasts).

- Ph. Courtier understood that **the problem comes from the fine-tuning of the background error covariance matrix**, which is excessively complex.

  « At this moment, in 1993-94, I remember that B has been adjusted in 1985-86 by Hollingsworth & Lundberg to make the best use of wind data collected by airplanes. And I realize that this leads to a bad use of temperature data. The correlation functions were filtering out the temperature information of radiances. This is why the impact of 3D-VAR remains marginal. We change this. And it works. » (interview with Ph. Courtier, june 13, 2014).

$$J[x] = \tfrac{1}{2}(x-x^b)^T B^{-1}(x-x^b) + \tfrac{1}{2}(y^o-H[x])^T (E+F)^{-1}(y^o-H[x])$$

B the background error covariance matrix : « *There is a great deal of science that goes in it* » (JN Thépaut, 8/9/14). « *The forecast error covariance* ***B*** *is the most difficult error covariance to estimate, and has a crucial impact on the results* » (Kalnay, 2003, p. 161).

« You have to imagine the sweat and tears during all this long years » (JN Thépaut, Head of data division, ECMWF)

- During the long journey from **Algo$_1$ (data$_1$) to Algo$_2$ (data$_2$)** we observe **3 breakthroughs**
  - ***On data*** : from direct measurement to satellite soudings
  - ***On assimilation methods*** : from OI to 3D-VAR
  - ***On implementation*** : incremental method, recoding of the model, coding of its adjoint, more powerful supercomputers, etc.

- Keep in mind that Meteorology is a very « favorable » context.
  - Data-centric organization
  - Infrastructure available (research centers, supercomputers, WMO…)
  - Models & data are « known »
  - No debate on « value »
  - Huge operational constraints (2 forecasts / day)

  ⇒We know what we are looking for (better temperature measurements)

« You have to imagine the sweat and tears during all this long years » (JN Thépaut, Head of data division, ECMWF)

- During the long journey from **Algo$_1$ (data$_1$) to Algo$_2$ (data$_2$)** we observe **3 breakthroughs**
  - ***On data*** : from direct measurement to satellite soudings

- This raises important questions for the « big data » strategy of firms & the building of big data design capability
- Currently there is no equivalent of engineering department for algorithms design (people, models, design rules, knowledge…)
  - ➢ Who is in charge of algorithm design in industrial firms (and in data-centric organizations) ?
  - ➢ Links with the emerging Data Science ?
  - ➢ Development of the technical infrastructure ?
  - ➢ Transfer of design methods to data ?

  => Lots of exciting work ahead…

slenfle@hotmail.com
akin.kazakci@mines-paristech.fr

http://crg.polytechnique.fr/home/lenfle/FR