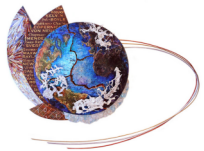
Comparing EnKF and 4DVAR:

OSSE and metrics

obs count · error model · cost function · posterior σ

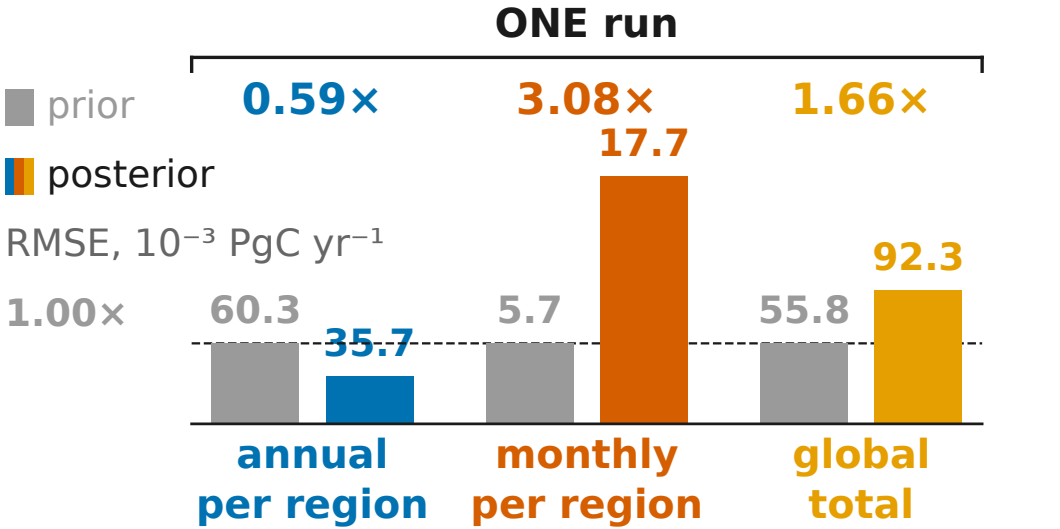
Ash Pera[1,2] · Lori Bruhwiler[2] · Youmi Oh[1,2] · Andy Jacobson[1,2] · John Miller[2]
[1] CIRES, University of Colorado Boulder [2] NOAA Global Monitoring Laboratory



Overview

- OSSEs: plant a truth, try to recover it
- Four metrics we all report
- Each one measured against a known answer
- What each is really sensitive to
  - observation count · error-model fit · cost-function decrease · posterior σ
  - CH_4 analytic inversion · CO_2 commercial aircraft · CH_4 twin vs TM5-4DVAR · $\Delta^{14}C$ fossil

