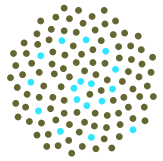


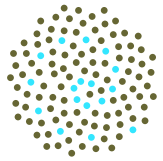
Statistics of Multiple Constraints: Analysis of Global Change and the Carbon Cycle

Ian G. Enting

MASCOS

The University of Melbourne

Acknowledgments



The ARC funds the Center of Excellence for Mathematics and Statistics (MASCOS).

My fellowship at MASCOS is supported by CSIRO through a sponsorship agreement.

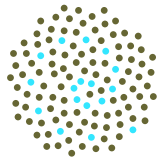
This version reflects feedback from participants at MSRI-NCAR carbon data assimilation workshop.

Collaborators:

Cathy Trudinger and Yingping Wang of CSIRO Marine and Atmospheric Research.

Roger Francey, Denis O'Brien, Peter Rayner, formerly of CSIRO Atmospheric Research.

Summary



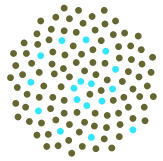
Theme: reviewing time series analysis for:

- improvement of modelling
- understanding issues of spatial analysis

Topics:

- interpreting the carbon cycle
- inversions as statistics
- digital filtering — resolution
- Kalman filter
- spatial analysis

Data for Carbon Cycle Studies

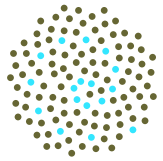


- Air sampling networks interpreted by inverse modelling;
- Satellite data, for quantities such as leaf-area index and phenology
- Terrestrial biosphere models;
- Convective boundary layer measurements;
- Stand-level flux networks;
- Ecosystem experiments;
- Small cuvettes.

From Canadell et al. 2000.

Also satellite data for CO_2 concentrations.

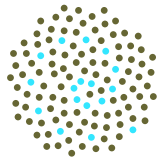
Key characteristics of statistics



- magnitude;
- degree of correlation between components;
- temporal correlation structure;
- spatial correlation structure;
- distribution;
- mismatches in averaging;
- contribution from model representativeness error.

From Raupach et al. 2005

Interpretation



Interpretation is an inverse problem, working backwards from results to causes.

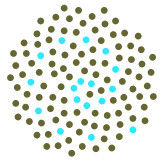
Two main inverse problems: calibration and data assimilation (deconvolution).

$$C(t) = C(0) = \int_0^t R(t - t')\, S(t')\, dt'$$

$$C(t) = C(0) = \int_0^t R(t'')\, S(t - t'')\, dt''$$

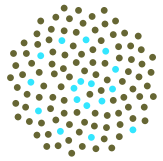
The problems of deducing model response, $R(t)$ and forcing term $S(t)$ are formally equivalent, but in practice differ in the characteristics of the statistics.

Inverse Problems as Statistics



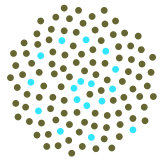
- Any uncertainty analysis needs to be based on statistics.
- Any statistical analysis is assuming (either implicitly or explicitly) some statistical model.
- *any variability that cannot be modelled deterministically . . . must be . . . modelled statistically* (Enting, 2002).

Aims of statistical analysis



- Compact characterisation of variability
- Design of techniques for data processing:
  statistical efficiency, robustness etc
- Formalism for propagating uncertainties through chain of calculations
- Design of new experiments:
  How much is uncertainty reduced?
- Testing statistical assumptions underlying data analysis techniques

Origins of Uncertainty

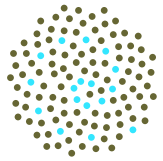


For empirical quantities: 'model' and 'data'

- Statistical variation
- Variability
- Inherent randomness
- Subjective judgement
- Linguistic imprecision
- Disagreement
- Approximation

From Morgan and Henrion (my split), *Uncertainty* (CUP, 1990).

Combining information



Bayesian: $\Pr(\underline{x}|\underline{z}) \propto \Pr(\underline{z}|\underline{x})\,\Pr_0(\underline{x})$

Multiple constraints:

$\Pr(\underline{x}|\underline{z}_1,\underline{z}_2) \propto \Pr(\underline{z}_2|\underline{x})\,\Pr(\underline{z}_1|\underline{x})\,\Pr_0(\underline{x})$

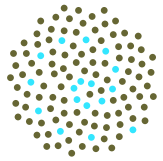
$\propto \Pr(\underline{z}_2|\underline{x})\,\Pr_1(\underline{x})$

For linear relations with multivariate normal distributions, constraints $\underline{z}_j = \underline{G}_j\underline{x} + \underline{e}_j$, with inverse covariances, $\underline{X}_j$, combine to give inverse covariance, $\underline{W}$, as:

$\underline{W} = \underline{W}_{\text{prior}} + \Sigma\,\underline{G}_j{}^{\mathsf{T}}\underline{X}_j\underline{G}_j$ for estimate

$\underline{\hat{x}} = [\underline{W}_{\text{prior}} + \Sigma\,\underline{G}_j{}^{\mathsf{T}}\underline{X}_j\underline{G}_j]^{-1}[\underline{W}\,\underline{x}_0 + \Sigma_j\,\underline{G}_j{}^{\mathsf{T}}\underline{X}_j\underline{z}_j]$

Digital filtering



Model is $z_k = s_k + n_k$ where signal has power spectrum, $f_s(\theta)$ and noise has spectrum $f_n(\theta)$.

Estimate signal as $\widehat{s}_k = \Sigma_j \Phi_j z_{k-j}$

Mean square error of estimate:

$\mathsf{E}[(\widehat{s}_k - s_k)^2] =$ bias + variance:

$$\mathsf{MSE} = \int_{-\pi}^{\pi} \left[|1 - \phi(\theta)|^2 f_s(\theta) + |\phi(\theta)|^2 f_n(\theta) \right] d\theta$$

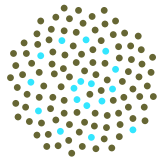
Optimal filter: $\phi(\theta) = f_s(\theta)/[f_n(\theta) + f_s(\theta)]$

with

$$\mathsf{MSE} = \frac{f_s(\theta) f_n(\theta)}{f_s(\theta) + f_n(\theta)} = \left[\frac{1}{f_s(\theta)} + \frac{1}{f_n(\theta)} \right]^{-1}$$

For MPI:BGC, Jena 2006

Characterising resolution



Use characteristic numbers:

N_{obs} How many observations?

N_{data} How many effectively independent observations?

K_{comp} How many components used in the calculations?

M_{signal} How many components needed to specify signal?

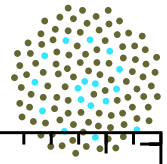
K_{synth} How many components used to fit the signal?

