Introduction to Data Assimilation

Alan O'Neill

Data Assimilation Research Centre

University of Reading

DARC

What is data assimilation?

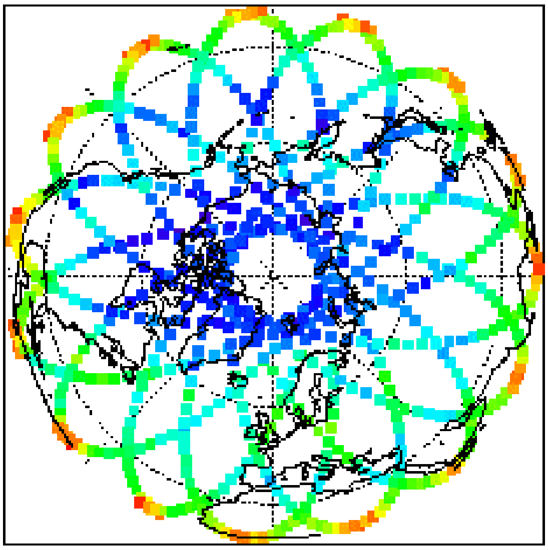
Data assimilation is the technique whereby observational data are combined with output from a numerical model to produce an optimal estimate of the evolving state of the system.

DARC

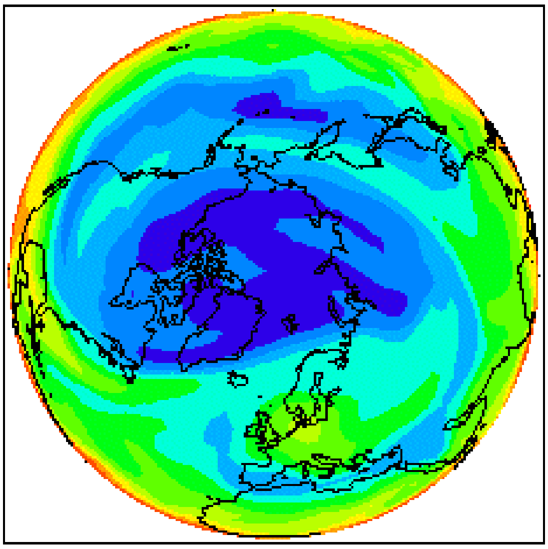
Why We Need Data Assimilation

- range of observations
- range of techniques
- different errors
- data gaps
- quantities not measured
- quantities linked

DARC



DARC



DARC

Some Uses of Data Assimilation

- Operational weather and ocean forecasting
- Seasonal weather forecasting
- Land-surface process
- Global climate datasets
- Planning satellite measurements
- Evaluation of models and observations