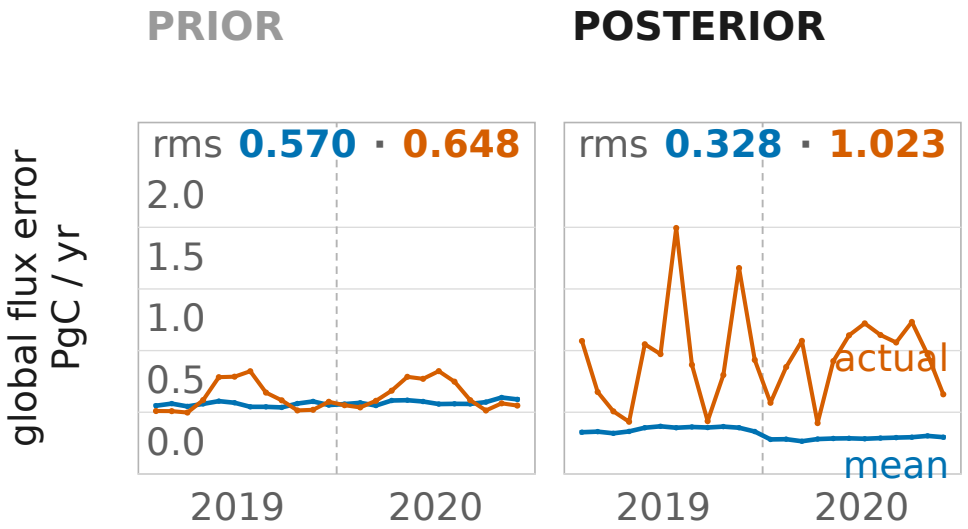
One solve, three RMSEs



each pair is scaled to its own prior — do not read across

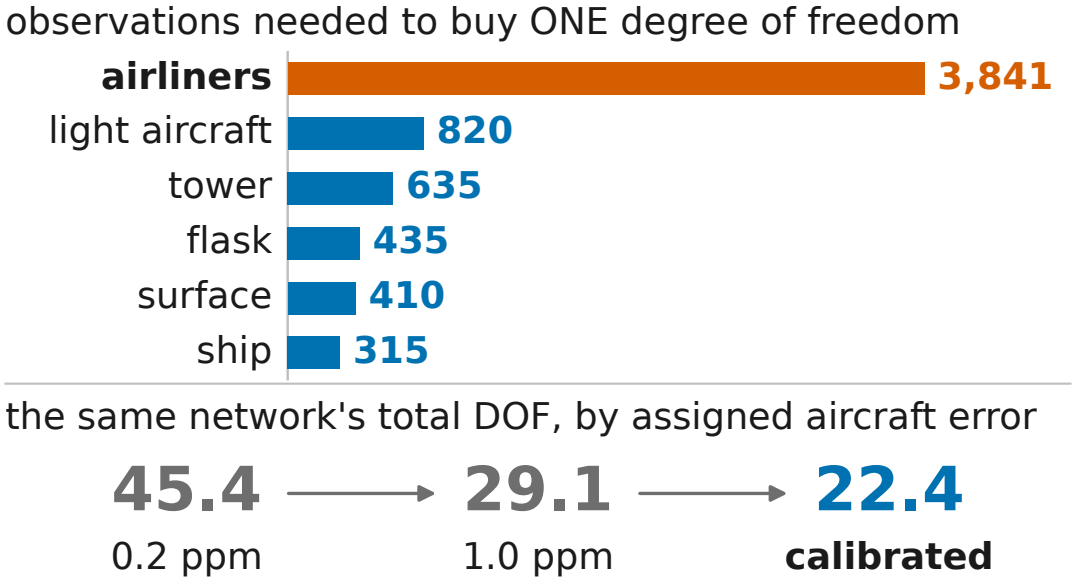
- CH_4 analytic inversion — 240 regions, 24 months
- RMSE = $\sqrt{\langle(\text{posterior} - \text{truth})^2\rangle}$ — averaged over what?
- Annual sums months first; global sums regions first

Annual gain, monthly noise



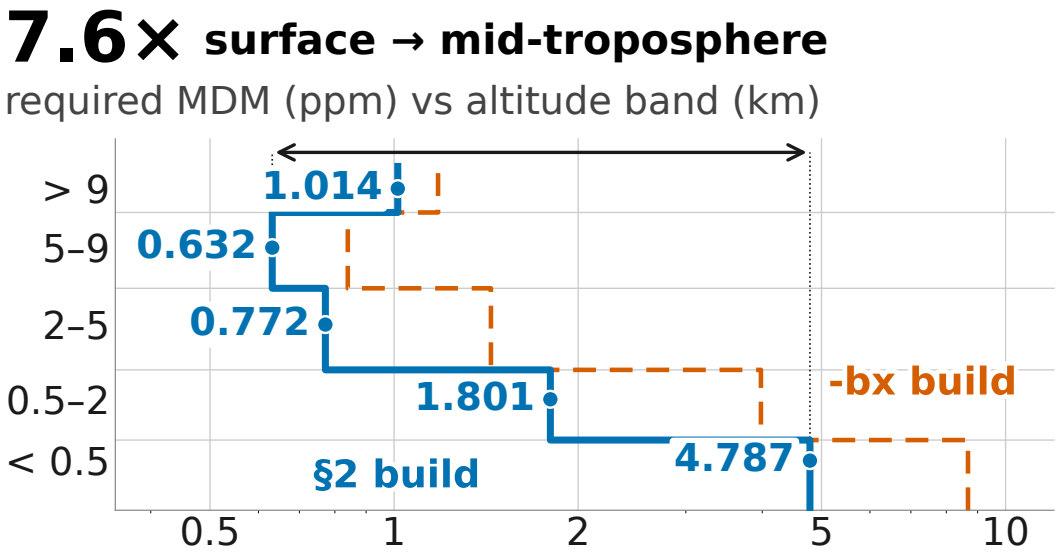
- Annual-mean error −43 %; the actual error ×1.58
- Inversions can require temporal damping
- A poorly behaved, non-production inversion
- Later arm: S_ann +0.429, S_mon −0.509