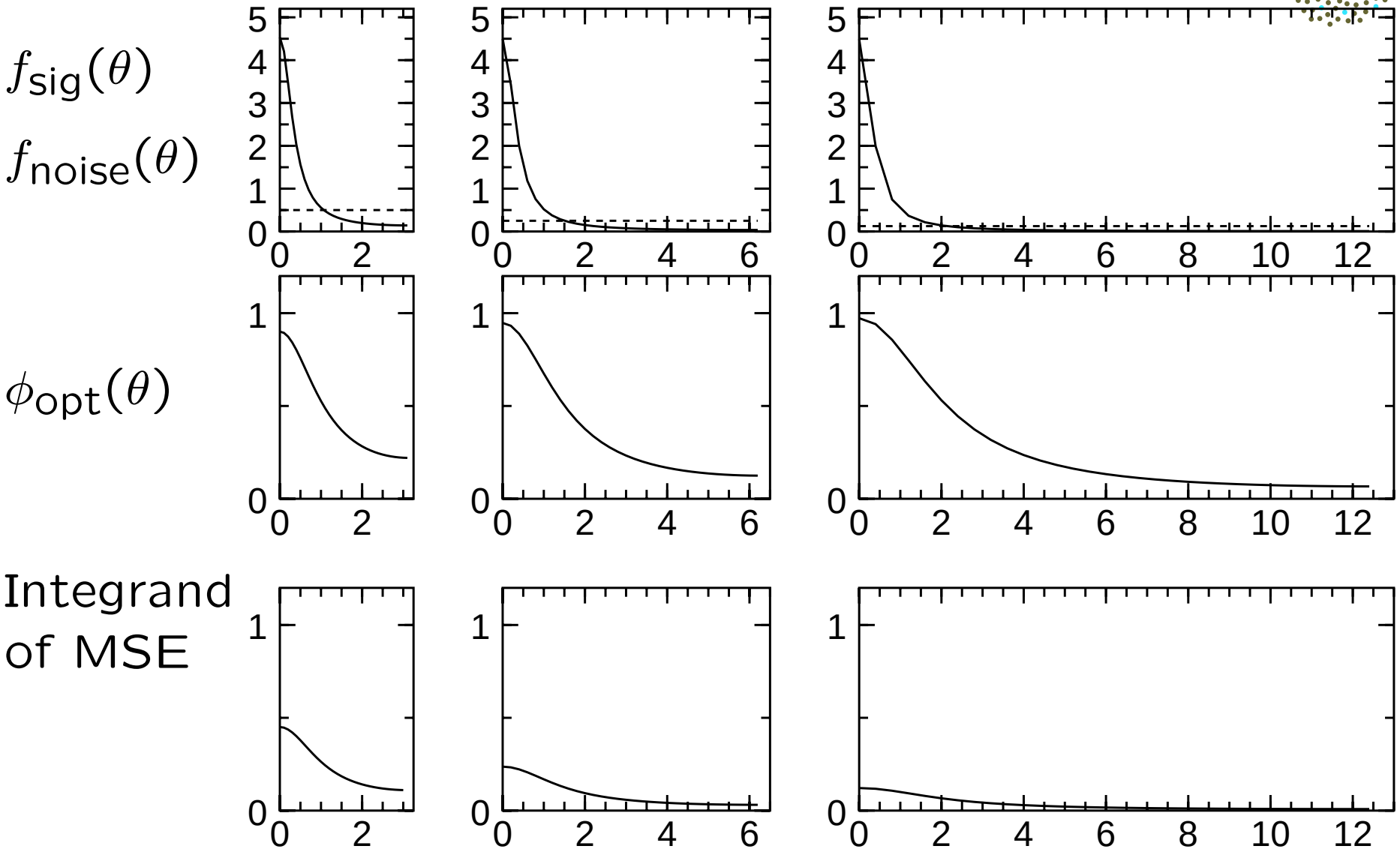
$M_{\text{s:n}}$ How many signal components exceed the noise level?

M_{target} How many signal components is one trying to estimate?

Expanded from Enting (2002: Section 8.3), including M_{signal} and distinguishing N_{obs} from N_{data}.