DARC

What We Want To Know

$\mathbf{x}(t)$ atmos. state vector

$\mathbf{s}(t)$ surface fluxes

$\mathbf{c}$ model parameters

$$\mathbf{X}(t) = (\mathbf{x}(t), \mathbf{s}(t), \mathbf{c})$$

DARC

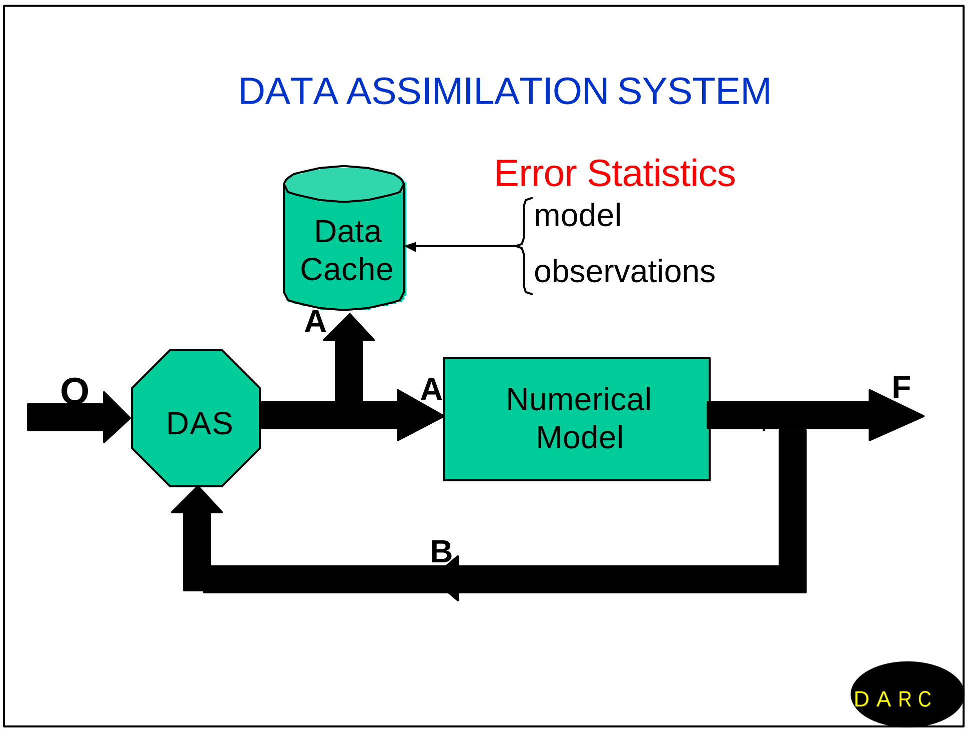
What We Also Want To Know

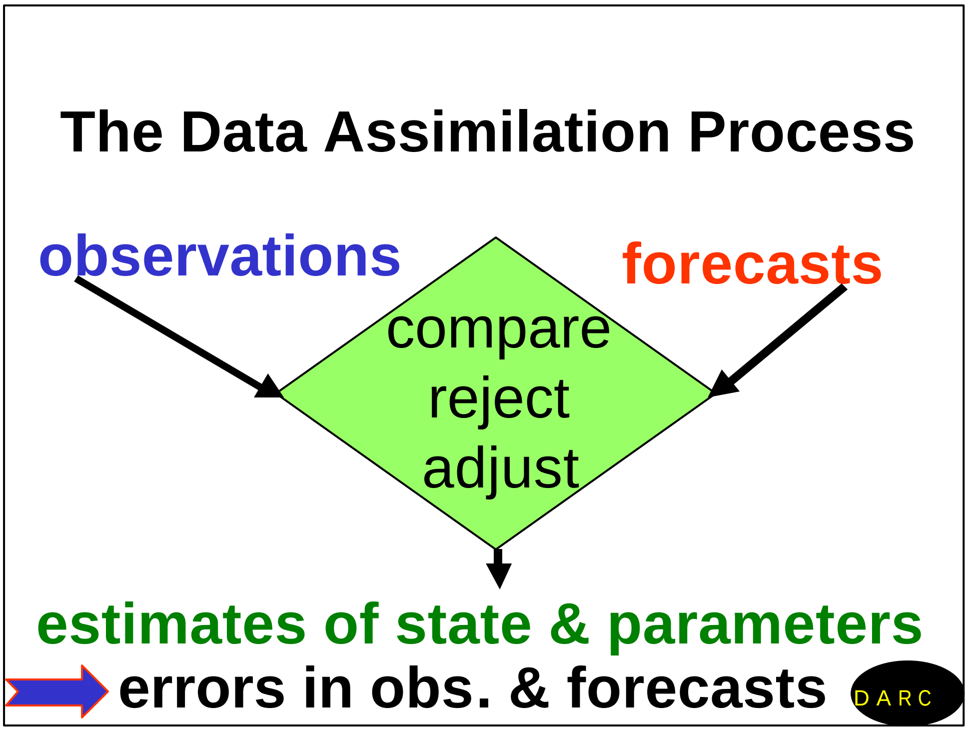
Errors in models

Errors in observations

What observations to make

DARC


DATA ASSIMILATION SYSTEM
Error Statistics
model
observations
Data Cache
A
O
DAS
A
Numerical Model
F
B
DARC


The Data Assimilation Process
observations
forecasts
compare
reject
adjust
estimates of state & parameters
errors in obs. & forecasts
DARC

Retrievals and Assimilation

- Problems with retrievals
  - *a priori* state and poorly known errors
- Problems with assimilation of radiance
  - systematic errors and cloud clearing
  - expensive (multi-channels)
- Need effective interface
  - Information content with error analysis

DARC

Statistical Approach to Data Assimilation

DARC

DARC

Minimum Variance

Combination of Data

Unbiased, Uncorrelated Errors

$$\tilde{x} = \alpha x_1 + \beta x_2 \qquad \alpha + \beta = 1$$

$$\mathrm{Var}(x_1) = \sigma_1^2 \qquad \mathrm{Var}(x_2) = \sigma_2^2$$

DARC

Minimum Variance Combination of Data Unbiased, Uncorrelated Errors

$$\mathrm{Var}(\tilde{x}) = \alpha^2 \sigma_1^2 + (1-\alpha)^2 \sigma_2^2$$

$$\frac{\partial}{\partial \alpha} \mathrm{Var}(\tilde{x}) = 2\alpha\sigma_1^2 - 2(1-\alpha)\sigma_2^2 = 0$$

$$\Rightarrow \alpha, \beta$$

DARC

Minimum Variance Combination of Data Unbiased, Uncorrelated Errors

$$\tilde{x} = \left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}\right)\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}$$

$$\mathrm{var}(\tilde{x}) \leq \min(\sigma_1^2, \sigma_2^2)$$

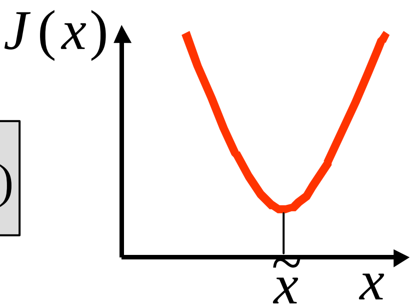
Best Linear Unbiased Estimate

DARC

Variational Method

$$J(x) = \frac{(x - x_1)^2}{\sigma_1^2} + \frac{(x - x_2)^2}{\sigma_2^2}$$

$\tilde{x}$ at $\min J(x)$



DARC

Maximum Likelihood Estimate

- Obtain or assume probability distributions for the errors
- The best estimate of the state is chosen to have the greatest probability, or maximum likelihood
- If errors normally distributed,unbiased and uncorrelated, then states estimated by minimum variance and maximum likelihood are the same

DARC

Multivariate Case

state vector $\mathbf{x}(t) = \begin{pmatrix} x_1 \\ x_2 \\ \bullet \\ \bullet \\ x_n \end{pmatrix}$

observation vector $\begin{pmatrix} y_1 \\ y_2 \\ \bullet \\ y_m \end{pmatrix}$

DARC

Variance becomes Covariance Matrix

- **Errors in x_i are often correlated**
  - **spatial structure in flow**
  - **dynamical or chemical relationships**
- **Variance for scalar case becomes Covariance Matrix for vector case** COV
- **Diagonal elements are the variances of x_i**
- **Off-diagonal elements are covariances between x_i and x_j**
- **Observation of x_i affects estimate of x_j**

DARC

Estimating Covariance Matrix for Observations, O

- O usually quite simple:
  - diagonal or
  - for nadir-sounding satellites, non-zero values between points in vertical only
- Calibration against independent measurements

DARC

Estimating Covariance Matrix for Model, B

- Use simple analytical functions constructed experimentally, e.g.

$$B_{ij} = \sigma_i \sigma_j \exp(-d_{ij})$$

where d_{ij} is the horizontal distance between $\mathbf{x}_i$ and $\mathbf{x}_j$

- Run ensemble of forecasts from slightly different conditions and analyse error growth.

DARC

Methods of Data Assimilation

- Optimal interpolation (or approx. to it)
- 3D variational method (3DVar)
- 4D variational method (4DVar)
- Kalman filter (with approximations)

DARC

Optimal Interpolation

observation

observation operator

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - H(\mathbf{x}_b))$$

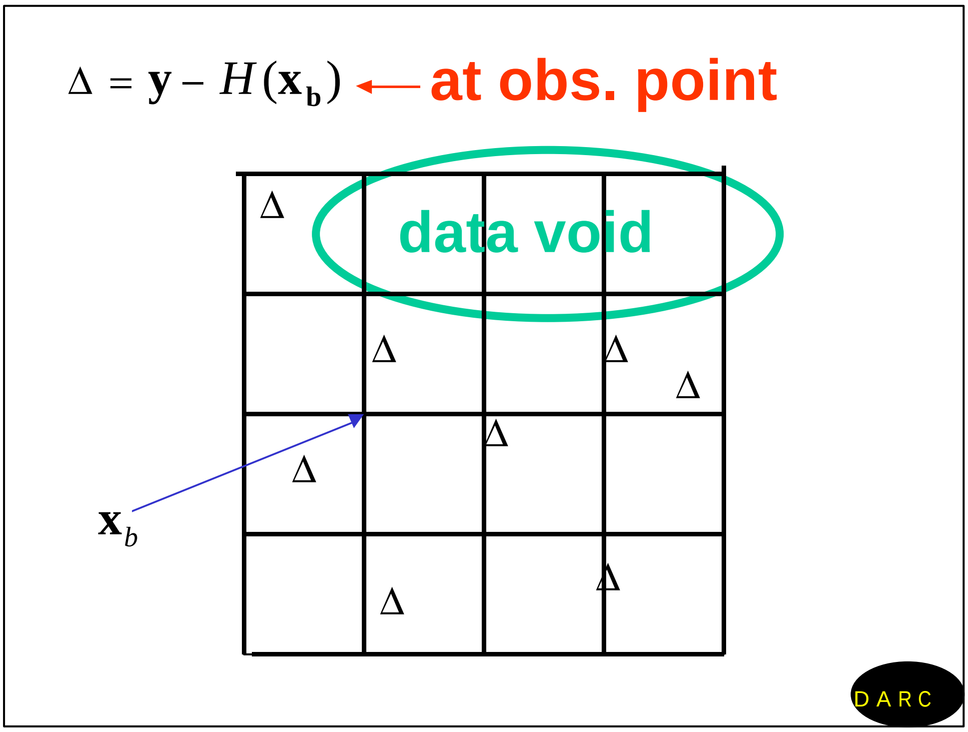
"analysis"