Observations, not information



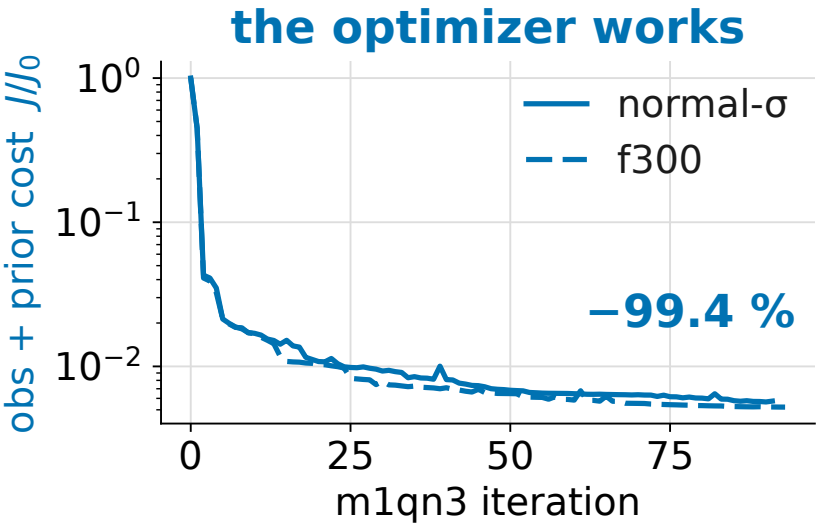
- CO_2 commercial-aircraft OSSE — 600 members, biweekly
- 28,892 aircraft obs buy 7.5 of 22.4 network DoF
- 3,841 observations per degree of freedom

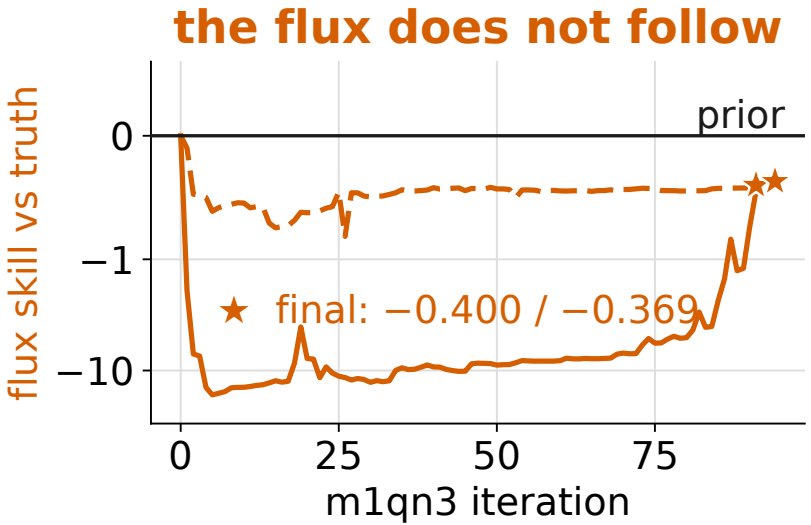
Error models have shape



- The same network scores 45, 29 or 22 DoF
- Radiocarbon: forecast the miss, then measured it
- Your QC sits inside your diagnostic

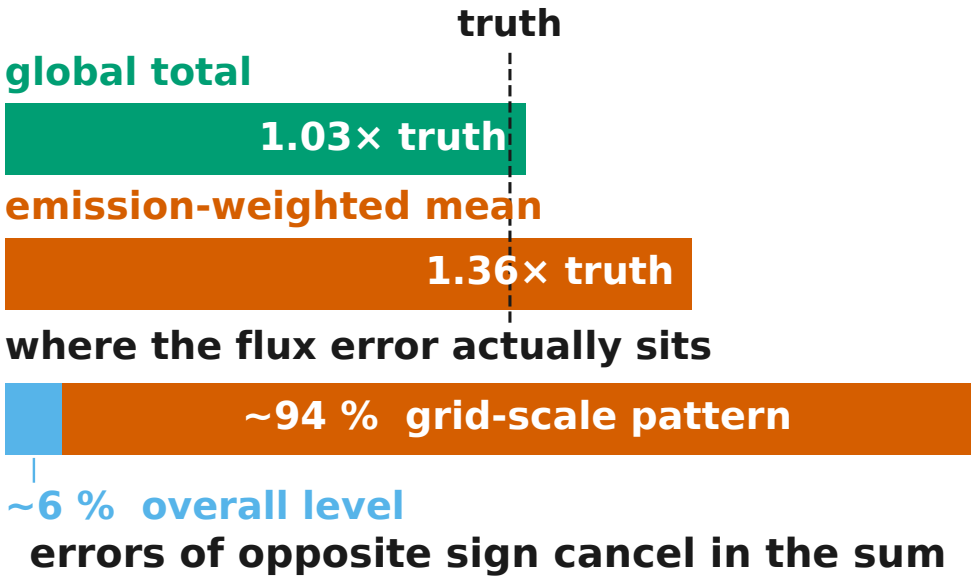
Fits beautifully. Wrong answer.





