Innovation, increments, and residuals: Definitions and examples

Patrick Heimbach
MIT, EAPS, Cambridge, MA

Outline

1. Why should we / do we care?

2. How does “(re-)analysis” really work?

3. Conclusions

Balance the global (momentum, enthalpy, freshwater) budget !

Goal: use closed heat budgets to understand causes heat anomalies

Term-by-term budget analysis of origin of heat content anomalies through time:

$$\underbrace{\int_0^t \frac{H_t}{\rho_o C_p V}\, dt}_{\equiv(T-T_o)} = \underbrace{\int_0^t \frac{C_{adv}}{\rho_o C_p V}\, dt}_{\equiv T_{adv}} + \underbrace{\int_0^t \frac{C_{diff}}{\rho_o C_p V}\, dt}_{\equiv T_{diff}} + \underbrace{\int_0^t \frac{Q_{net}}{\rho_o C_p V}\, dt}_{\equiv T_Q}$$

Example: decomposition of advective term:

$$C_{adv} = -\rho_o C_p \int_{-D}^{\eta} \boldsymbol{\nabla}\cdot(\overline{\boldsymbol{u}T} + \overline{\boldsymbol{u}^* T})\, dz = \underbrace{-\rho_o C_p \int_{-D}^{\eta} \boldsymbol{\nabla}\cdot(\overline{\boldsymbol{u}}\overline{T})\, dz}_{\equiv C_{lin}} \underbrace{-\rho_o C_p \int_{-D}^{\eta} \boldsymbol{\nabla}\cdot(\overline{\boldsymbol{u}'T'} + \overline{\boldsymbol{u}^* T})\, dz}_{\equiv C_{bol}},$$

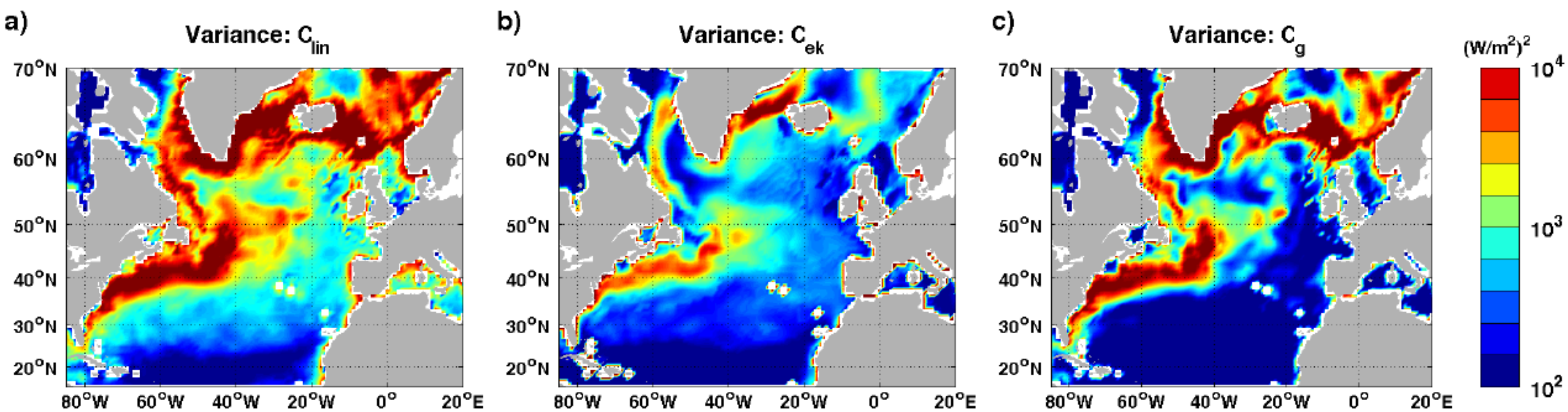
Buckley et al. (2013)

Goal: use closed heat budgets to understand causes heat anomalies

Another example:
decomposition of Ekman and geostrophic components:

$$C_{ek}(\mathbf{u}_{\mathbf{ek}}, w_{ek}, \theta) = C_{ek}(\bar{\mathbf{u}}_{\mathbf{ek}}, \overline{w}_{ek}, \bar{\theta}) + \underbrace{C_{ek}(\mathbf{u}_{\mathbf{ek}}{}', w'_{ek}, \bar{\theta})}_{C^{v}_{ek}} + \underbrace{C_{ek}(\bar{\mathbf{u}}_{\mathbf{ek}}, \overline{w}_{ek}, \theta')}_{C^{T}_{ek}} + \underbrace{C_{ek}(\mathbf{u}'_{\mathbf{ek}}, w'_{ek}, \theta')}_{C^{vT}_{ek}}$$