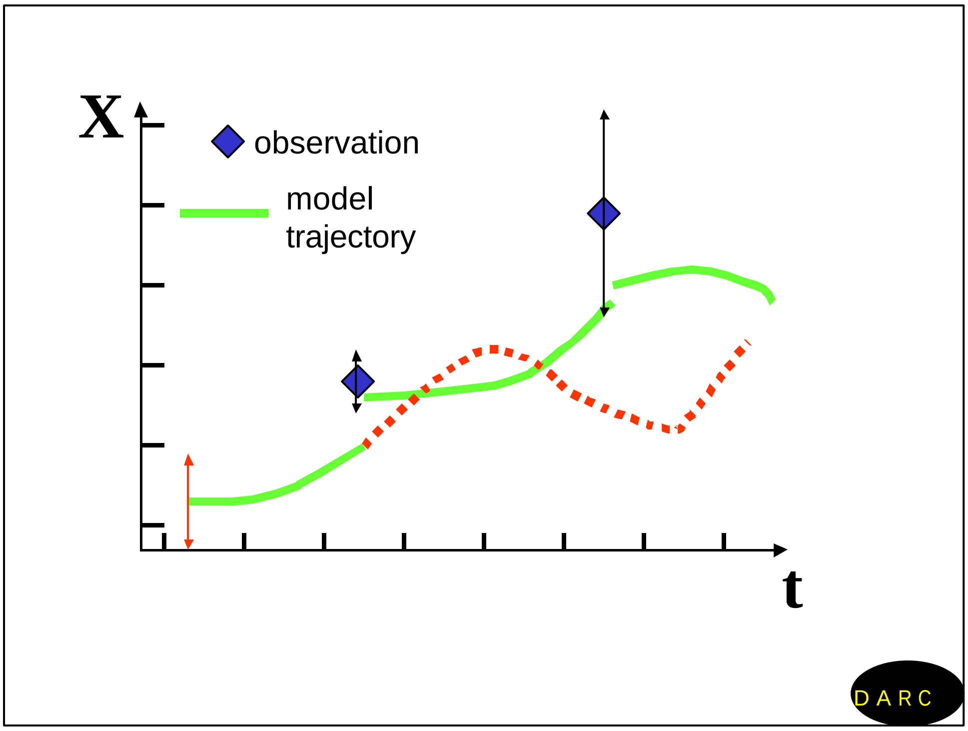
"background" (forecast)

- linearity $H \rightarrow \mathrm{H}$
- matrix inverse
- limited area

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{O})^{-1}$$

DARC


$\Delta = \mathbf{y} - H(\mathbf{x}_b)$ ← at obs. point
data void
Δ
$\mathbf{x}_b$
DARC


X
observation
model
trajectory
t
DARC

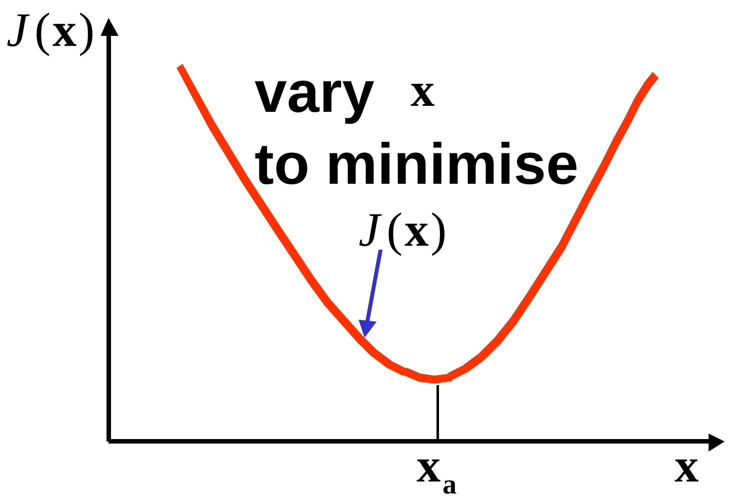
Data Assimilation: an analogy

Driving with your eyes closed:

open eyes every 10 seconds and correct trajectory

DARC

Variational Data Assimilation



DARC

Variational Data Assimilation

$$J(\mathbf{x}) =$$
$$(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}-\mathbf{x}_b) +$$
$$(\mathbf{y}-H(\mathbf{x}))^T \mathbf{O}^{-1} (\mathbf{y}-H(\mathbf{x}))$$

