Has the data outstripped the models?

Or:

What can possibly go wrong?

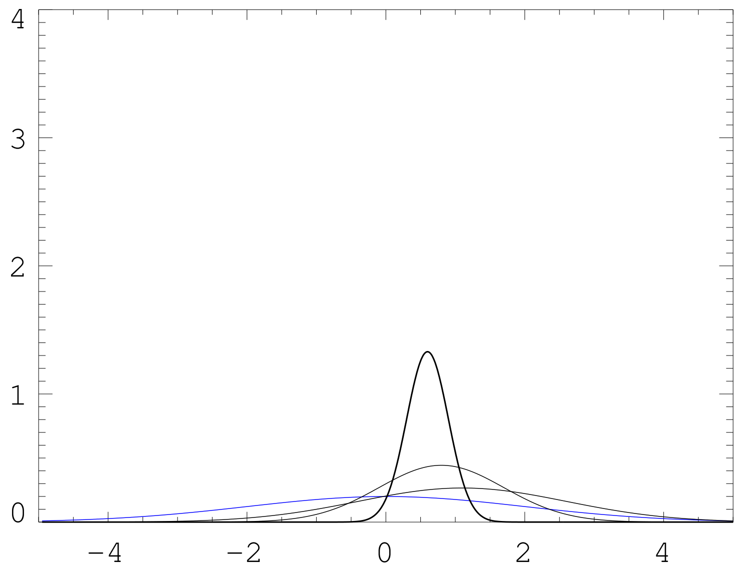
Recap from yesterday

- Data assimilation is an example of Bayesian Inference;
- BI itself follows from rules for combining PDFs;
- Techniques like least squares minimisation are special cases for particular types of PDF
- Most approaches such as Kalman Filtering and 4dVar can be expressed with this formalism.

Outline

- What *should* happen;
- The black triangle revisited;
- Things to watch for;
- Some real world cases;
- What can we do?

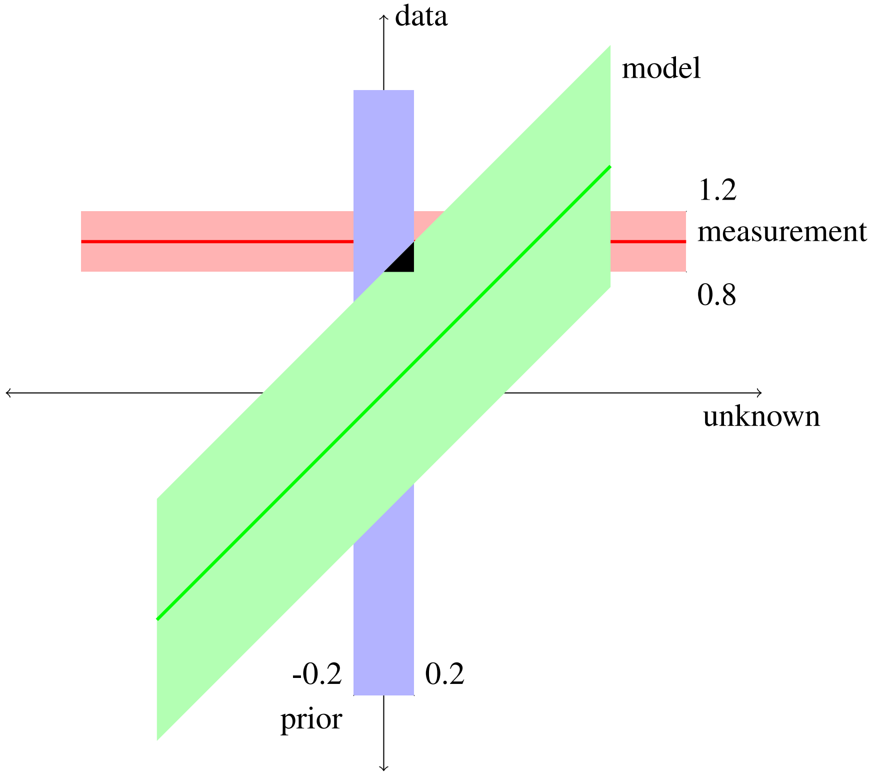
What *should* happen



PDFs for parameter with prior (blue) and after 1, 2 and 3 observations.