$$C_{g}(\mathbf{u}_{\mathbf{g}}, w_{g}, \theta) = C_{g}(\bar{\mathbf{u}}_{\mathbf{g}}, \overline{w}_{g}, \bar{\theta}) + v\, \underbrace{C_{g}(\mathbf{u}'_{\mathbf{g}}, w'_{g}, \bar{\theta})}_{C^{v}_{g}} + \underbrace{C_{g}(\bar{\mathbf{u}}_{\mathbf{g}}, \overline{w}_{g}, \theta')}_{C^{T}_{g}} + \underbrace{C_{g}(\mathbf{u}'_{\mathbf{g}}, w'_{g}, \theta')}_{C^{vT}_{g}}$$



Buckley et al. (subm.)

Need adequate model output

Basic principles for sampling model simulation

Maintain high degree of integrity of output

E.g.:

Products of time-dependent fields should be time averaged as a product, using all model time steps to build the average

Griffies et al. (2009)

INTERNATIONAL COUNCIL FOR SCIENCE

INTERGOVERNMENTAL OCEANOGRAPHIC COMMISSION

WORLD METEOROLOGICAL ORGANIZATION

WORLD CLIMATE RESEARCH PROGRAMME

Sampling Physical Ocean Field in WCRP CMIP5 Simulations: CLIVAR Working Group on Ocean Model Development (WGOMD) Committee on CMIP5 Ocean Model Output

February 2009

ICPO Publication Series No.137

WCRP Informal Report No. 3/2009

Combining the knowledge reservoirs: "data assimilation", "reanalysis"

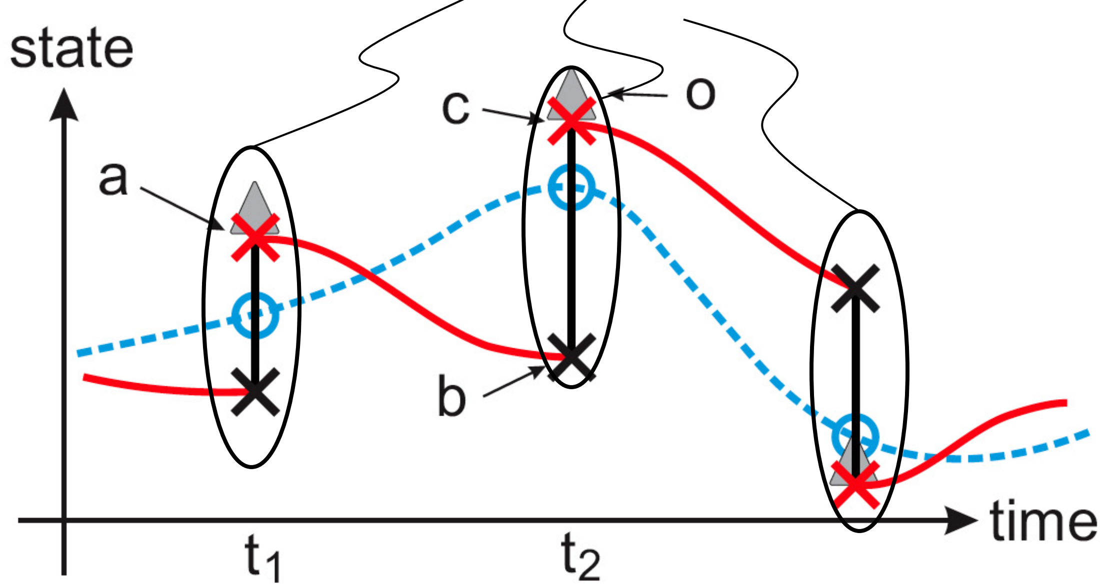
The *estimation (interpolation)* vs. *forecasting (extrapolation)* problem

- **Atmosphere**
  - Relatively abundant data sampling of the 3-dim. atmosphere
  - Most DA applications target optimal forecasting
  - ➔ find initial conditions which produce best possible forecast;
  - ➔ *dynamical consistency or property conservation *NOT* required*
- **Ocean**
  - Very sparse data sampling of the 3-dim. ocean
  - Understanding past & present state of the ocean is a major issue all by itself, the forecasting dependent upon it