More data: $M_{\text{s:n}} < N_{\text{data}}$

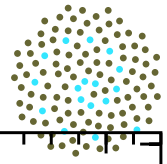


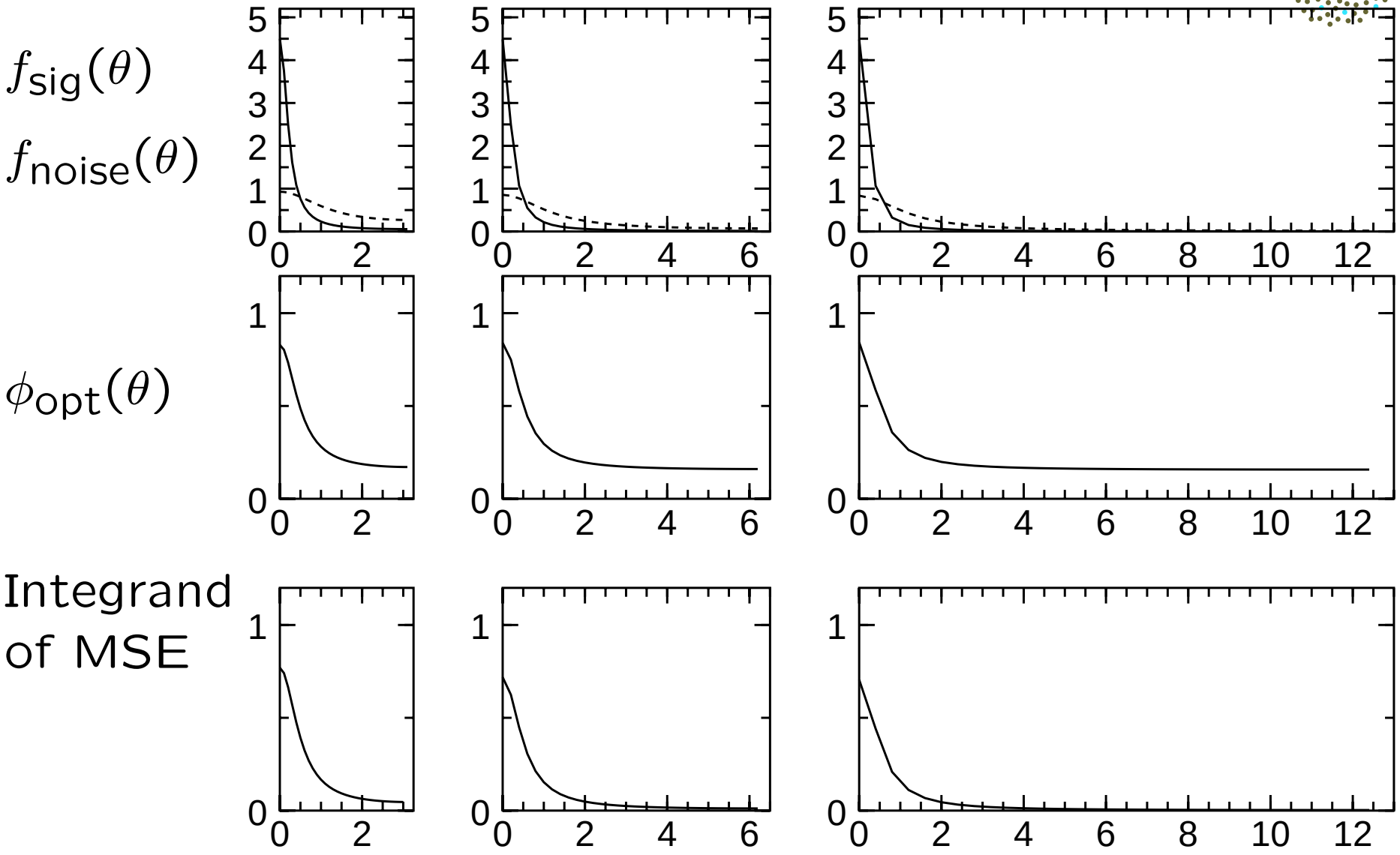


MSE $\sim 1/N_{\text{data}}$ due to reduced aliasing in noise.

For MPI:BGC, Jena 2006

Correlated data: $N_{\text{data}} < N_{\text{obs}}$

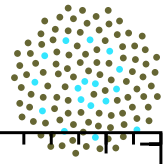


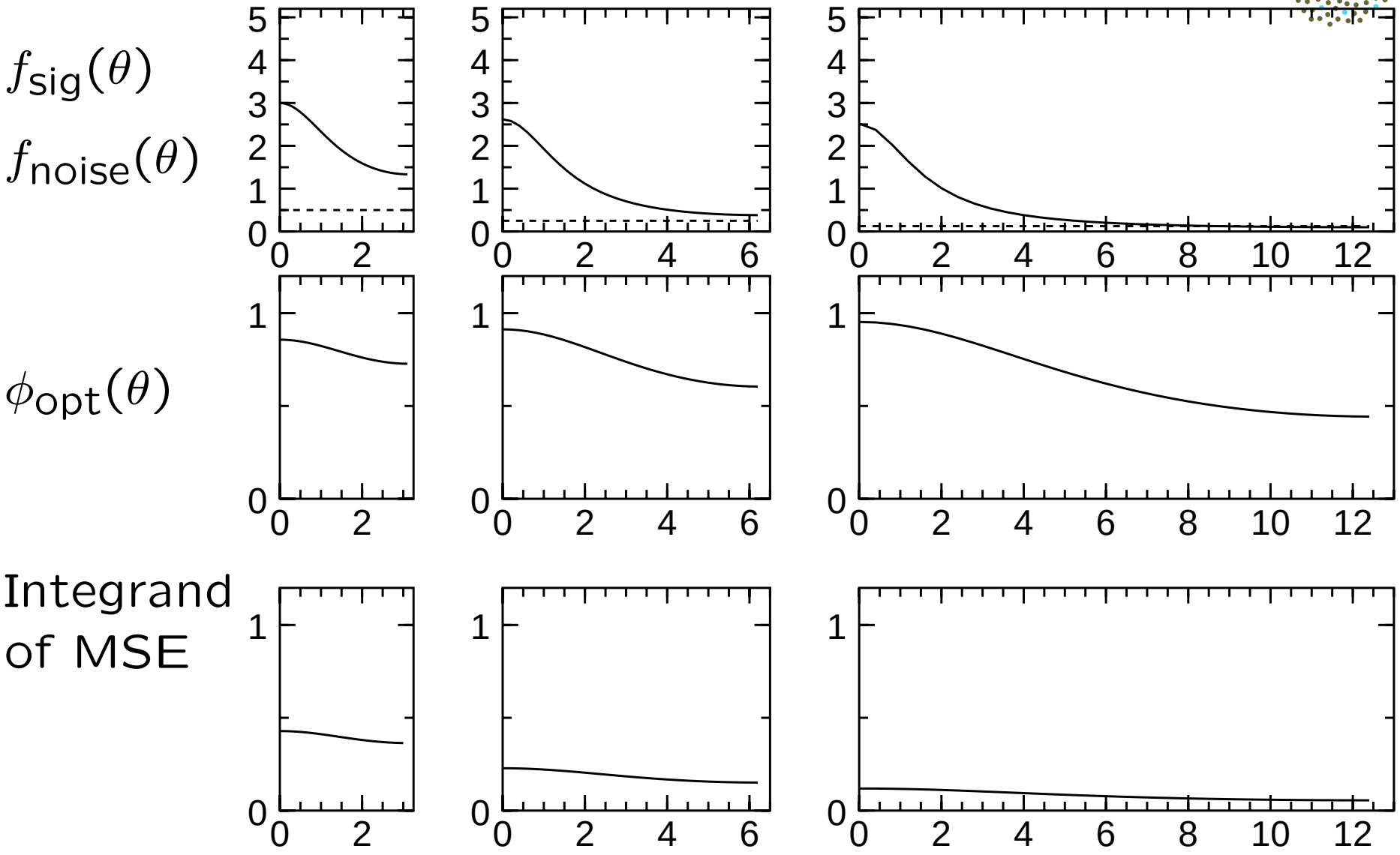


No change in MSE unless increasing N_{obs} increases N_{data}.