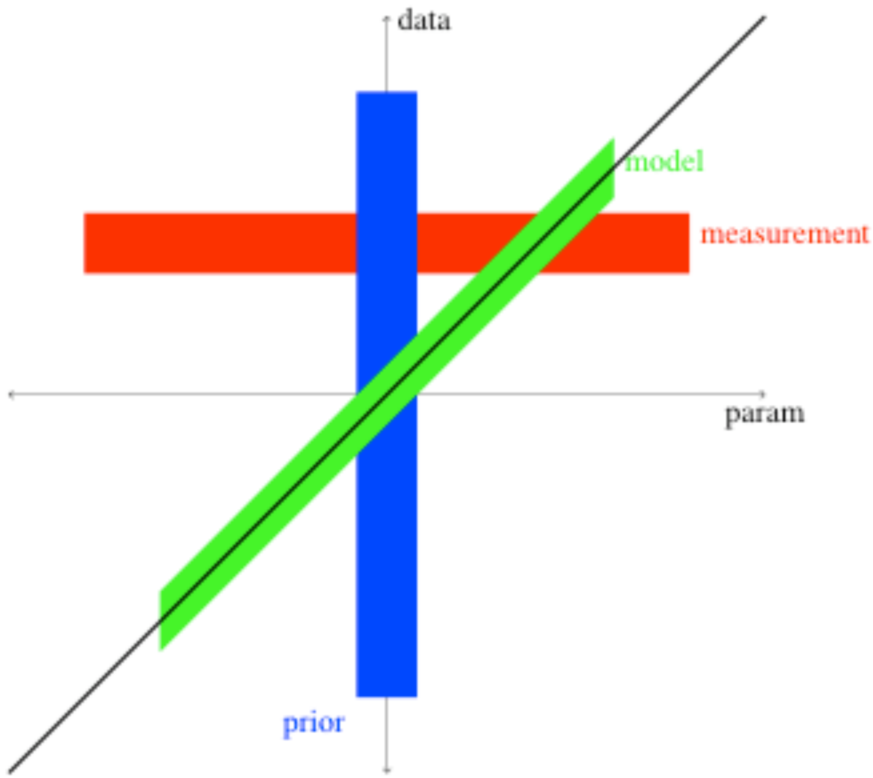
- Prior broad distribution hence weak constraint;
- Each observation refines the estimate (sharper peak);
- Each estimate is consistent with the previous ones;
- Final estimate casts doubt on prior estimate *but not the PDF*.

The black triangle revisited



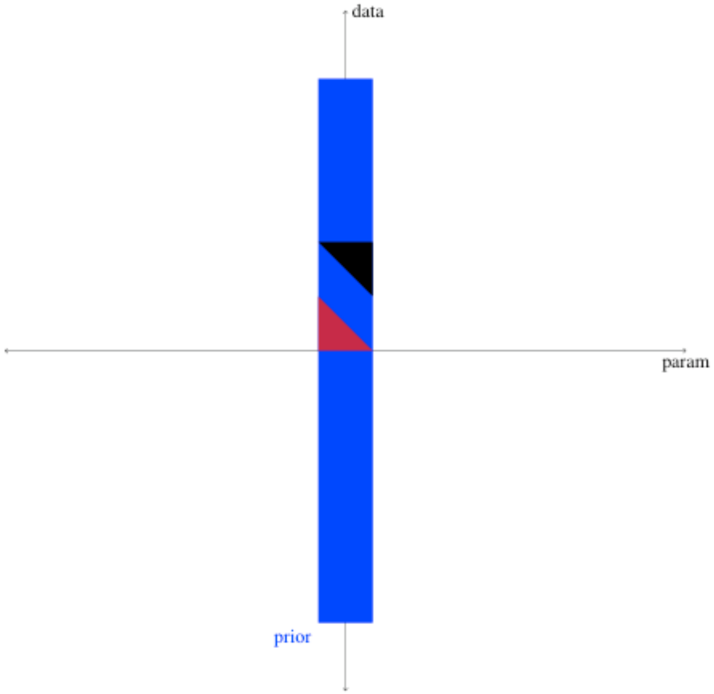
- unknown on X-axis, obs on Y-axis;
- Light-blue = prior unknown;
- Light-red = obs;
- Green = model;
- Black = solution.

A problematic Case



- No overlap means no solution;
- Fundamental problem is at least one PDF is wrong;
- With a Gaussian we always get a solution but sometimes of very low probability;
- How can we tell?

Another problematic Case



- Parameter constrained by measurements of two different quantities;
- Either measurement alone is consistent with prior;
- 2 measurements + prior + model has no solution.

What to do?

- Look hard at each of the 3 input PDFs;
- Check the assumed PDF against the sample generated in the inversion;
- Check with independent data (often called cross-validation);
- A lot of examples.

First check the Priors

- Approach will differ depending on the unknowns;
- Sometimes you calculate actual PDFs, sometimes you use algorithms;
- Cases like weather prediction you can test these rules every day.

Example from Flux Inversions

- Chevallier et al. GRL, 2006;
- Compare ORCHIDEE at 50km resolution to CO_2 flux measurements;
- STD-dev of differences $\approx$ 2.5 respiration;
- No spatial correlations in error;
- Temporal correlations of about 1 month.