"jumps" are unphysical

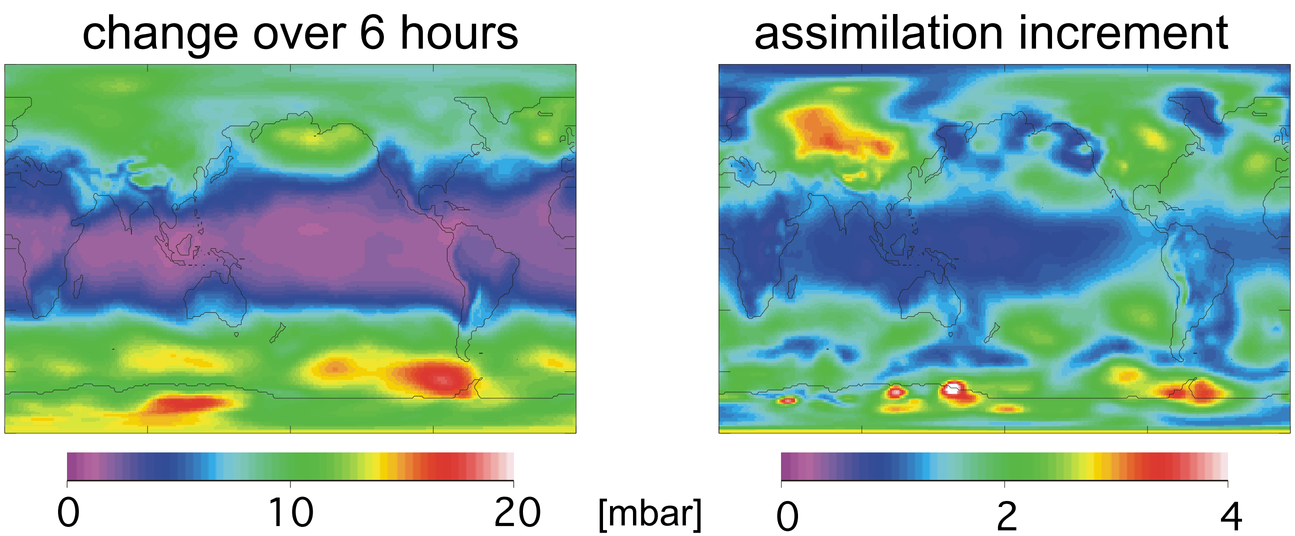
➔ use observations in an optimal way to extract max. information

➔ *dynamic consistency & property conservation *ESSENTIAL** for climate



Dynamical consistency: why does it matter:

(Large) imbalances in air-sea fluxes from atmospheric re-analysis products



Standard deviation of NCEP surface pressure shows that, on average, 24% of its mass change is physically unaccounted for.

reanalysis product	net fresh water imbalance [mm/year] "+" for ocean volume increase		net heat flux imbalance [W/m²] "+" for ocean cooling	
	ocean-only	global	ocean-only	global
NCEP/NCAR-I 1992-2010	159	62	-0.7	-2.2
NCEP/DOE-II (1992-2004)	740	-	-10	-
ERA-Interim (1992-2010)	199	53	-8.5	-6.4
JRA-25 (1992-2009)	202	70	15.3	10.1
CORE-II (1992-2007)	143	58		

Problem statement & definitions

The **state & parameter estimation problem** statement

$$
\begin{aligned}
J =& \sum_{m=0}^{M} \Big[y(t) - \mathbf{E}\tilde{x}(t) \Big]^T \mathbf{R}^{-1}(t) \Big[y(t) - \mathbf{E}\tilde{x}(t) \Big] \\
&+ \sum_{m=0}^{M-1} \tilde{u}(t)^T \mathbf{Q}^{-1} \tilde{u}(t), \qquad t = m\,\Delta t
\end{aligned}
$$

differs from the **data assimilation problem**, as stated in NWP (a.k.a. **analysis** or **re-analysis**):

$$
\begin{aligned}
J_1 =& \Big[\tilde{x}(t_{now}) - \tilde{x}(t_{now}, -) \Big]^T \mathbf{P}^{-1}(t_{now}, -) \Big[\tilde{x}(t_{now}) - \tilde{x}(t_{now}, -) \Big] \\
&+ \Big[y(t_{now}) - \mathbf{E}(t_{now}) x(t_{now}) \Big]^T \mathbf{R}^{-1}(t_{now}) \Big[y(t_{now}) - \mathbf{E}(t_{now}) x(t_{now}) \Big]
\end{aligned}
$$

Problem statement & definitions

- Prior or background state estimate $\tilde{x}(t_{now}, -)$ obtained by running (in forecast mode) model over time τ
- New information/observations $y(t_{now})$ arrive at timt t_{now}

$$\mathbf{E}(t_{now})\, x(t_{now}) \,+\, n(t_{now}) \,=\, y(t_{now}),$$

$E(t_{now})$ observation matrix,
$n(t_{now})$ observation/representation error ("noise")