For MPI:BGC, Jena 2006

Aliasing: $N_{\text{data}} < M_{\text{signal}}$

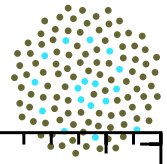


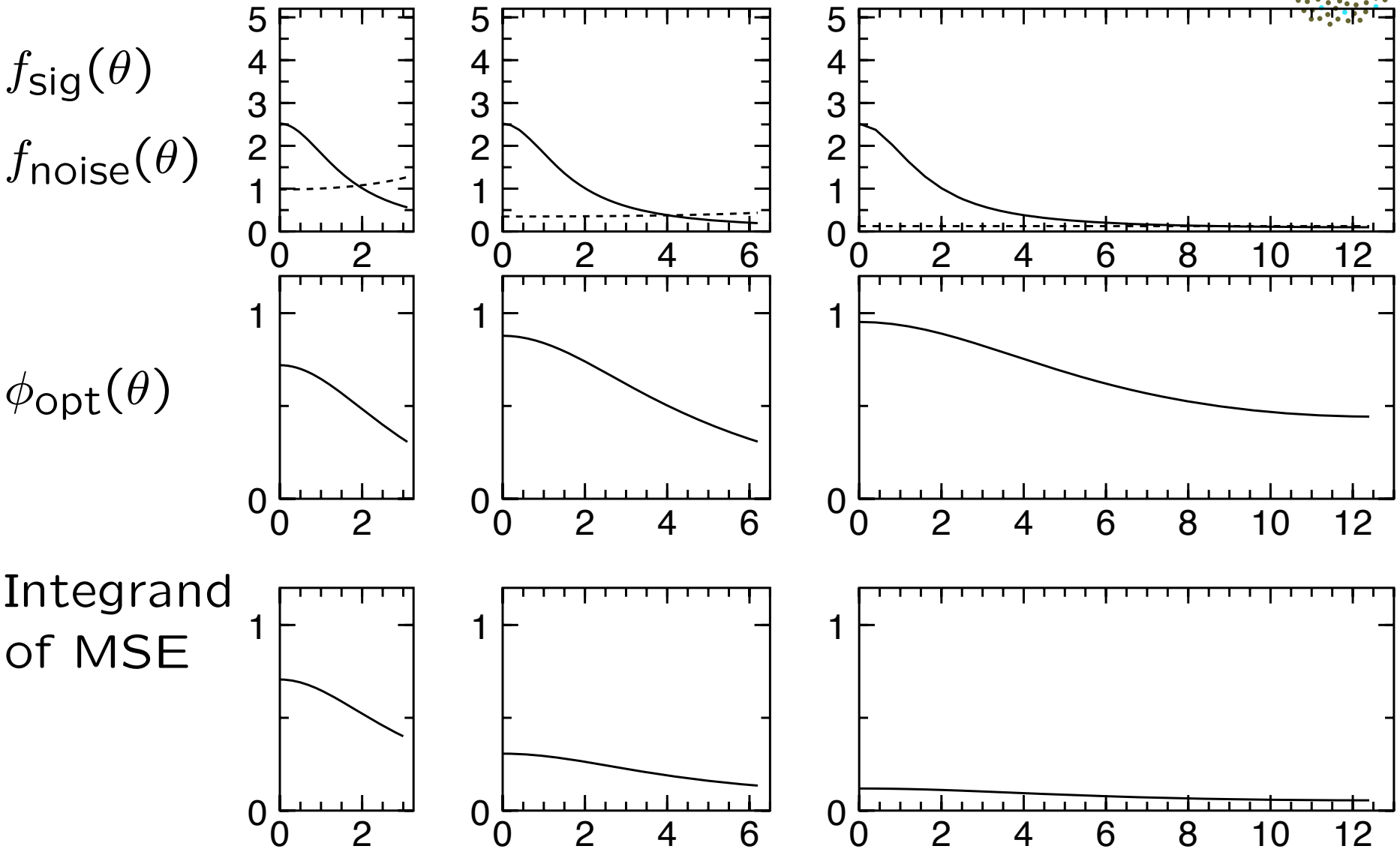


MSE too low, due to ignoring aliasing (truncation error).

For MPI:BGC, Jena 2006

Aliasing: $N_{\text{data}} < M_{\text{signal}}$

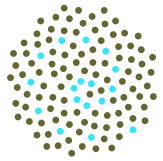




Treating aliased signal as an error contribution.

For MPI:BGC, Jena 2006

Smoothing splines



Fit a set of data, z_n, with a smooth curve, $f(t)$ chosen to minimise

$$\Theta = \sum_j [z_j - f(t_j)]^2 + \lambda \int_{t_1}^{t_N} [f''(t)]^2 \, dt$$

Spline acts as approximate digital filter with

$$\phi(\theta) = 1/[1 + (\theta/\theta_{0.5})^4]$$

where 50% attenuation occurs at:

$$\theta_{0.5} = 2\pi/T_{0.5} = [\lambda \Delta t]^{-1/4}$$

Fit is linear in data, so data uncertainties can be linearly propagated through calculations.