nonlinear operator

assimilate y directly

global analysis

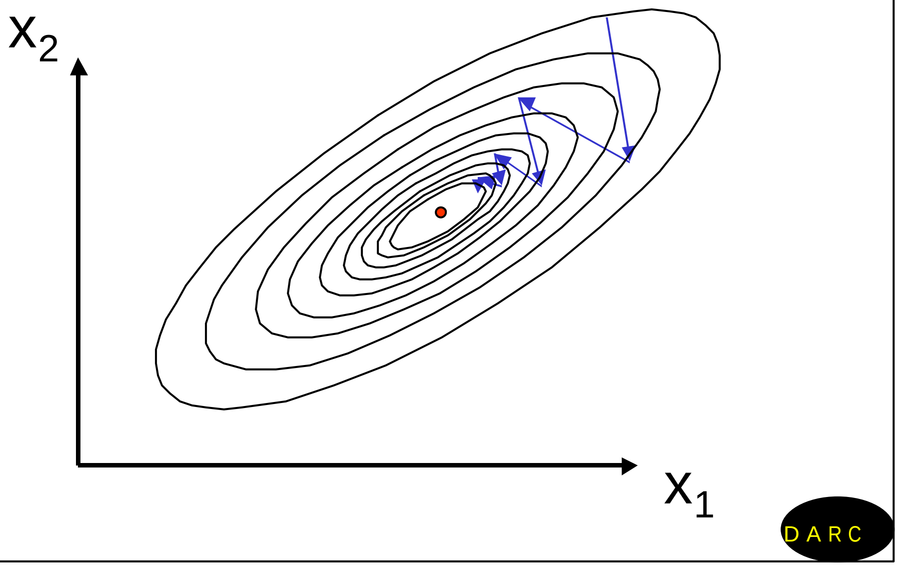
DARC

Choice of State Variables and Preconditioning

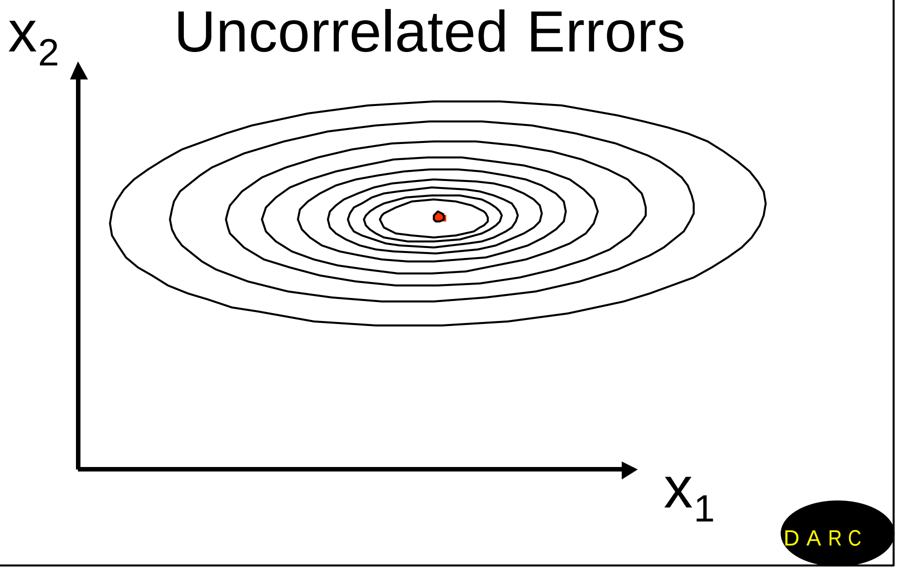
- **Free to choose which variables to use to define state vector, *x(t)***
- **We'd like to make B diagonal**
  - **may not know *covariances* very well**
  - **want to make the minimization of J more efficient by "preconditioning": transforming variables to make surfaces of constant J nearly spherical in state space**

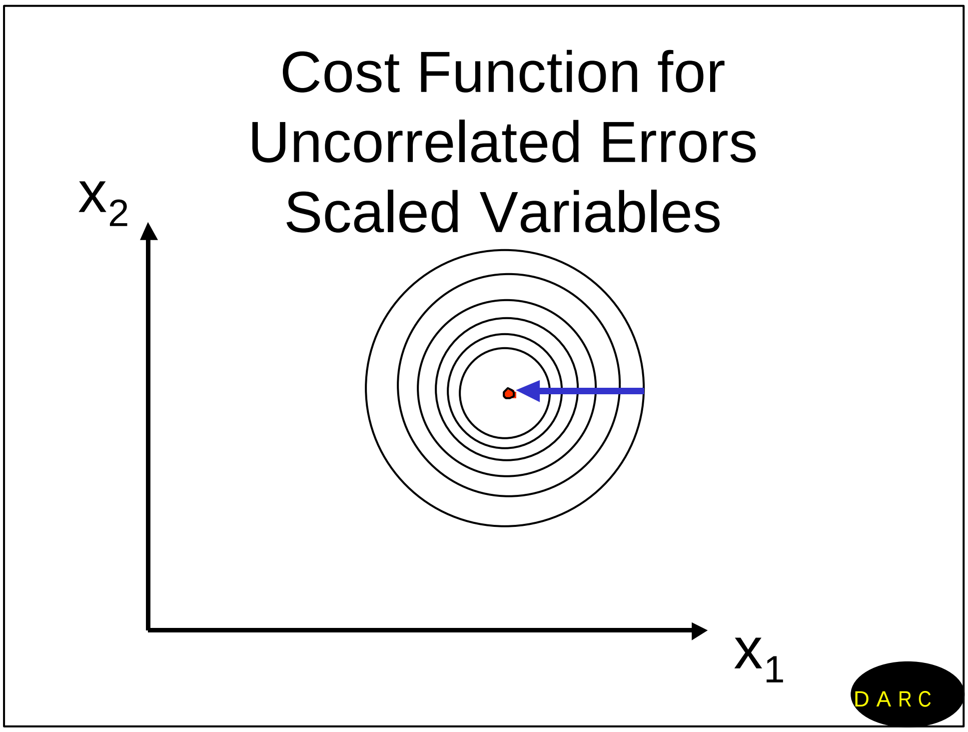
DARC

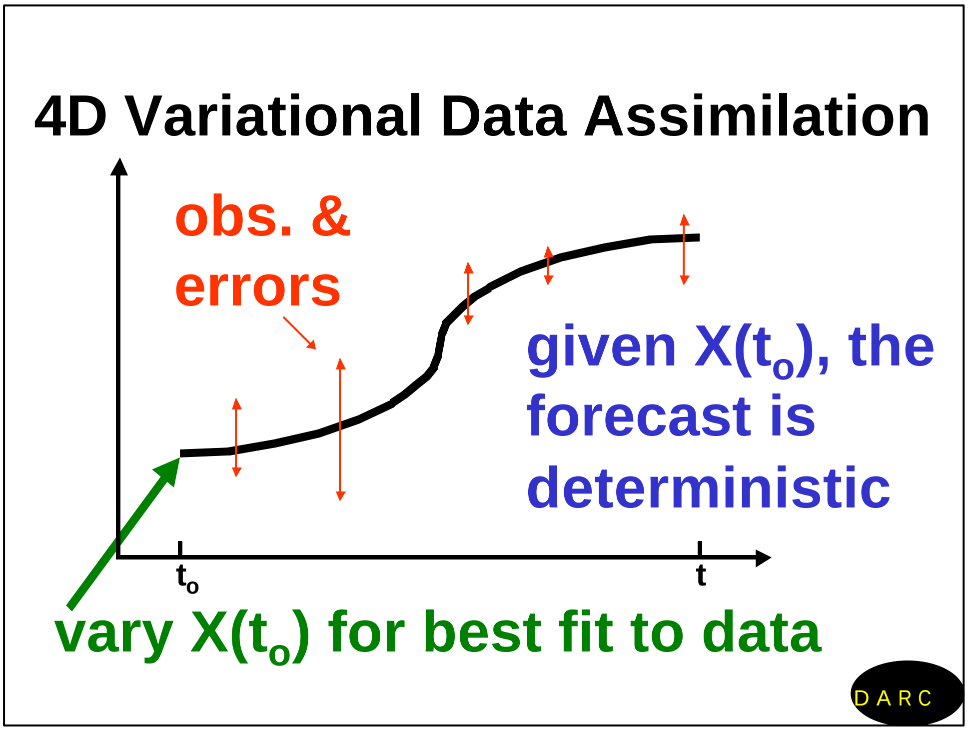
Cost Function for Correlated Errors



Cost Function for Uncorrelated Errors




Cost Function for
Uncorrelated Errors
Scaled Variables
X2
X1
DARC


4D Variational Data Assimilation
obs. &
errors
given X(to), the
forecast is
deterministic
to
t
vary X(to) for best fit to data
DARC