Obtain best linear unbiased estimator (BLUE), called **the analysis** (e.g., Kalman Filter, Optimal Interpolation):

$$\tilde{x}(t_{now}) \,=\, \tilde{x}(t_{now}, -) + \mathbf{P}^b \mathbf{H}^T \left[\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R}\right]^{-1} \Big(y(t_{now}) - \mathbf{E}x(t_{now}, -)\Big)$$

Problem statement & definitions

$$\tilde{x}(t_{now}) = \tilde{x}(t_{now}, -) + \mathbf{P}^b \mathbf{H}^T \left[\mathbf{H} \mathbf{P}^b \mathbf{H}^T + \mathbf{R} \right]^{-1} \Big(y(t_{now}) - \mathbf{E}\, x(t_{now}, -) \Big)$$
$$= \tilde{x}(t_{now}, -) + \mathbf{K} \cdot \Big(y(t_{now}) - \mathbf{E}\, x(t_{now}, -) \Big)$$

- K: **gain matrix**
- $\tilde{y}(t_{now}) = y(t_{now}) - \mathbf{E}\, x(t_{now}, -)$: **innovation vector** or **residuals**: new information contained in the observations

Interpretation:
Gain matrix **K** *weighs* the innovation vector $\tilde{y}(t_{now})$
according to observational & prior uncertainties
to produce the analysis $\tilde{x}(t_{now})$

Problem statement & definitions

For example (limiting cases):

- large obs. error $\longrightarrow$ small weights $\longrightarrow \tilde{x}(t_{now}) \approx x(t_{now}, -)$
- small obs. error $\longrightarrow$ large weights $\longrightarrow \tilde{x}(t_{now}) \neq x(t_{now}, -)$

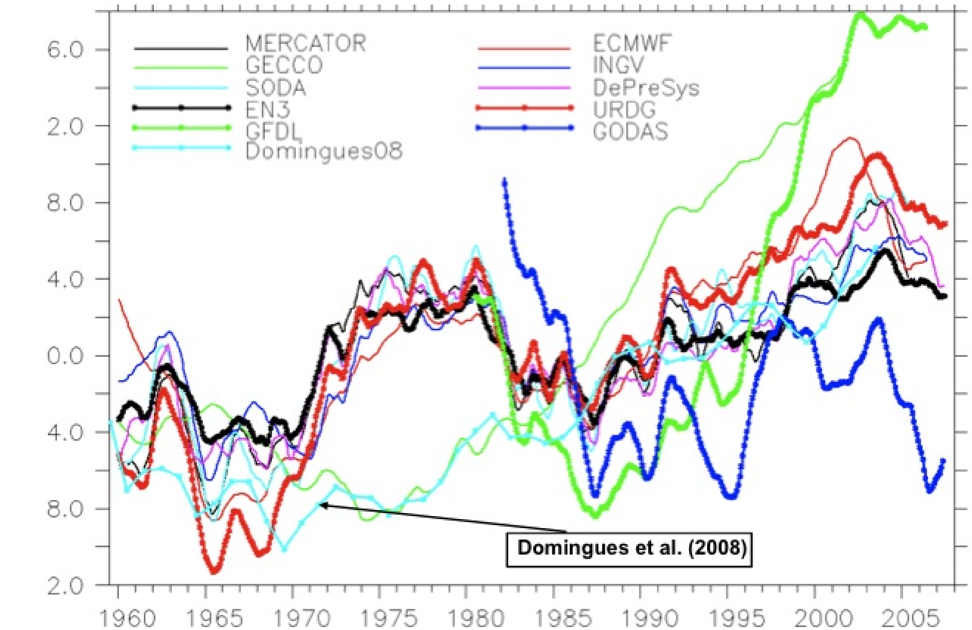
Key point to understand: The analysis step

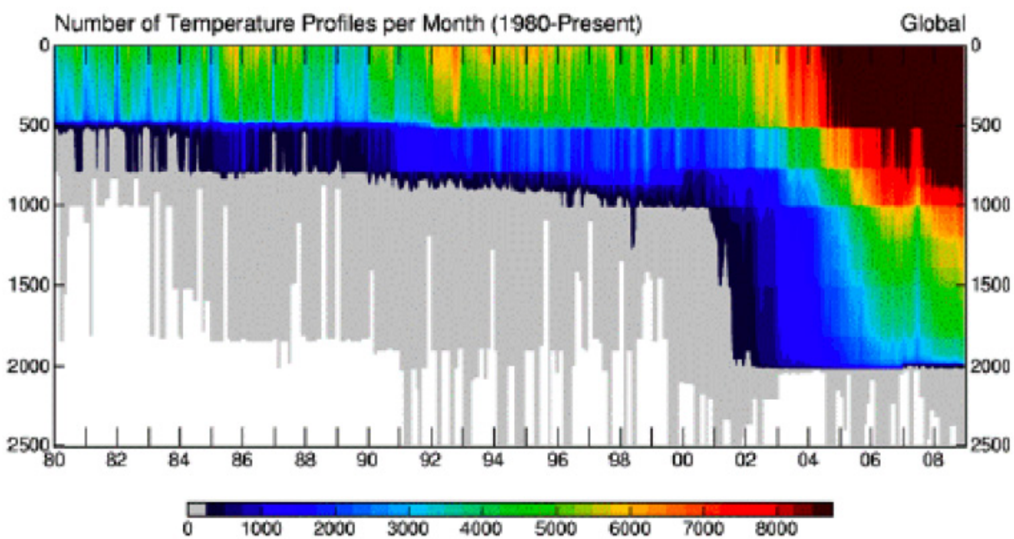
$$x(t_{now}, -) \quad \longrightarrow \quad x(t_{now}) = x(t_{now}, -) + \mathbf{K} \cdot \Big(y - \mathbf{E}\, x \Big)$$