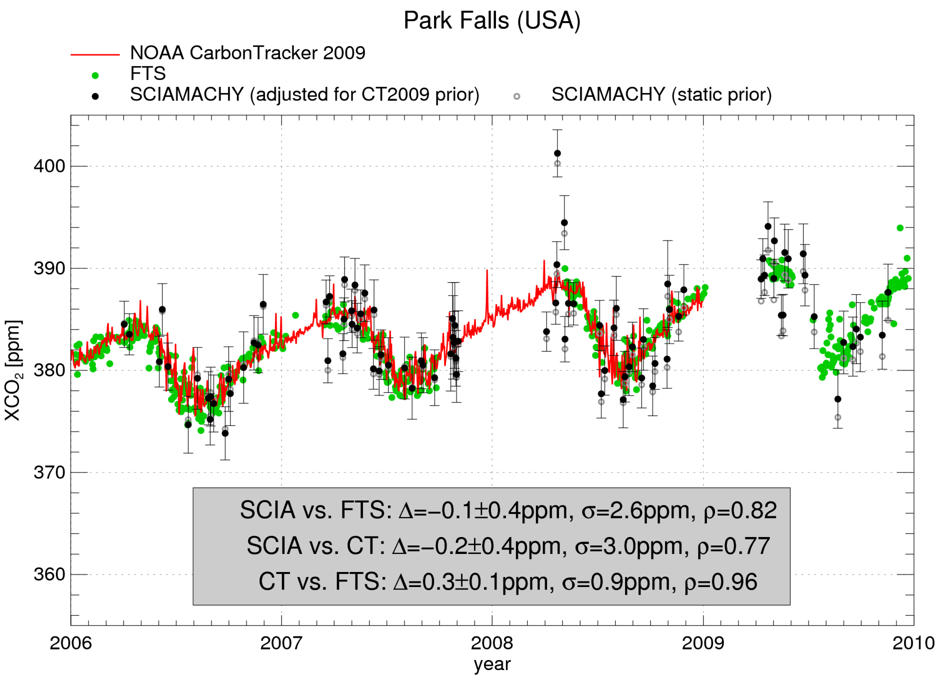
The Measurement PDF

- PDF is that of the *true* value;
- Known errors (often called biases) must be removed first;
- This does *not* say there are no mean errors left, just that we don't know what they are.

Get to Know your Measurements

- Many “measurements” are themselves products of a model;
- Worry much more about the systematics than the noise;
- Systematics can be treated as correlated errors;
- Small signals on long records are the hardest things we do;
- Independent data is precious.

Comparison of SCIAMACHY and TCCON



Comparison of SCIAMACHY and ground-based spectrometer measurements of CO_2 at Park Falls Wisconsin

- Reuter et al., 2010, (in prep.);
- SCIAMACHY satellite on board ESA ENVISAT;
- Ground-based Solar Fourier Transform Spectrometer part of Total Carbon Column Observing Network (TCCON);
- Random and systematic errors but data are approaching usable.

Validating TCCON

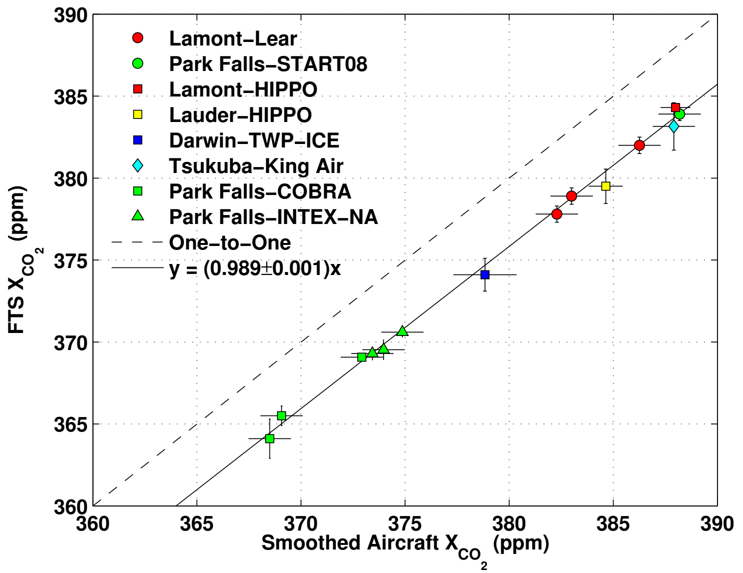


Fig. 4. The TCCON calibration curve for CO_2. The smoothed aircraft value is $\hat{c}_s$ from Eq. (7).

Comparison of TCCON measurements with simultaneous aircraft profiles

- D.Wunch et. al., (2010) discussion paper, Atmos. Meas.Tech.
- Aircraft measurements traceable to primary standards;
- Very good correlation but not one-one;
- Can correct but contributes to overall error.

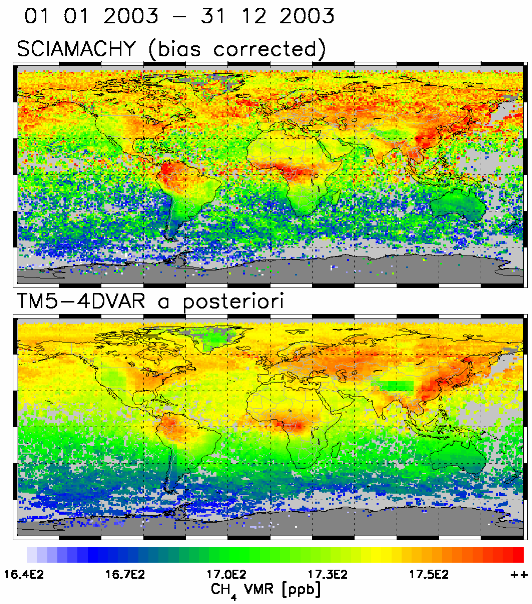
Including measurement errors in Model

- Often you don't have independent data;
- Add extra unknowns to account for systematic errors, e.g.

$$q = q^* + \phi(\text{latitude})$$

  to deal with consistent errors in latitude
- This sacrifices some information in q.