4D Variational Data Assimilation

- **Advantages**
  - consistent with the governing eqs.
  - implicit links between variables
- **Disadvantages**
  - very expensive
  - model is strong constraint

DARC

Problem

model error covariances

DARC

Kalman Filter
(*expensive*)

Use model equations to propagate B forward in time.

B ⟶ B(t)   KAL

Analysis step as in OI

DARC

What are the benefits of data assimilation?

- Quality control
- Combination of data
- Errors in data and in model
- Filling in data poor regions
- Designing observational systems
- Maintaining consistency
- Estimating unobserved quantities

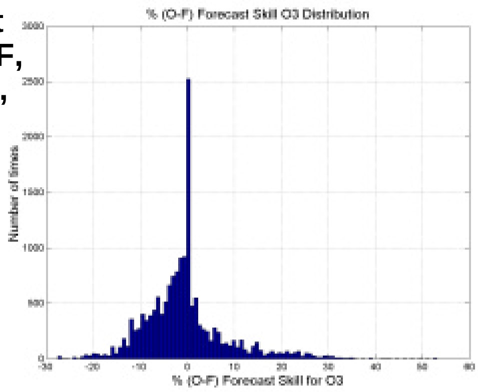
DARC

Some Applications of Data Assimilation

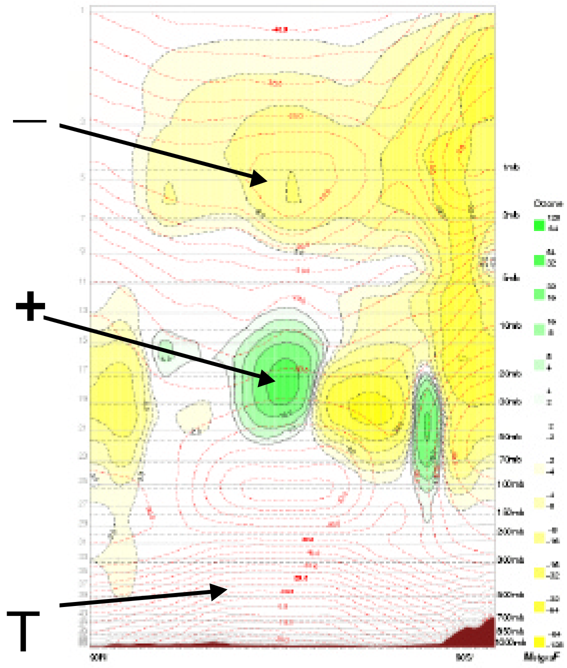
DARC

Skill Measures: Observation Increment, (O-F)

- The difference between the forecast from the first guess, F, and the observations, O, also known as observed-minus-background differences or the innovation vector.
- This is probably the best measure of forecast skill.



DARC

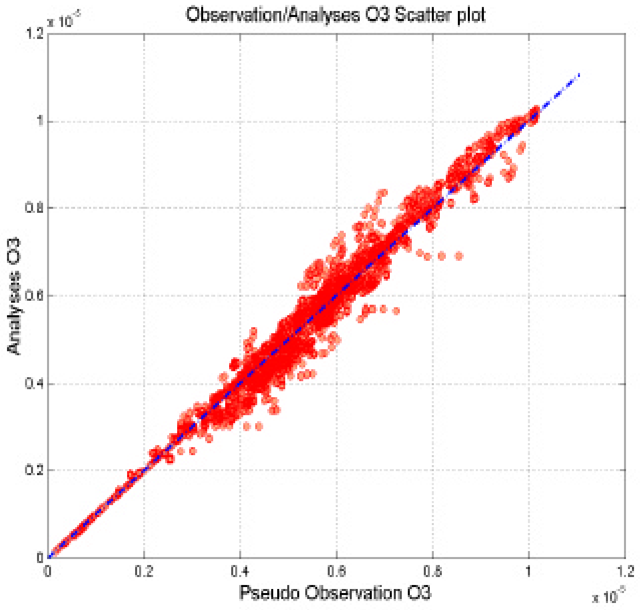


ECMWF

Ozone monthly-mean analysis increment.