violates the conservation equations for tracers and momentum!

Challenges (e.g. summarized in several OceanObs' 09 whitepapers)

Comparison of different multi-decadal "re-analyses" (CLIVAR/GSOP)



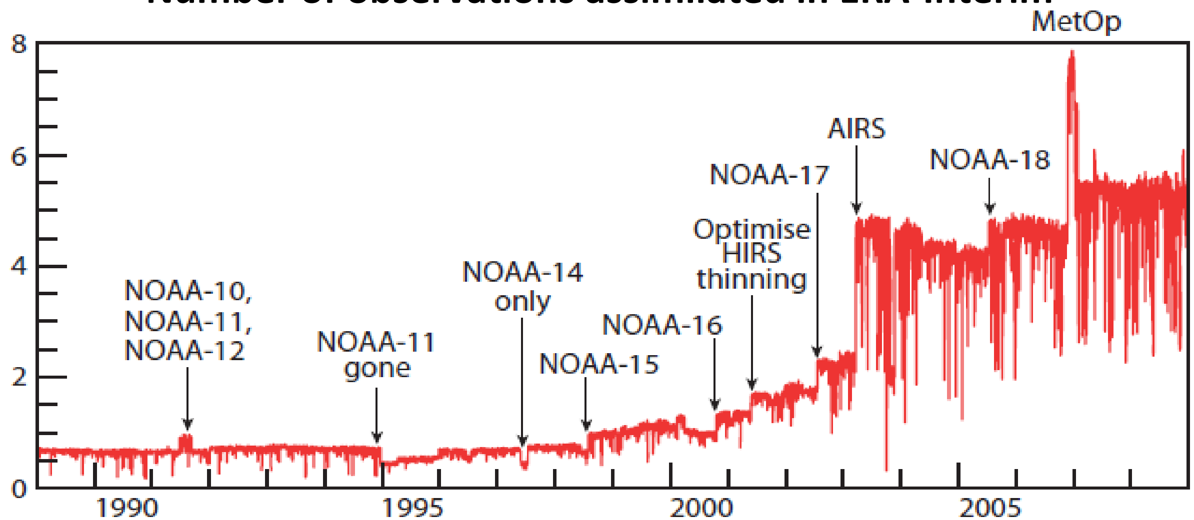


- Convergence of estimates with increasing number of observations?
- If not, why not?
- Consequences for attempt to understand circulation changes…
- Narrowing of transport uncertainties to "*useful*" numbers remains hard

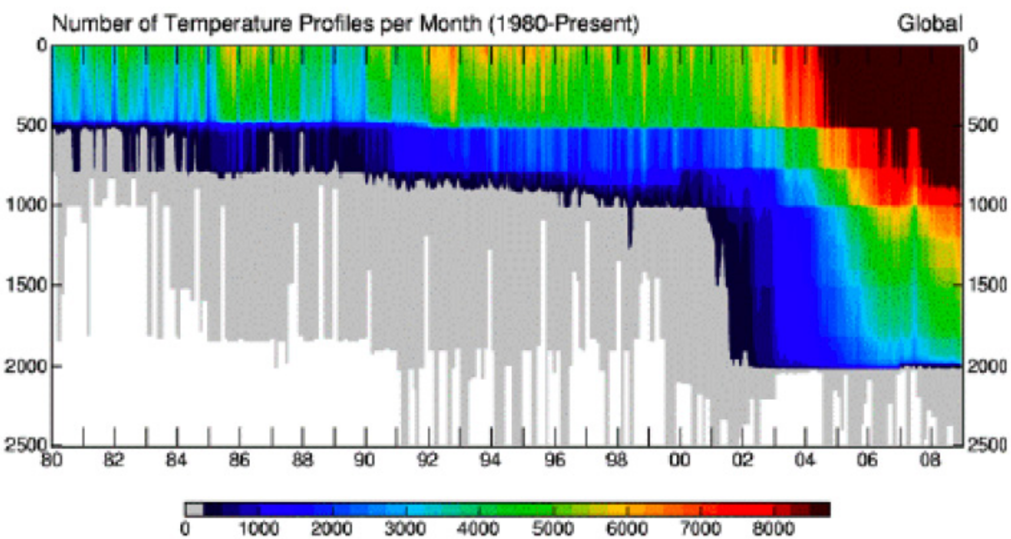
Recent(!) changes in the atmosphere and ocean observing systems