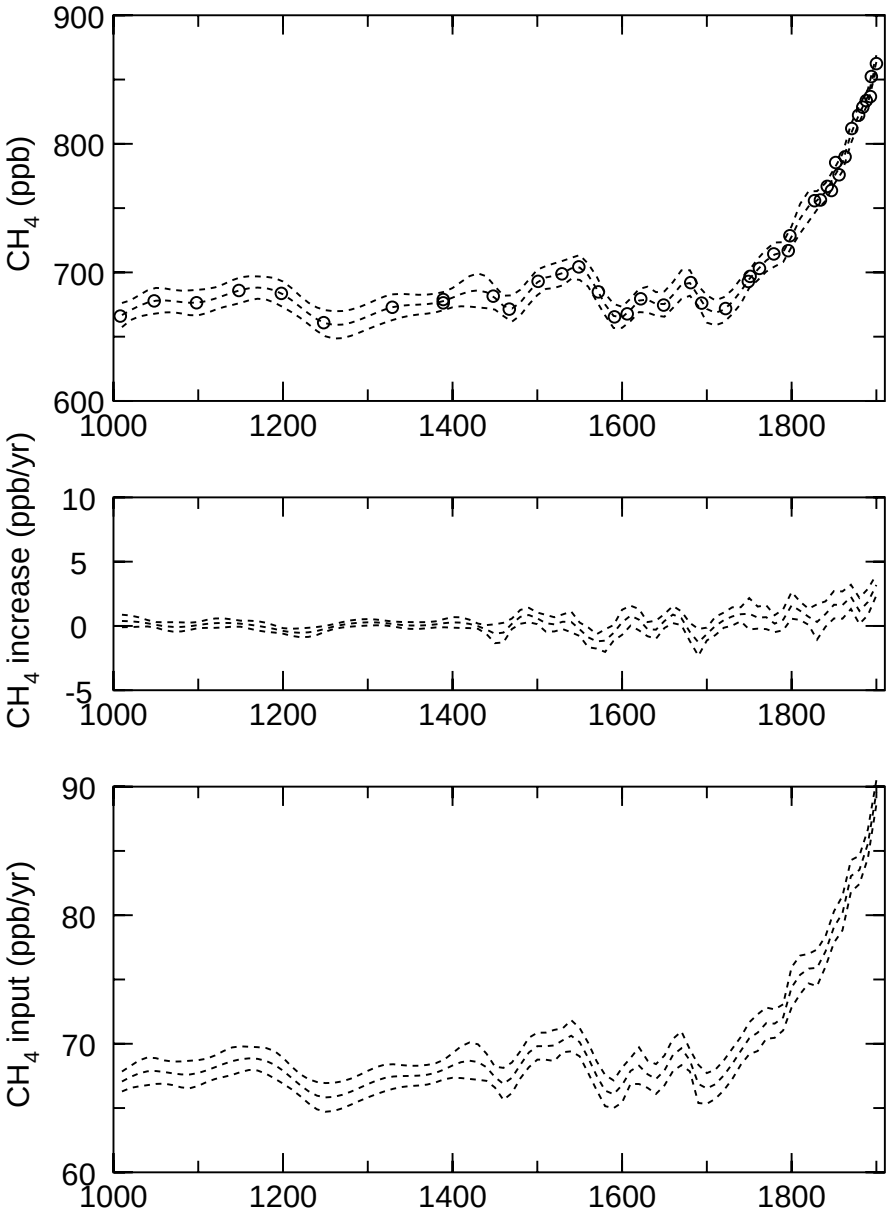
Spline example: Law Dome CH_4

Spline fit, $\widehat{f}(t)$, with ± 2 s.d. data uncertainty.
Growth rate, $\dot{\widehat{f}}(t)$, and source estimate,
$\dot{\widehat{f}}(t) + \dot{\widehat{f}}(t)/\tau$, also ± 2 s.d.

Lower pre-1500 data density implies smoother spline.

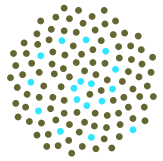
Uncertainties are uncertainty in spline (smooth part of source), not uncertainty in complete source function.

Have chosen K_{target} !!



For MPI:BGC, Jena 2006

Kalman filter paradigm



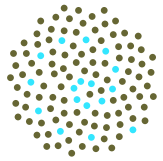
Mixed deterministic-stochastic model assumes evolution of a state, $\underline{x}$ by: $\underline{x}(n+1) = \underline{F}(n)\underline{x}(n) + \underline{u}(n) + \underline{w}(n)$ with indirect noisy observations: $\underline{z}(n) = \underline{H}(n)\underline{x}(n) + \underline{e}(n)$

Where $\underline{F}$, $\underline{H}$ and $\underline{u}$ are taken as known, and $\underline{w}$ and $\underline{e}$ are zero-mean multivariate normal with known covariances $\underline{Q}$ and $\underline{R}$.

Kalman filter formalism gives the optimal estimates, $\underline{\widehat{x}}(n)$ of state, given $\underline{z}(1)$ to $\underline{z}(n)$.

Combines mulitvariate normal distributions of observations, $\underline{z}(n)$, and projection $\underline{\widehat{x}}(n|\underline{z}_1 \ldots \underline{z}_{n-1})$.

Modelling for Kalman filter



State-space model for methane from ice cores:

x_1 is methane concentration, x_2 is source.

$$\underline{F} = \begin{bmatrix} 1 - \Delta t/\tau & a\Delta t \\ 0 & 1 \end{bmatrix} \qquad \underline{Q} = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}$$

$\underline{H} = [1, 0]$ (or ([0, 0] if no data).

Data uncertainty R, unit conversion factor, a.

Thus, source is modelled as a 'random walk'.

Simplified from Trudinger et al, 2002.

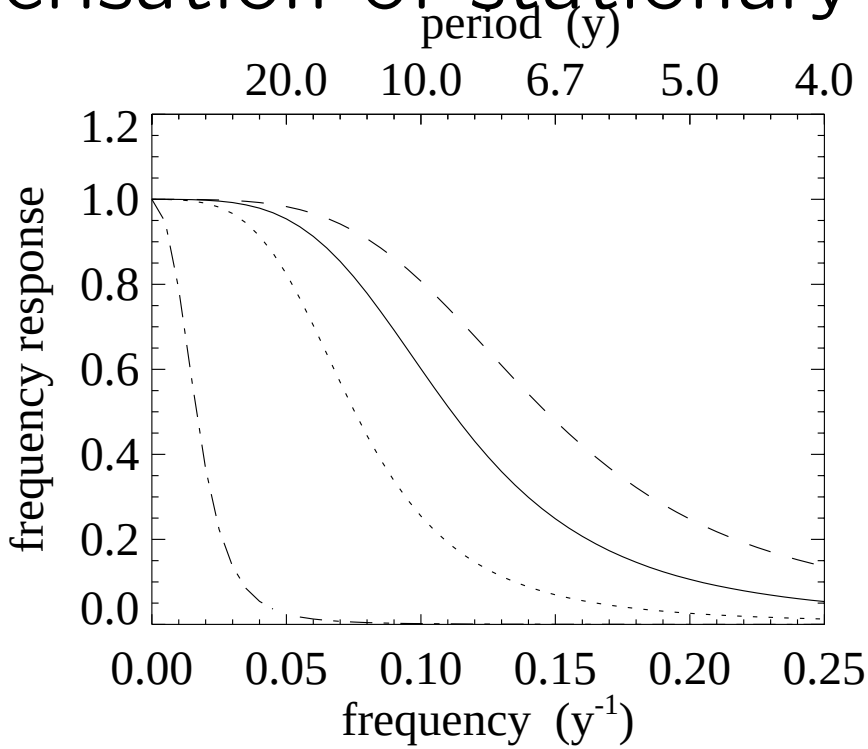
Kalman filter response

Frequency domain characterisation of stationary case of Kalman filter

E.g. noise as white noise, $f_n = R/2\pi$

Random walk model of forcing

$f_s \propto 1/(1-\cos\theta)$



R/Q ratio (9/81, 25/81, 49/36 for $\Delta t = 2$, and 49/1 for $\Delta t = 20$, right to left) changes the relative weights of 'prior' and observations in frequency-dependent manner.