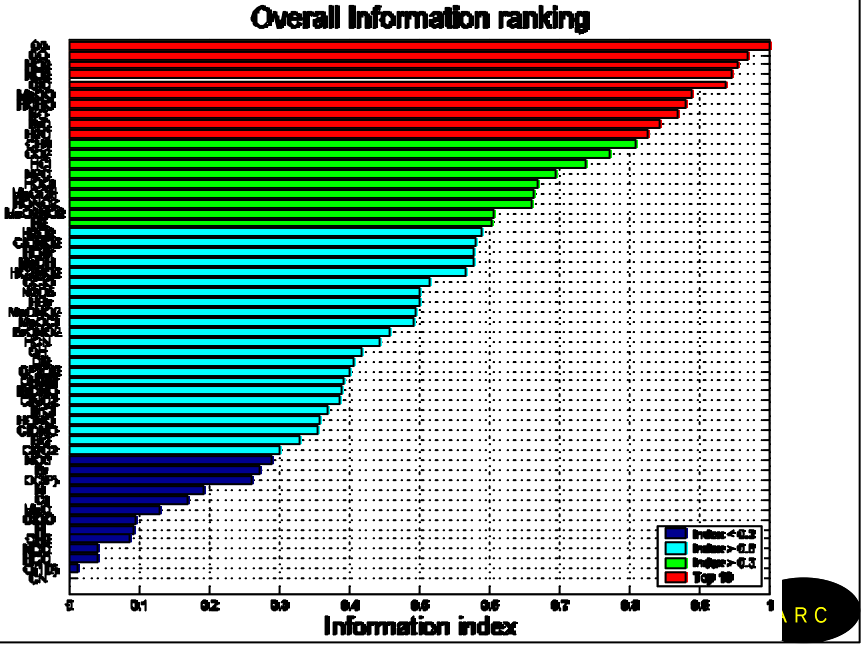
Sept. 1986

DARC

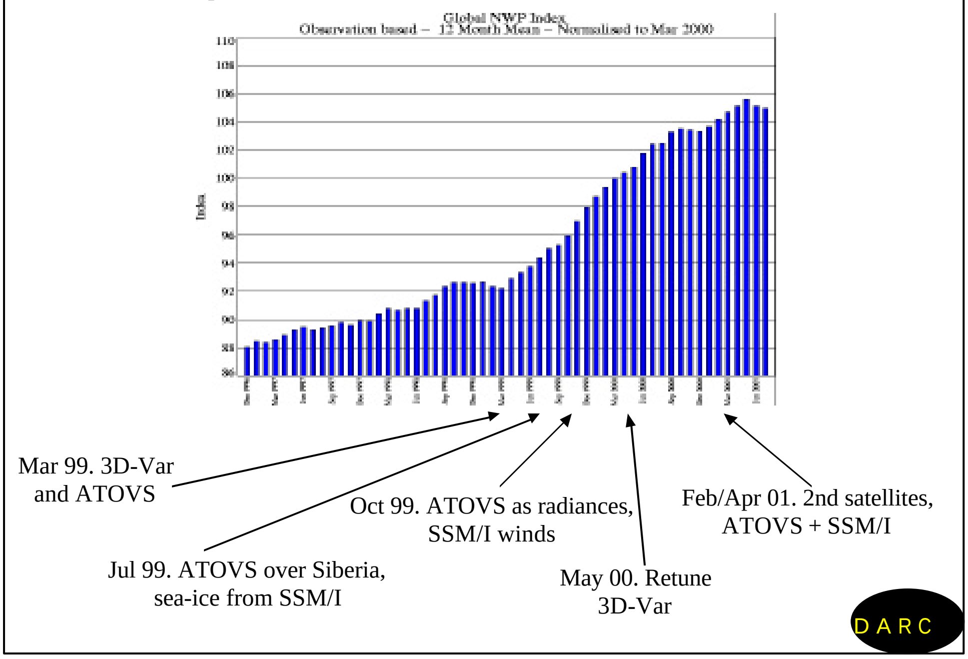


DARC

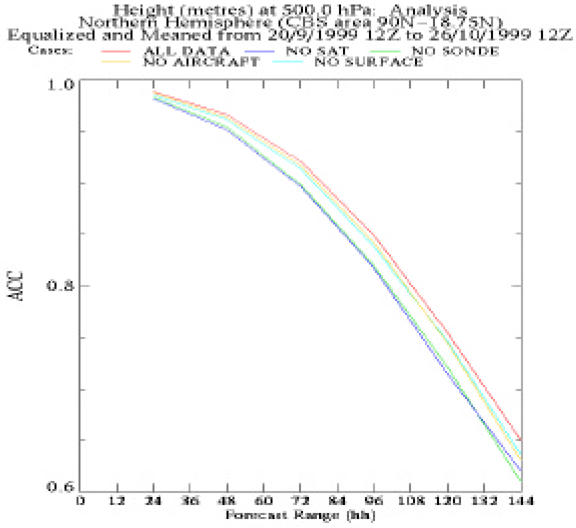
Best observations to make to characterize the chemical system