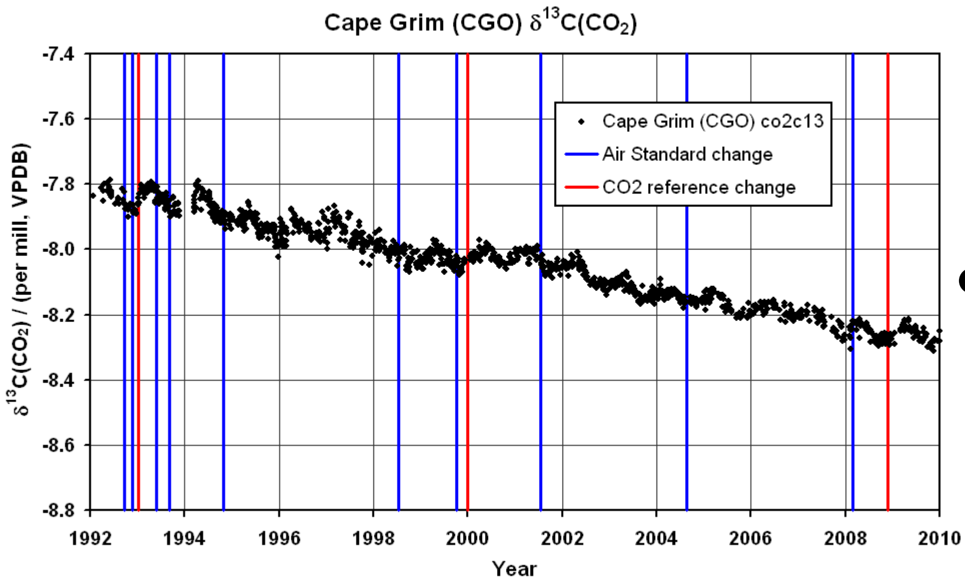
Combining SCIAMACHY and in situ CH_4 data



Annually-averaged, column-integrated methane mixing ratio from SCIAMACHY alone (top) and from SCIAMACHY and surface data assimilated into a single flux inversion (bottom).

- Bergamaschi et al., JGR 2007;
- SCIAMACHY methane from Frankenberg05;
- Uses modelled CO_2 as reference;
- Bias-corrected by simultaneously assimilating surface data into flux inversion.

Accounting for Discontinuities



$\delta^{13}CO_2$ measurements from Cape Grim, Tasmania.

- Two external references necessary for final measurement;
- Great effort made to maintain continuity across changes but never perfect;
- Can be handled either with extra unknowns or correlations.

Checking the Model PDF

- Need PDF of simulated value given *true* value for unknowns;
- Rarely have such cases;
- Use model ensembles as proxy; *risky*;
- More tomorrow.

T1 fossil and biosphere

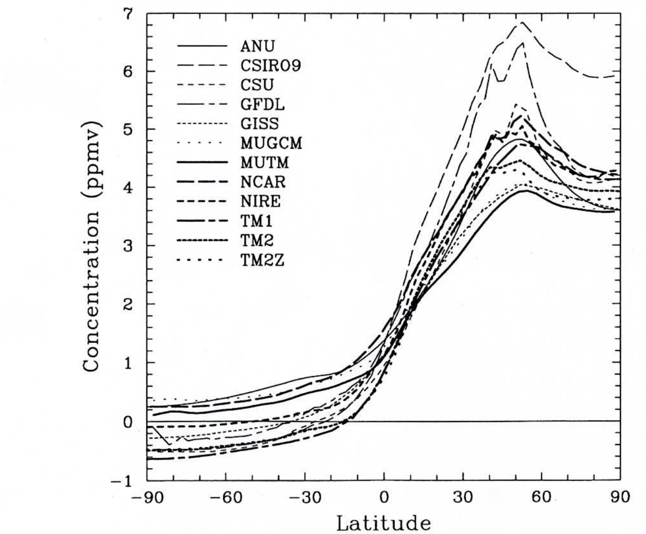


Fig. 3.1: Zonal annual surface mean concentration in ppmv due to fossil emissions.