- TM5-4DVAR on our truth and pseudo-obs
- Obs cost fell to almost nothing; flux got worse
- A good fit is not a good answer

Why totals look fine



- Global total lands within a few percent
- Emission-weighted mean runs a third high
- A total can be right for wrong reasons

Sinks are not optimised

the same run, two sink packages

CH_4-only GCP 4DVAR

Wang Cl · TransCom OH · scale 1.0

$\delta^{13}C$ production

Hossaini Cl · Spivakovsky OH · scale 0.901

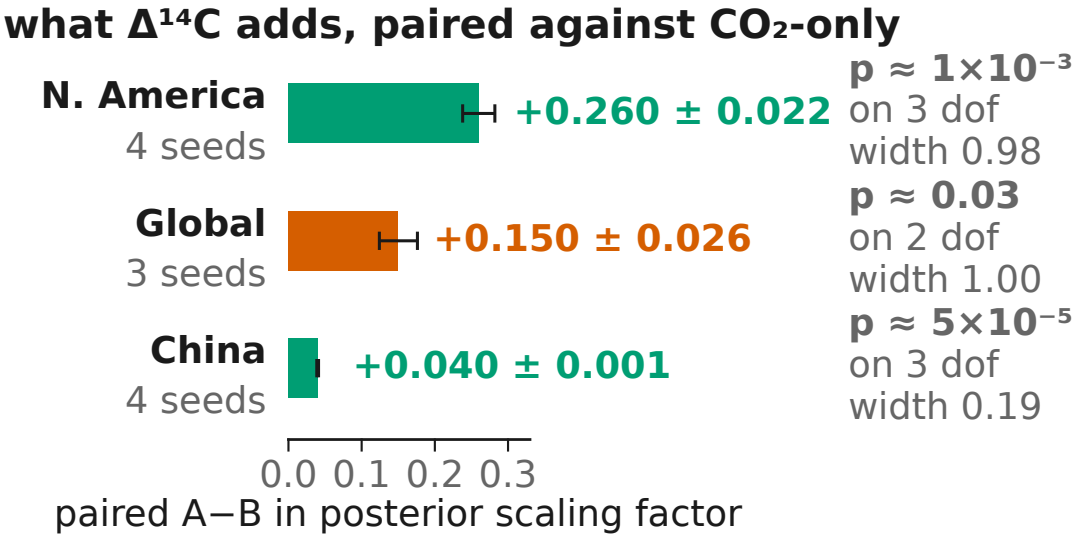
↓

the emissions absorb it

sinks are fixed inputs — optimize.parameter : F

- Two lineages, two sink packages
- The chlorine is the small term; the OH field and the scale are not
- Sinks are not in the state vector