An important issue is that changes in the global observing system aren't just a thing of the past (pre-1980s), but remain relevant today.

Number of observations assimilated in ERA-Interim



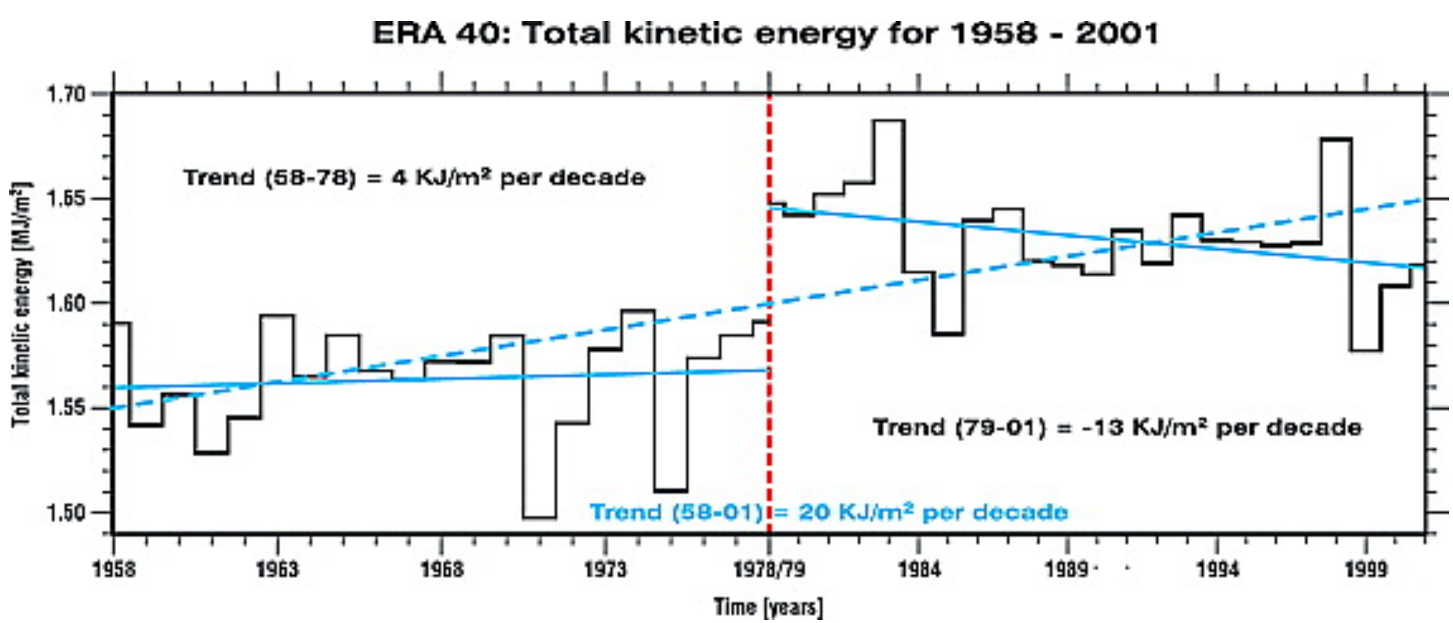
[Dee et al., 2009, *ECMWF Newsletter* (119)]



Lessons from atmospheric reanalysis:
"Can climate trends be calculated from reanalysis data?"