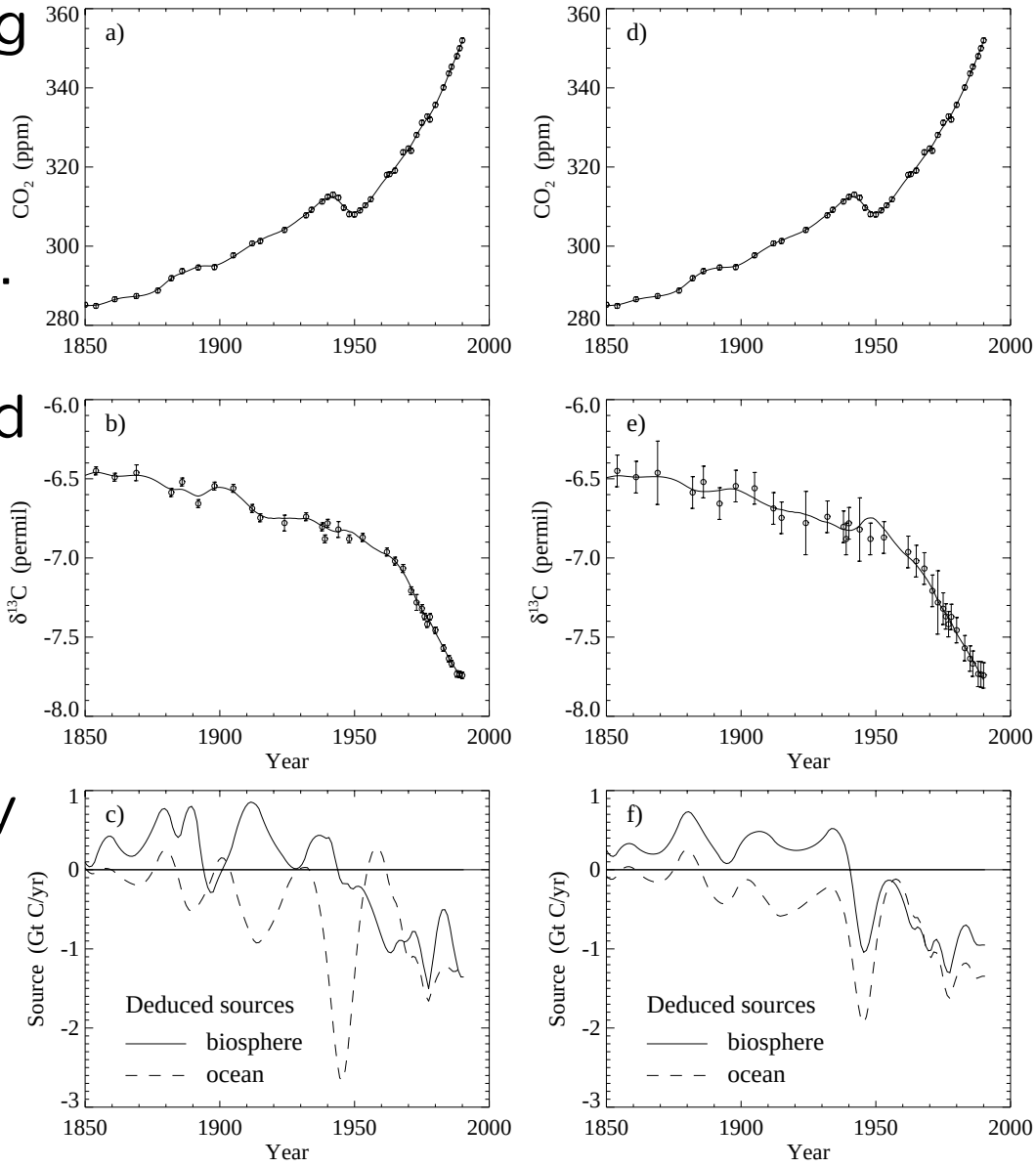
Kalman filter on CO2

Combined model, including CO_2 and $^{13}CO_2$, using concentration values corrected for firn diffusion.

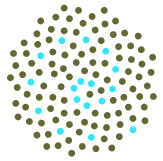
Case on left uses published $\delta^{13}C$ uncertainties.
Case on right multiplies these by four.

Again, scale of uncertainty affects the optimal smoothing.

(From Trudinger et al., 2002).



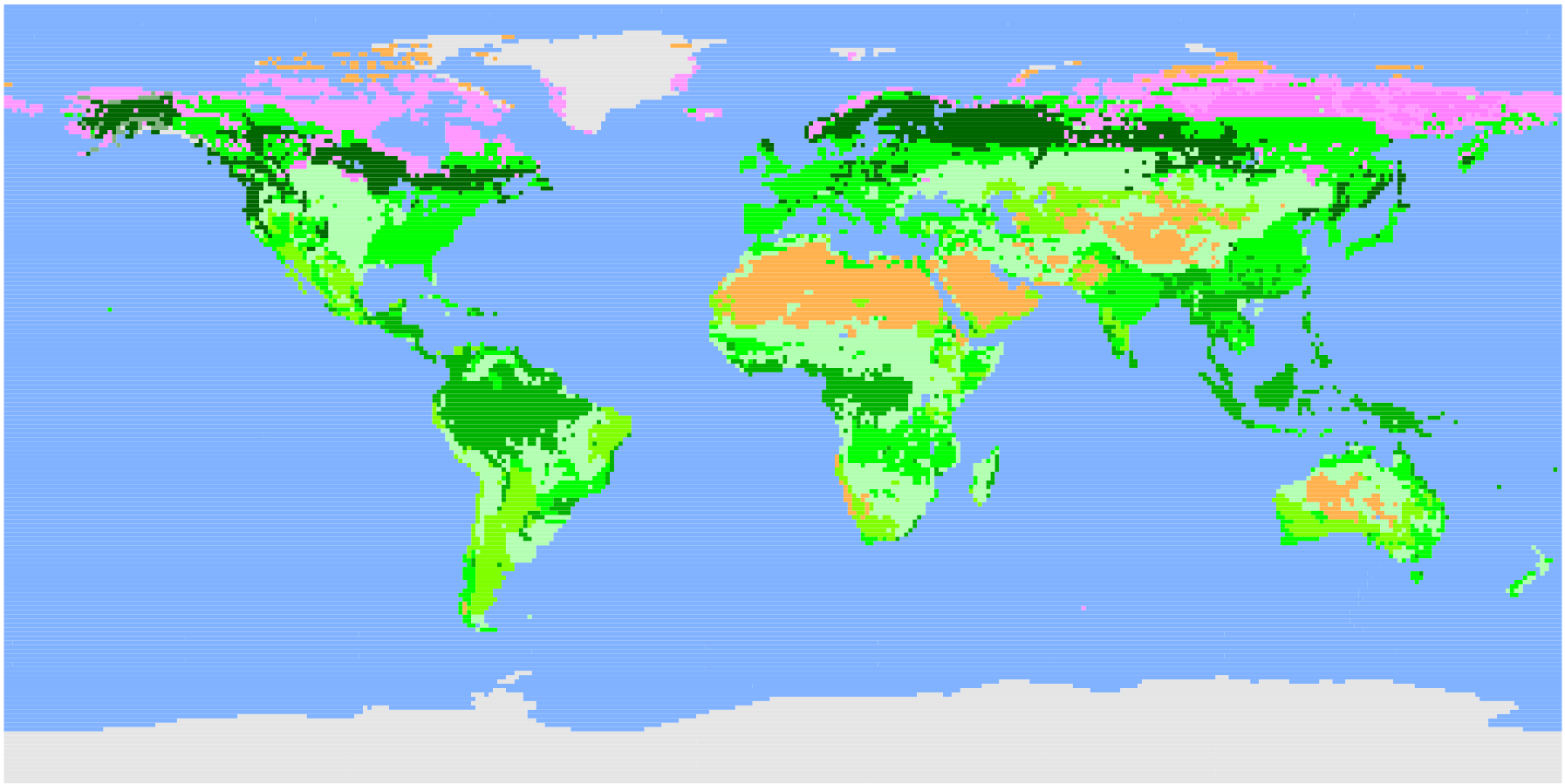
Time-dependent CO_2 inversion



As with simple Kalman filter case, statistical assumptions about structure of errors in time can greatly influence what is estimated:

- Synthesis in terms of independent monthly pulses (Rayner et al 1999), effectively assumes no long-term systematic error.
- Representing prior information as 'mean-plus-anomaly' avoids artificially-low uncertainty on posterior mean.