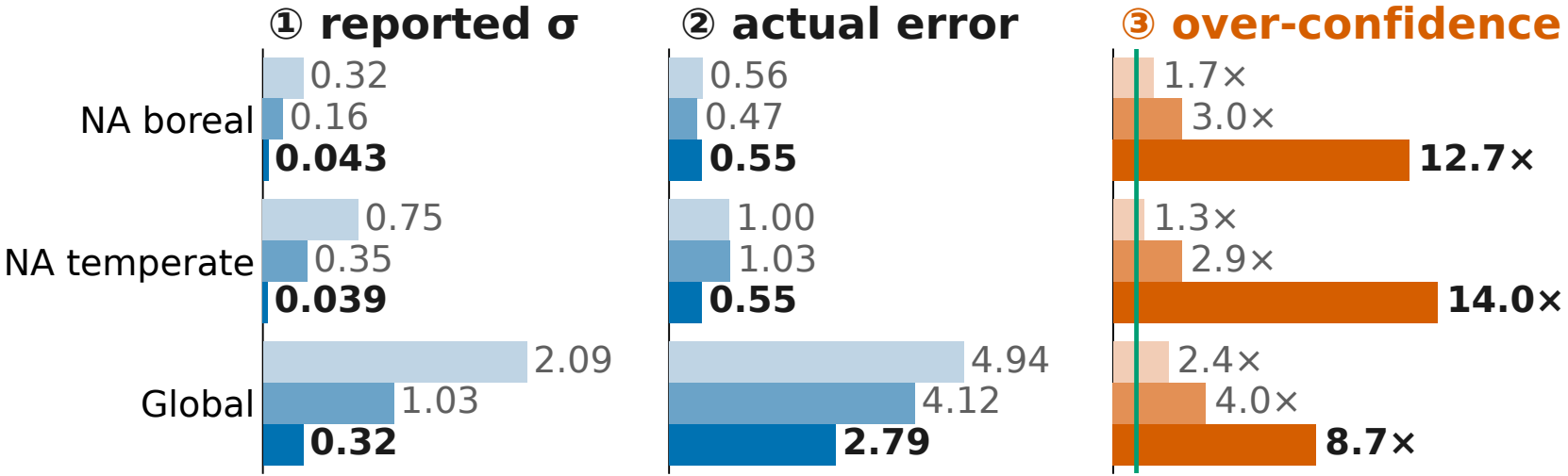
What the isotope adds



Global: the flagship and the weakest — 3 seeds, 5 %

- Paired against CO_2-only, same truth, same seeds
- Strength as p on its own dof — never a bare σ
- The flagship row is the weakest: 3 seeds, 5 %

Uncertainty reduction



prior → no aircraft → + aircraft · PgC yr^{-1} on the regional total · 1× = honest

- Back to the CO_2 aircraft OSSE — the same 600-member chains
- Global: reported σ falls 6.5×, the error it makes falls 1.8×
- The gap grows as observations are added

Uncertainty nobody checks

the σ we ship	analysis expects	the truth shows
CH_4 EnKF twin the ensemble σ	**0.17 % outside** **$\chi^2_{flux} = 1$**	**12.7 % outside** **χ^2_{flux} 6.49**
CH_4 twin, no-loc the ensemble σ	**$\sigma_{ens}/\sigma_{exact} = 1$** **$\chi^2_{flux} = 1$**	**1.09×** ✓ **χ^2_{flux} 8.56** ✗
^{14}C — real data ensemble + its R	**outside 0.0132** **mean rank 75.0**	no flux truth **obs: 0.0121** ✓ **rank 35.8** ✗
CH_4 budget σ 1.01 Tg yr^{-1}	**≥ 2.61 Tg yr^{-1}** Desroziers	**nothing to check**

- Twin: 12.7 % outside, 0.17 % expected
- The exact σ agrees — and still fails
- Real data: nothing to check it against

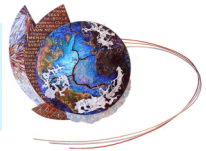
What to report instead

you report	it is really sensitive to	report instead
observation count	redundancy you never measured	degrees of freedom — ours cost 3,841 obs each
χ^2_obs	the obs you kept — your QC sits inside it	pooled, all obs — and name the stream
cost-function decrease	the prior and the error model — never the truth	the pattern, against a planted truth
posterior σ	the prior you chose — 21× over the σ_B sweep	reduction vs an external anchor, never the level

never quote a recovery % without saying what the denominator is.

Conclusions

- Every metric moved without the answer moving
- Aggregation, error model, and prior set the number
- An OSSE is where you can catch it
  - Never quote a recovery % without its denominator



Backup

Backup — recovery axes

six axes MEASURED to move the same reported number — one truth, one network, one window

recovery %
the number you report

- **prior width σ_B (R fixed)**
  $-0\ \%\ \longrightarrow\ 172.9\ \%$ across $\sigma_B \in [0.1,\ 4]$
- **in-region obs coverage**
  INVERSELY: 2.7 % cov → 66.3 %, 99.6 % → 7.9 %
- **truth construction** *twin vs inventory-forward*
  ≈ 3–10×, coverage-partitioned; decisive cell aborted
- **posterior-file alignment**
  one indexing convention moved EVERY number
- **observation integrity**
  a corrupted pseudo-obs set voids every score built on it
- **where in the window you read**
  the SAME run scores 33.8 % or 66.3 %

…and NOT to the truth it claims to score

planted signal ×1

15.4 %

5× LESS planted signal

15.6 %

ZERO overlap, UNMATCHED
width 0.22 vs plant's 0.98

15.8 %

ZERO overlap, MATCHED
matched in cells and fossil weight

the metric cannot tell these apart

−0.6 % *the scorer CAN return zero*

never quote a recovery % without saying what the denominator is

report instead: a pre-registered paired contrast, equally covered, with a null arm.