Impact on NWP at the Met Office

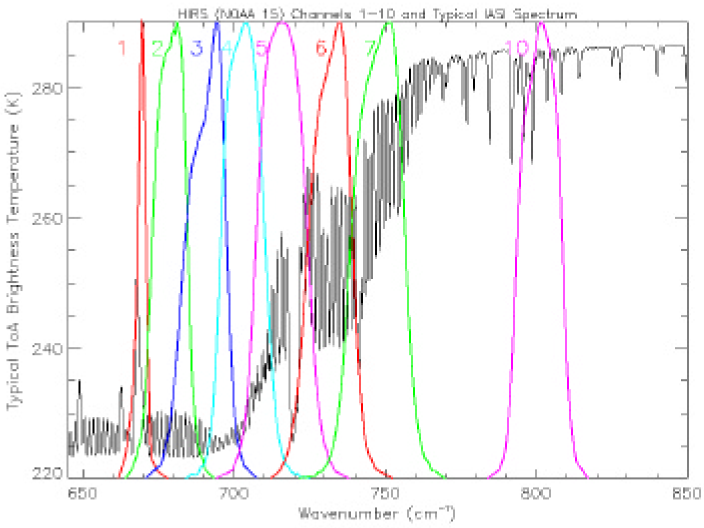


Impact of satellite data in NWP

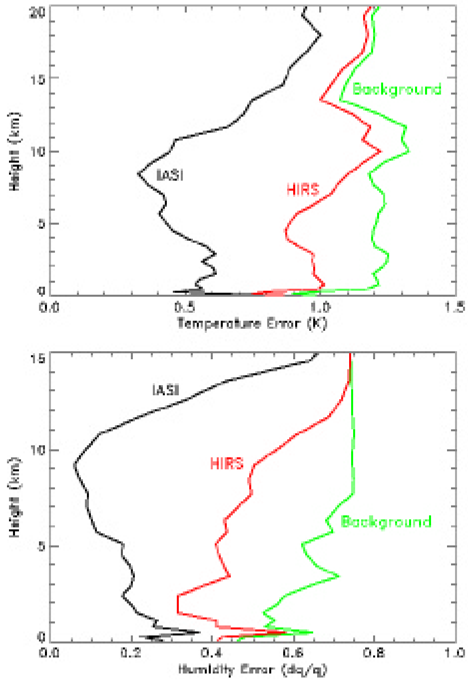


DARC

Future Advanced Infrared Sounders

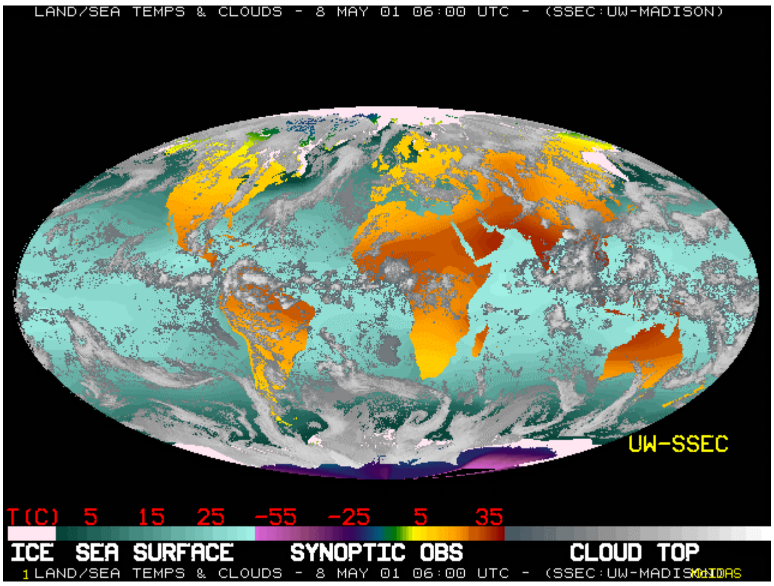


DARC

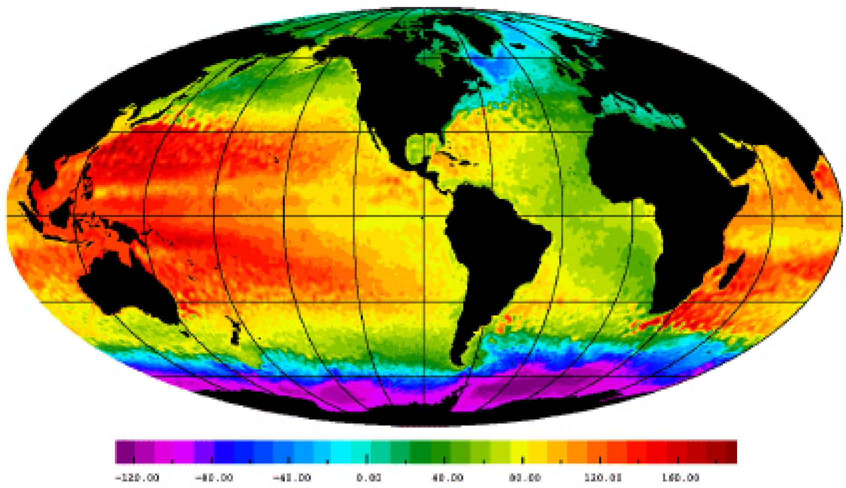


IASI vs HIRS

DARC

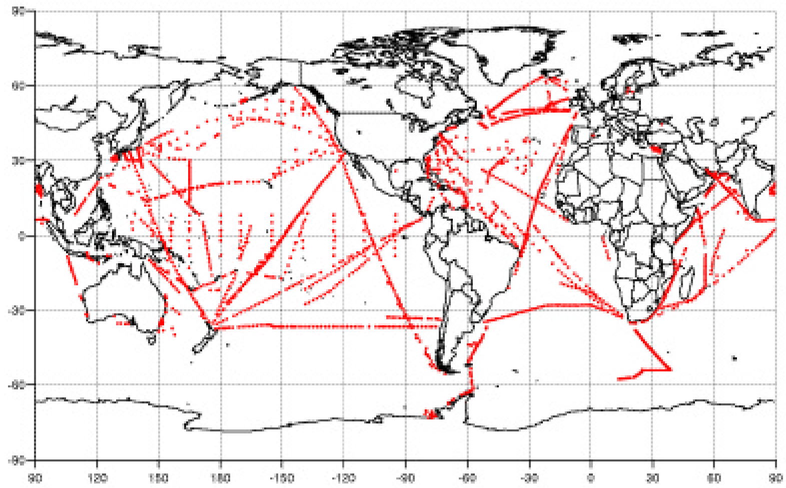


DARC

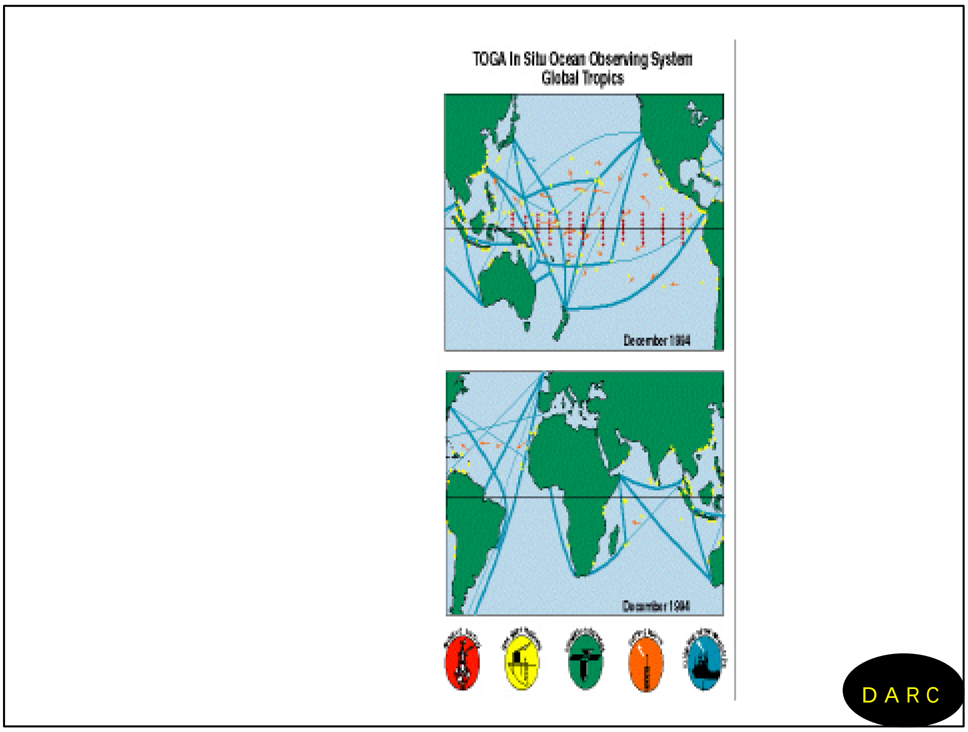


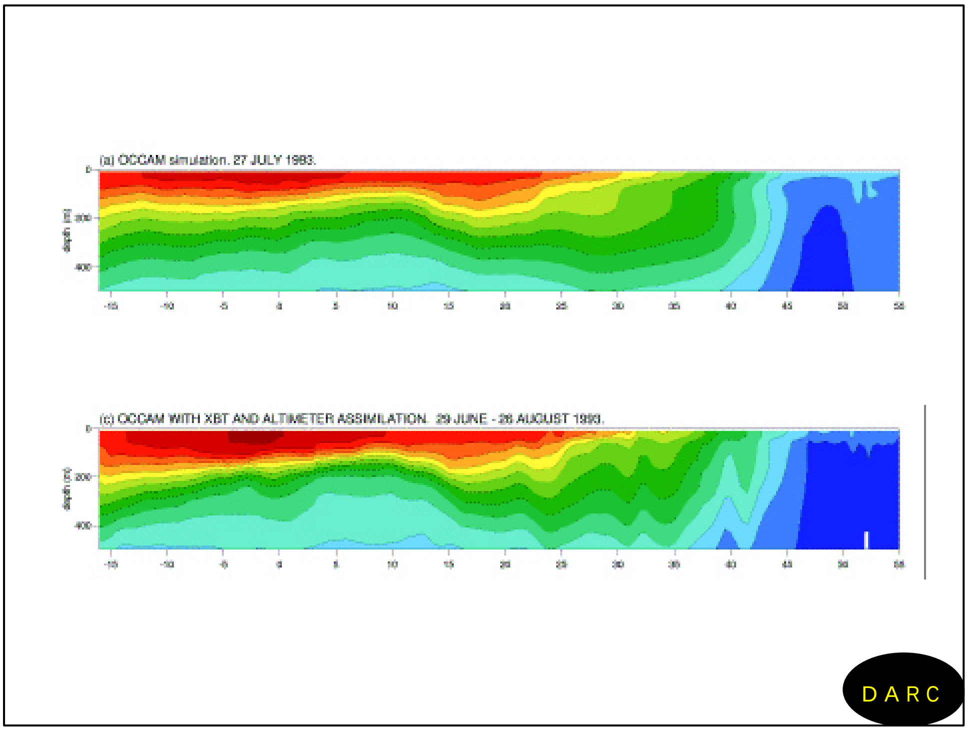
DARC

XBT data assimilated in March 1996.

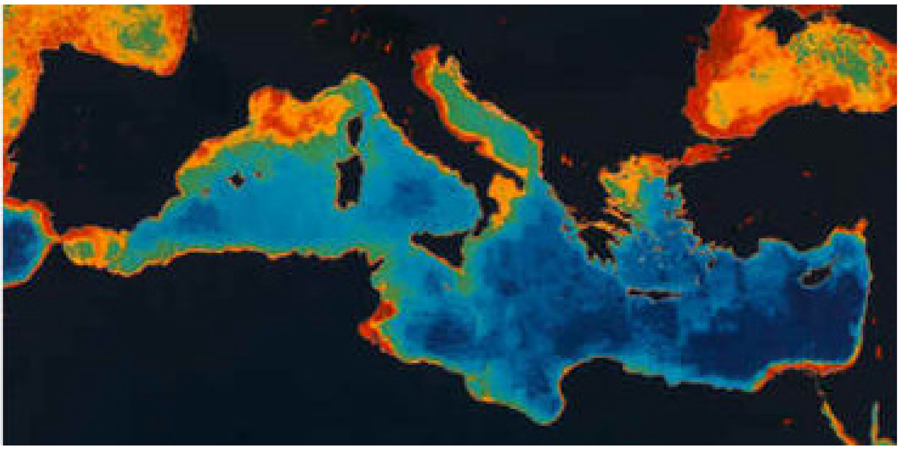


DARC


TOGA In Situ Ocean Observing System
Global Tropics
December 1994
December 1994
DARC

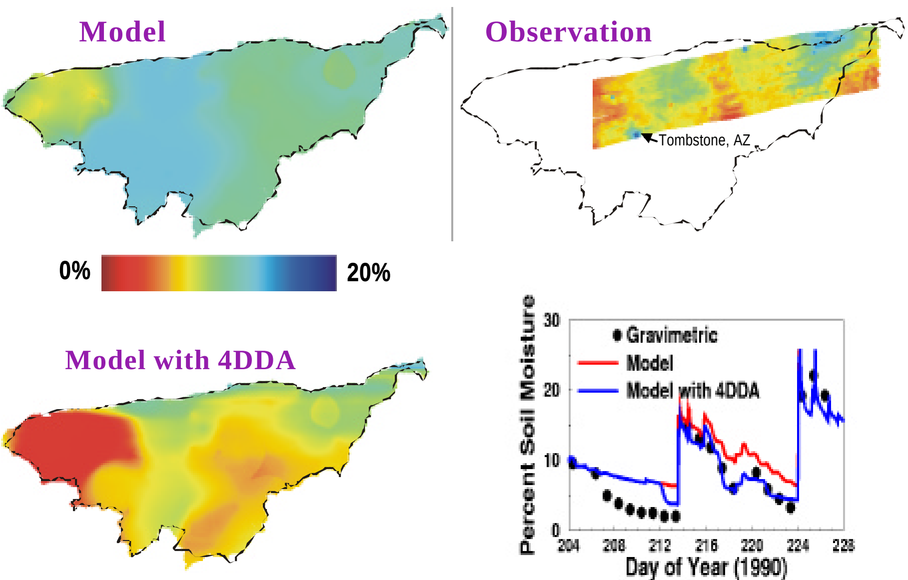

(a) OCCAM simulation. 27 JULY 1993.