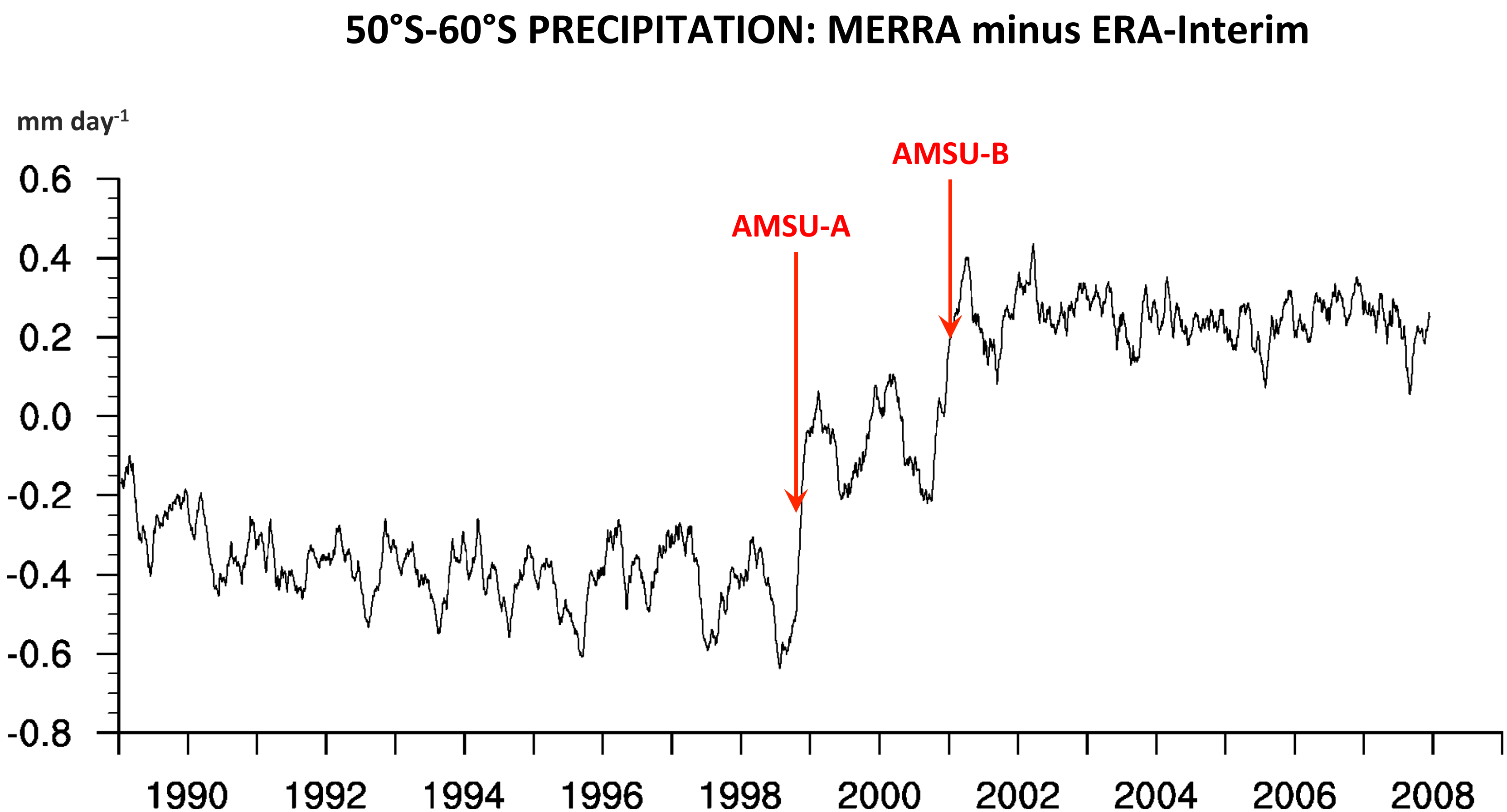
- Large warming trend in ERA40 an artifact of large changes in observational coverage at the end of the 1970s
- Large uncertainty in the calculation of trends from present reanalyses
- Present observing system was set up to support weather forecasting, not directly suitable for climate monitoring
- Systematic errors in the assimilating models add complications
- Limited resources currently devoted to address these problems!!

Why should the ocean estimation problem be any different?



Bengtsson et al. (2004)

Spurious trends in MERRA precipitation

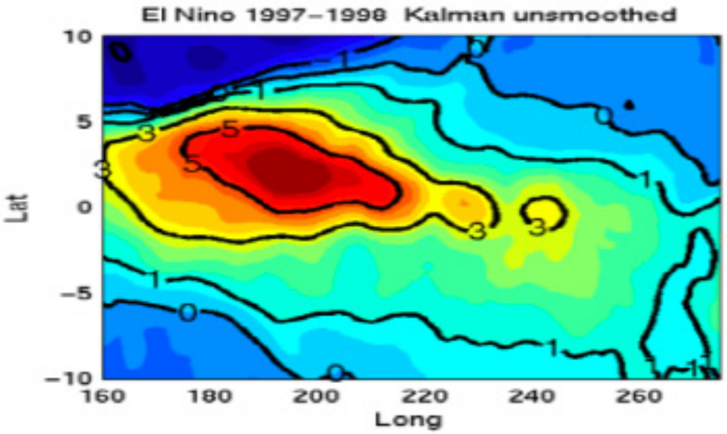


2-month running average difference between forecast daily precipitation from MERRA and from ERA-Interim, spatially averaged over the 50°S-60°S latitude band.