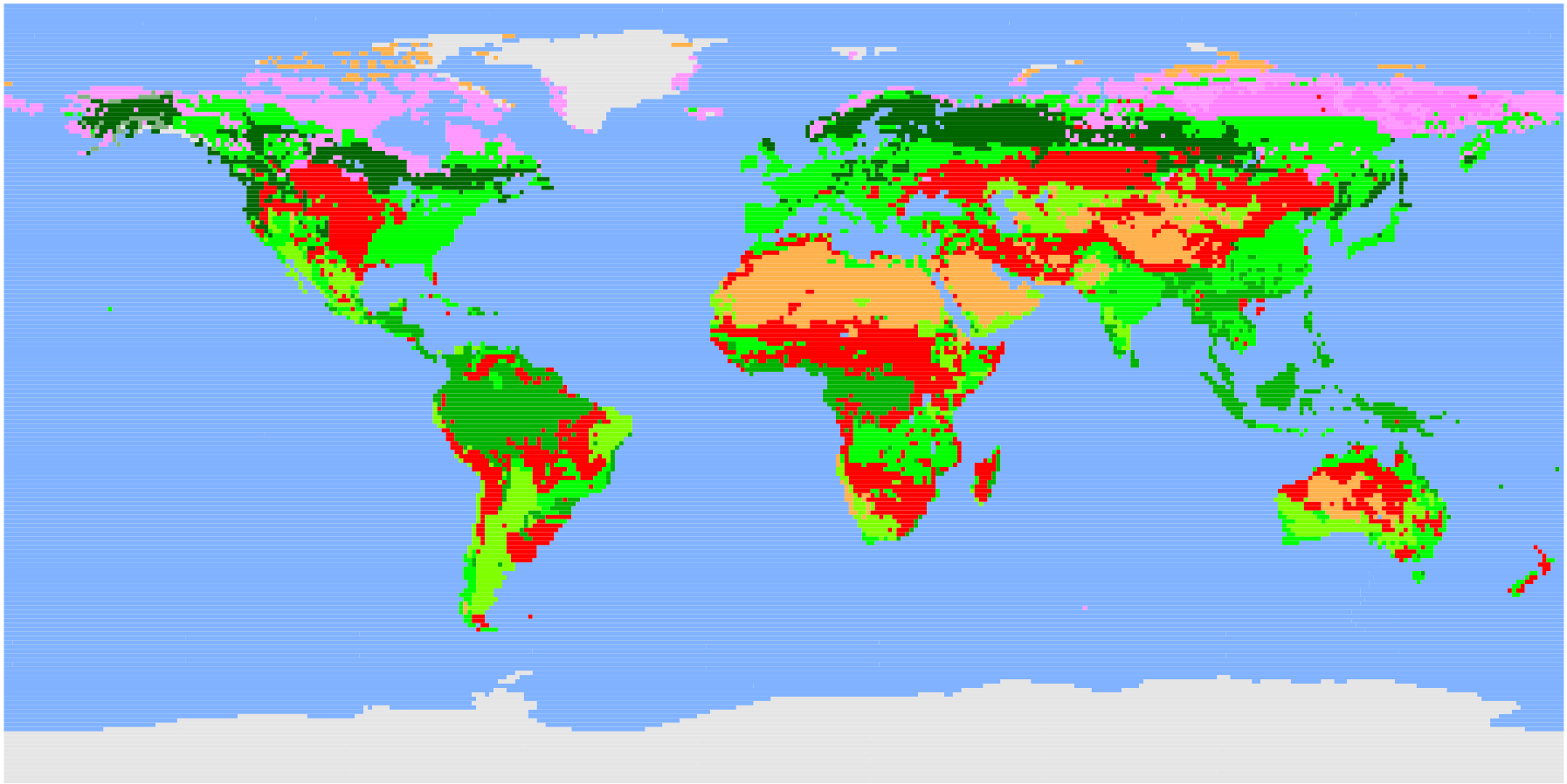
Terrestrial distribution



In process inversions, biome-specific distributions will be modulated by climatic variations.
Map data from Matthews 1983.

For MPI:BGC, Jena 2006

Grassland footprint

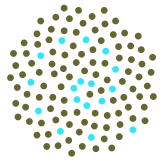


Distribution of grasslands gives a ^{13}C signal due to C3-C4 differences.

Extracting signal from sparse samples problematic.

For MPI:BGC, Jena 2006

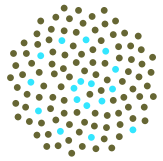
Number of modes



In an ill-conditioned inverse problem, only a limited number of modes will be resolved by the data.

- these may not be the modes that you want to know about
- you don't get to choose which modes are resolved
- c.f. biased estimates by Fan et al. by putting fine source discretisation in region (Nth America) without corresondingly fine data set.

Designing inversions



An exploratory signal-to-noise analysis doesn't need to be as precise as actual inversion.

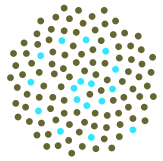
$$\widehat{x} = \sum_j a_j z_j \quad \text{observations } z_j \text{ with variance } R_j$$

typically with lots on near-cancellations in an ill-conditioned problem.

$$\text{However} \qquad \mathsf{E}[(\widehat{x} - x)^2] = \sum_j |a_j|^2 R_j$$

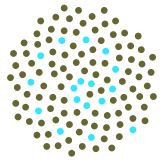
and so the variance calculation is much less sensitive to errors in the a_j (i.e. model error).

Approach



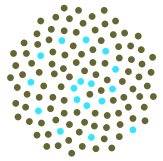
- Use a 'toy' model with purely diffusive transport
- Amplitude of responses for full 3-D model are approximated well by the $1/n$ response to latitudinal variation (Reality check)
- Inversion factor $\propto n$ comparable to numerical differentiation,
- Actual inversions confirm that latitudinally-integrated flux has much less correlation, than the actual flux estimates.