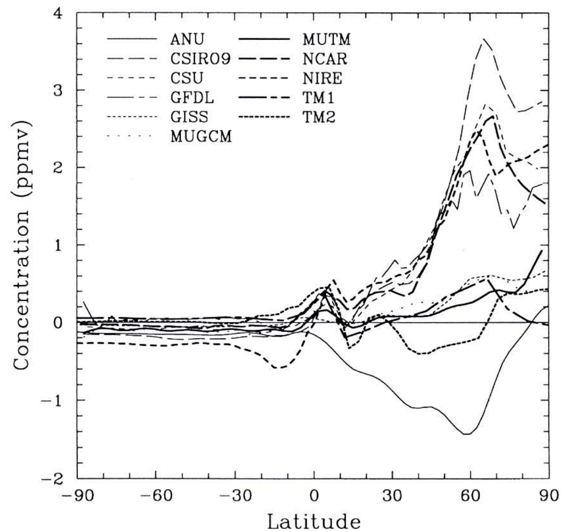


Fig 4.6: Zonal annual mean concentration in ppmv for the bios

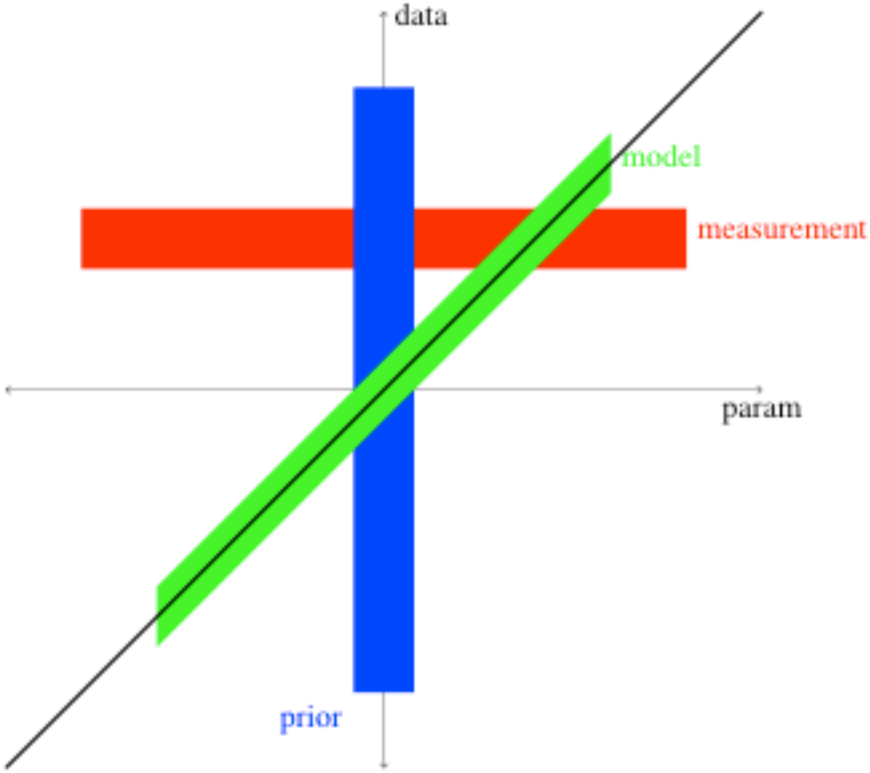
Zonal mean concentration from fossil fuel source

Zonal mean response to annually balanced biosphere source

Checking Posterior PDFs

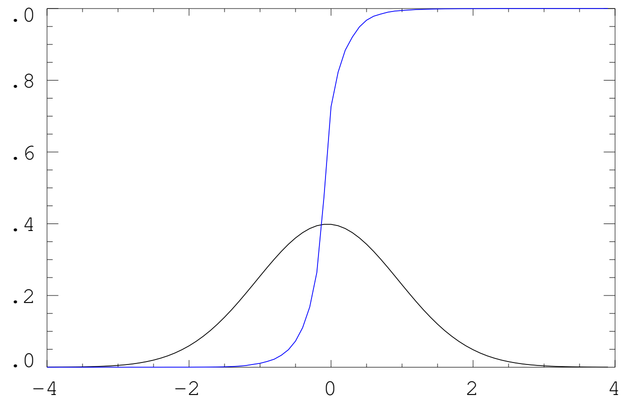
- Basic assumption is that the samples of posterior values for unknowns and simulated observations are drawn from the relevant populations;
  - Posterior $-$ prior $\leftrightarrow$ prior PDF
  - Model $-$ observed $\leftrightarrow$ data PDF
- Must hold for all aspects of the PDF;
- Take note of sample size.

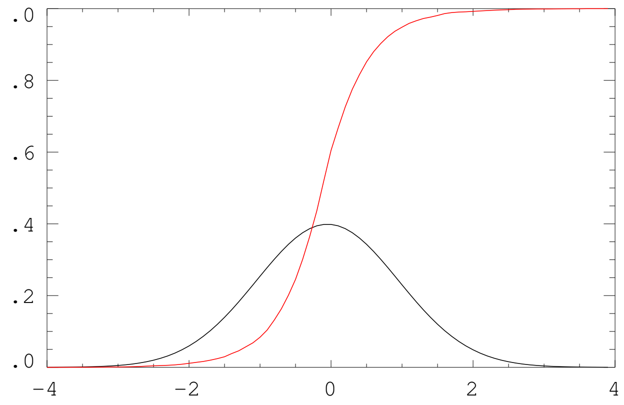
A problematic Case



- No overlap means no solution
- With a Gaussian we always get a solution but sometimes of very low probability;
- How can we tell?
- Fix by increasing uncertainties but which ones and how much?

Plot the Residuals





- Plot of normalised innovations (posterior − prior)/(prior-uncertainty) and normalised residuals (simulation − obs)/(data-uncertainty) for flux inversion;
- Use cumulative frequency rather than raw PDF, easier to look at;
- Compare with standard normal distribution;
- The steep slope corresponds to smaller variance;
- Also numerical tests.