(c) OCCAM WITH XBT AND ALTIMETER ASSIMILATION. 29 JUNE - 28 AUGUST 1993.
DARC

MERIS ocean colour



DARC

Regional Scale: Walnut Gulch (Monsoon 90)

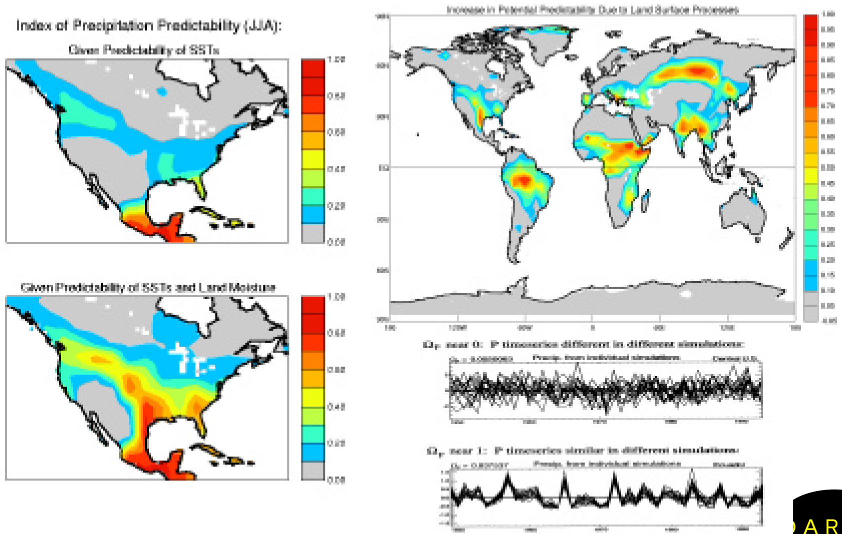


Houser et al., 1998

DARC

Land Initialization: Motivation

- Knowledge of soil moisture has a greater impact on the predictability of summertime precipitation over land at mid-latitudes than Sea Surface Temperature (SST).



ARC

Conclusions

- Data assimilation should be essential part of “ground-segment” of all satellite missions.
- Aim should be to provide all data in “near real time”.
- Need to find optimal ways to assimilate data – efficient interfaces.
- Challenges: errors, resolution, data.

DARC