Generalities



- Satellite data don't have the $1/\sqrt{n(n+1)+\alpha m^2}$ attenuation factor (but vertical averages are still attenuated relative to surface distribution)
  Low precision can give valuable constraints.
- For comparable spatial resolution, E-W sampling density possibly needs to be greater than N-S density (however, consider re-visiting this, with 'toy model' including 'solid-body rotation', for advective term).

From point to globe



Observations:

Fluxes, forcings and proxies

↓

Local parametric model

↓

← Global distribution

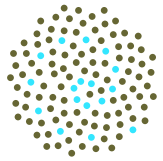
of forcing and/or proxy

↓

Global model

Contributions to uncertainty in global model are:
global distribution * parametric uncertainty +
parametric sensitivity * uncertainty in distribution

On-going application: ^{14}C

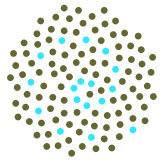


Revisit analysis by Randerson et al, 2002.
Look at seasonal modulation of ^{14}C after nuclear testing, due to seasonality in:

* transport from stratosphere to troposphere
* transport between hemispheres
* dilution of atmospheric signal by biotic respiration (initially with low ^{14}C).

What constraints does this really place on terrestrial model?

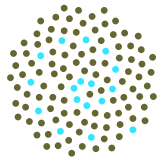
Concluding thoughts:



Take-home questions:

- What are your calculations really estimating?
- Can what you want really be estimated given limits to resolution imposed by model, data, ill-conditioning and signal-to-noise ratio?
- Are the results, including residuals, consistent with the statistical assumptions?

Further Information

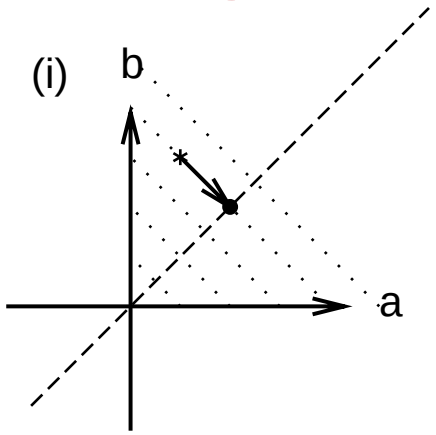


- I. Enting: *Characterising the Temporal Variability of the Global carbon Cycle*. CSIRO Atmospheric Research, Technical paper 40.
- I. Enting: *Inverse Problems in Atmospheric Constituent Transport*. 2002, CUP.
- C. Rödenbeck: Estimating CO_2 sources and sinks .... MPI-BGC Technical Report 6.

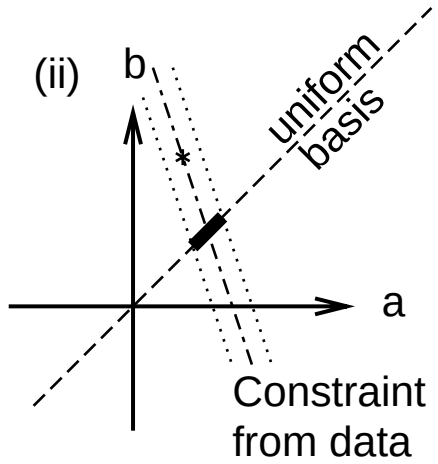
See also:

- Trudinger et al (2002a,b): Kalman filter analysis of ice-core data: 1 and 2. *JGR*
- Enting, Trudinger and Etheridge: Propagating data uncertainty through smoothing spline fits. *Tellus*: (2006).

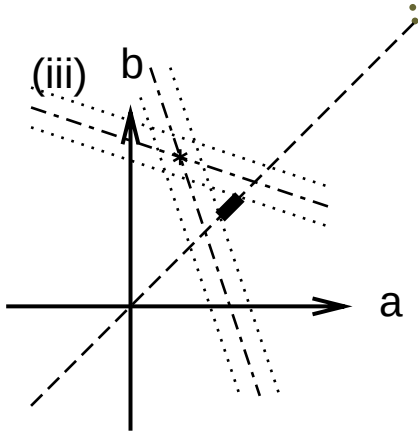
Aliasing from Truncation error



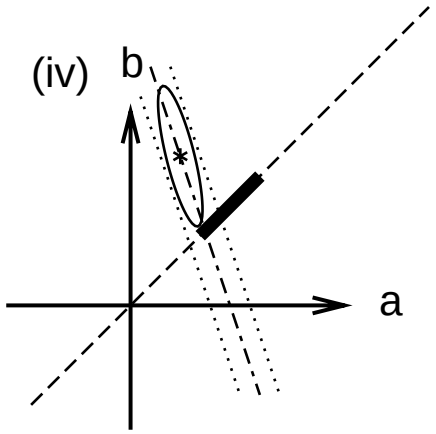
desired solution



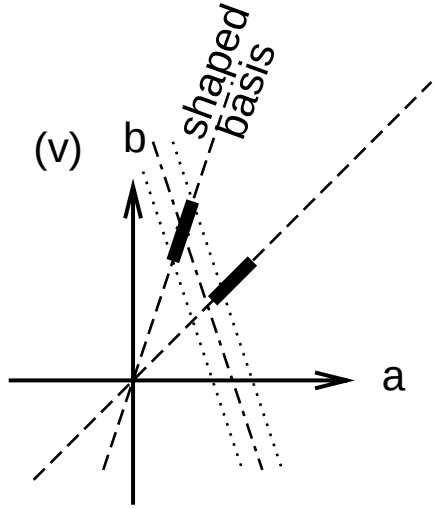
solution from biased data



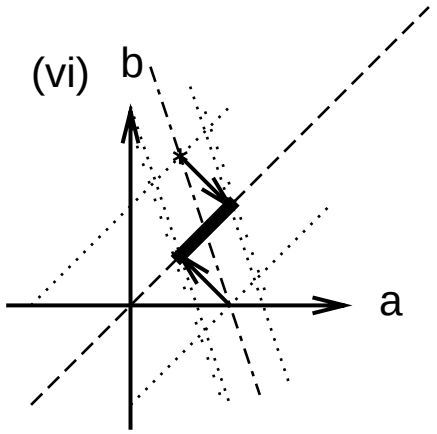
solution from unbiased data



project from full space



work in target space



apply truncation error to data

(i) objective, (ii) risk, (iii) hope, (iv) Wunsch, (v) strong priors, (vi) correction.

For MPI:BGC, Jena 2006