Value of the cost function

Minimise

$$J = (\vec{x} - \vec{x}_0)^T \mathbf{C}^{-1}(\vec{x}_0)(\vec{x} - \vec{x}_0) + (\mathbf{M}\vec{x} - \vec{y})^T \mathbf{C}^{-1}(\vec{y})(\mathbf{M}\vec{x} - \vec{y})$$

Yields

$$\vec{x} = \vec{x}_0 + \mathbf{C}(\vec{x}_0)\mathbf{M}^T \left[\mathbf{M}\mathbf{C}(\vec{x}_0)\mathbf{M}^T + \mathbf{C}(\vec{y})\right]^{-1} (\vec{y} - \mathbf{M}\vec{x}_0)$$

Substituting

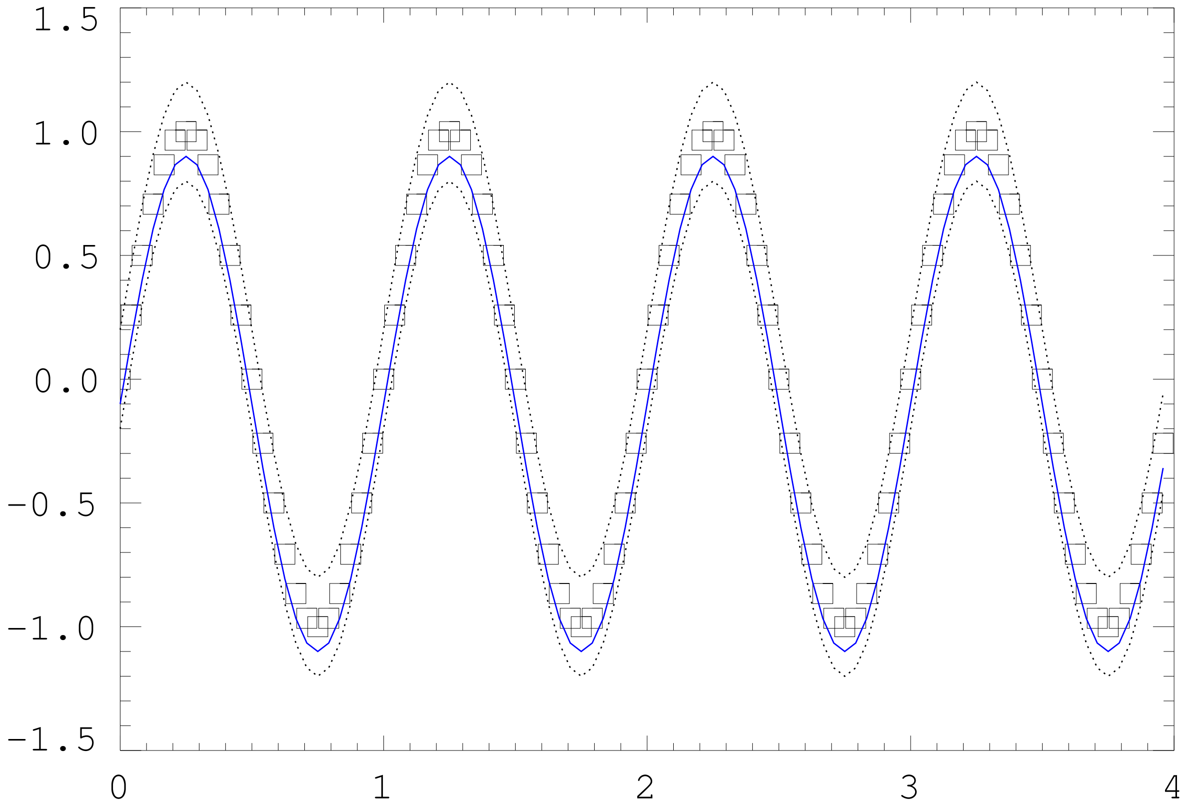
$$J_{MIN} = (\vec{y} - \mathbf{M}\vec{x}_0)^T \left[\mathbf{M}\mathbf{C}(\vec{x}_0)\mathbf{M}^T + \mathbf{C}(\vec{y})\right]^{-1} (\vec{y} - \mathbf{M}\vec{x}_0)$$