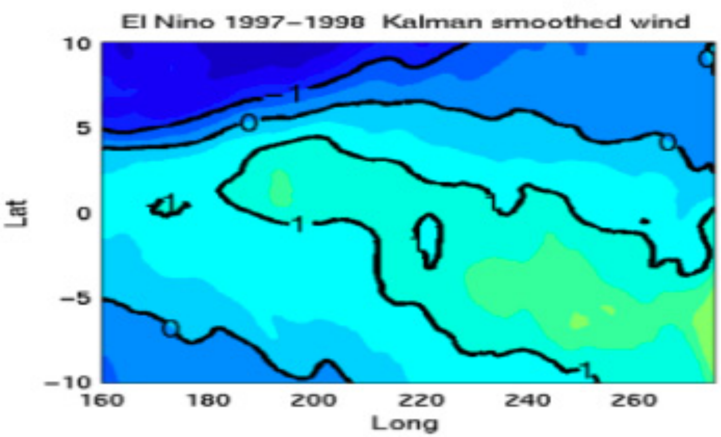
Dynamical consistency: why does it matter:

Driving offline tracer simulations (e.g. CO_2)

Filtered estimate of CO_2 flux during 97-98 El Niño (mol/m^2/yr)

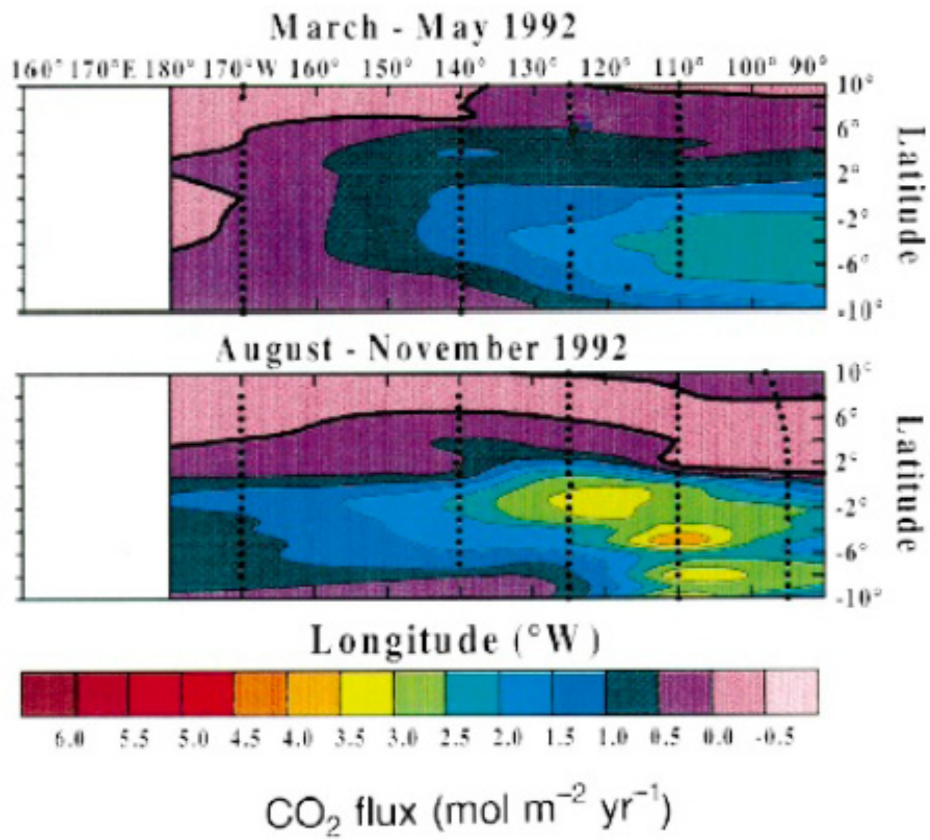


Smoothed estimate of CO_2 flux



McKinley, 2002

Observed estimate of CO_2 flux during 92-93 El Niño



Feely et al., 1999

Conclusions

- *sequential/filtering* methods solve not the same problem as *estimation/smoother* methods
  - break dynamical consistency
  - tracer conservation no longer fulfilled
  - NWP is not the same thing as "reconstruction"

- for climate science applications need to account for and understand the analysis increments

- if needed, at a minimum make analysis step / innovation vector part of the basic diagnostics and budget calculations
  - also has value to diagnose systematic model deficiencies