Properties

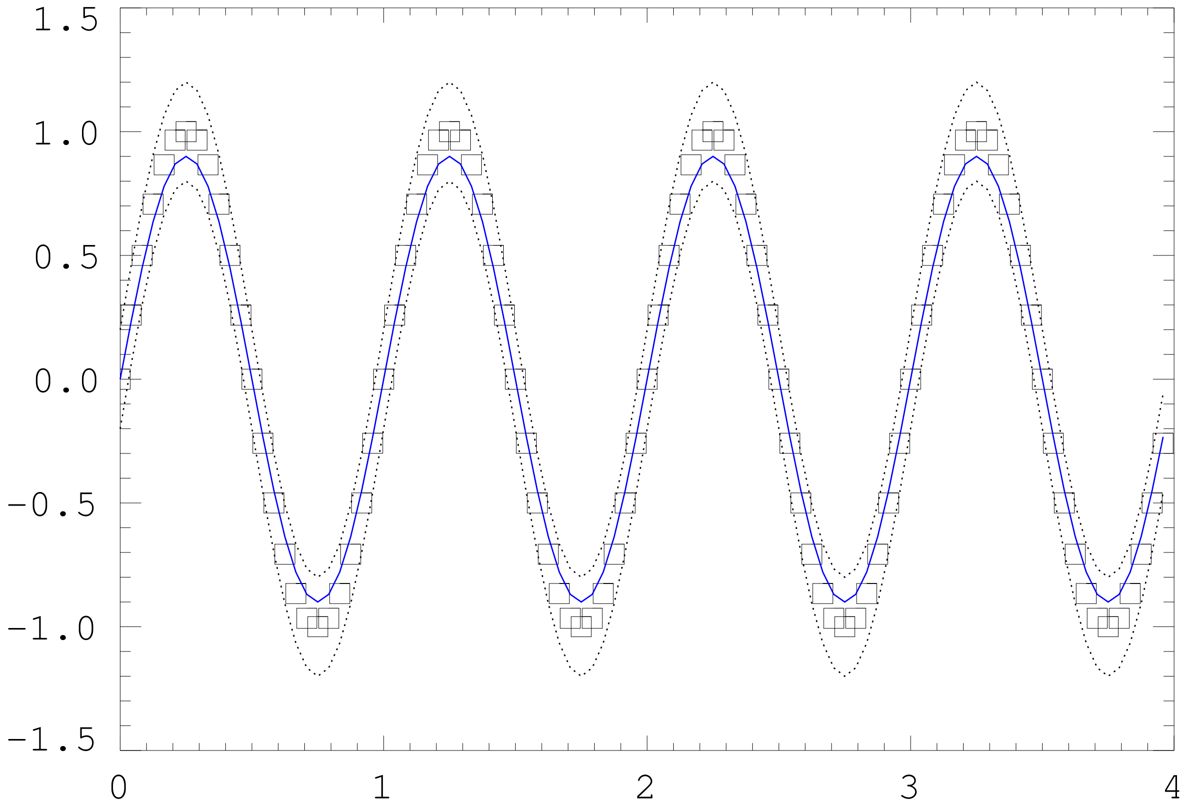
$$J_{MIN} = (\vec{y} - \mathbf{M}\vec{x}_0)^T \left[\mathbf{MC}(\vec{x}_0)\mathbf{M}^T + \mathbf{C}(\vec{y})\right]^{-1} (\vec{y} - \mathbf{M}\vec{x}_0)$$

- Numerator difference between obs and prior simulation;
- Denominator uncertainty in that quantity;
- Should be consistent $J_{MIN} \approx N_{OBS}$, if not, posterior uncertainty inconsistent with inputs;
- Michalak et al., JGR, 2005 has algorithm for scaling uncertainty.

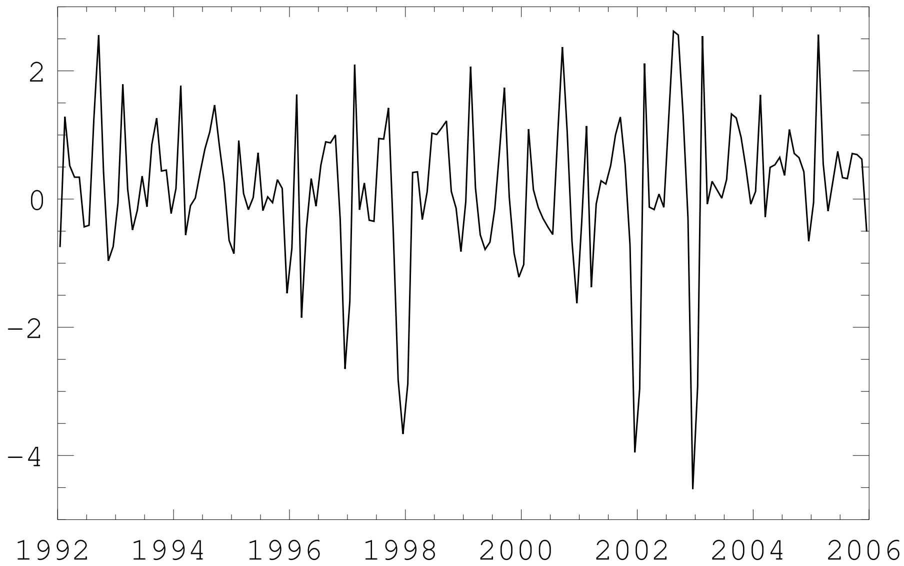
Example of Bias



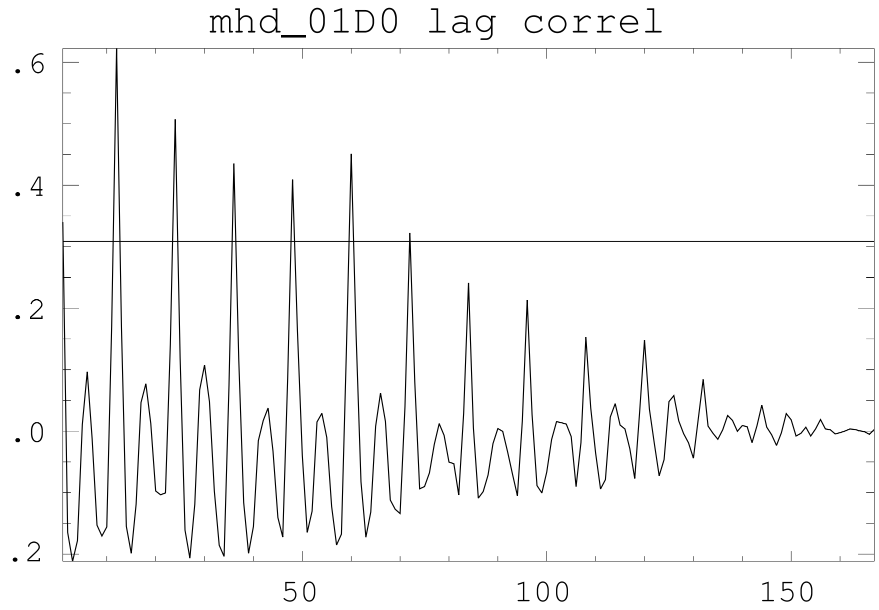
Amplitude Example



Residuals and their correlations



Residual (Model − observed) CO_2 concentration for Mace Head, Ireland

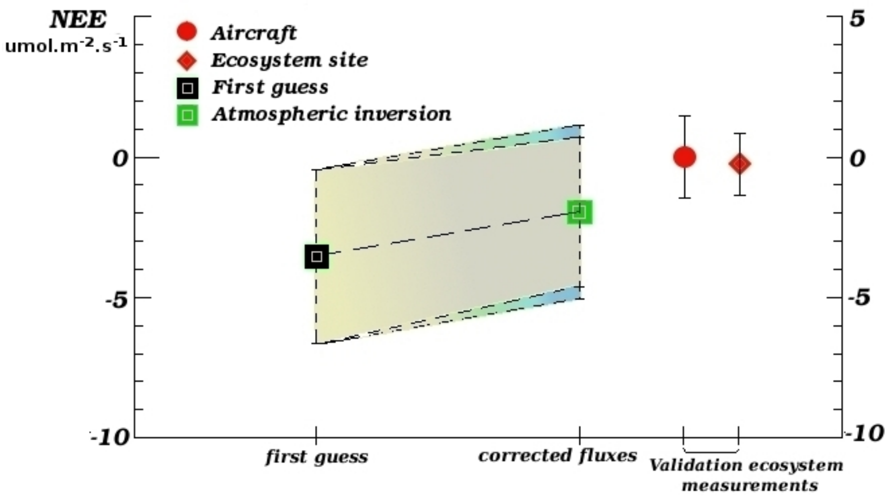


Lagged correlation of residuals

Cross Validation

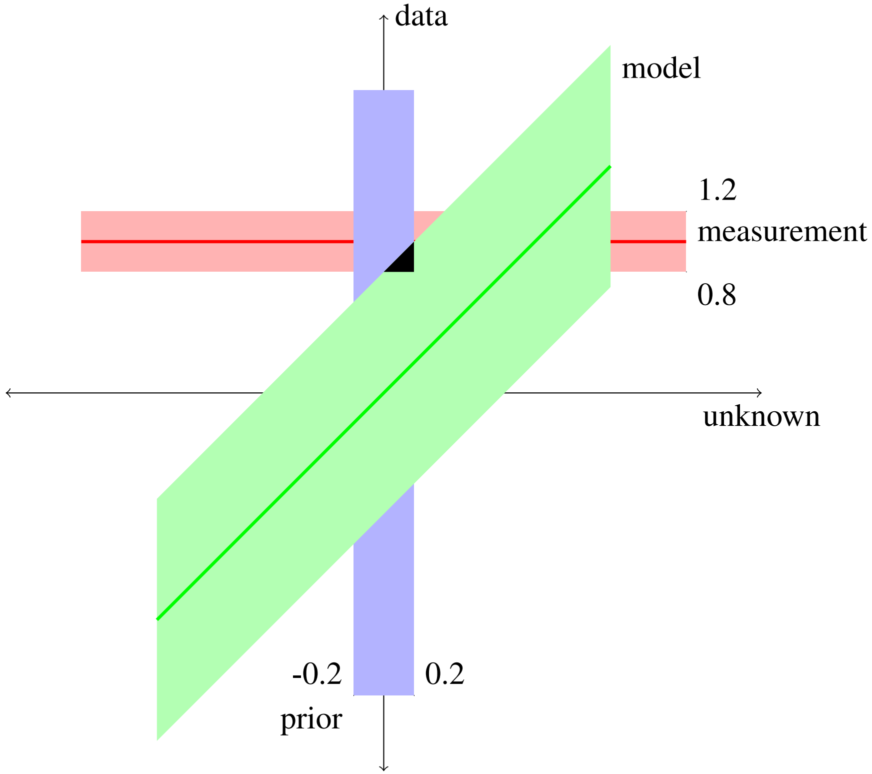
- Use of independent data to test results of assimilation;
- If assimilation is for state data is rare but if for function we can apply the model elsewhere;
- Sometimes independent data is for the unknowns but usually other observables;
- As always, need to consider the problem statistically.

Independent Measurements of the unknowns



- Lauvaux et al., GRL, 2009;
- Compare inverse fluxes with independent measurements from aircraft;
- Posterior estimates closer to aircraft fluxes.

That — triangle again



- Unknown on X-axis, obs on Y-axis;
- Now imagine light blue was posterior PDF from previous inversion;
- If just used central value (0) would not overlap obs;
- Must consider posterior uncertainty in unknowns when comparing to other obs.

Summary

- Problems with data assimilation usually sign of incorrectly specified statistics;
- Where possible check input statistics against independent data;
- Check output statistics against assumed PDFs;
- Check as many elements as possible, not just quality of fit;
- Uncertainties are a necessary component of cross-validation.