Uncertainty Quantification in CO_2 Retrieval

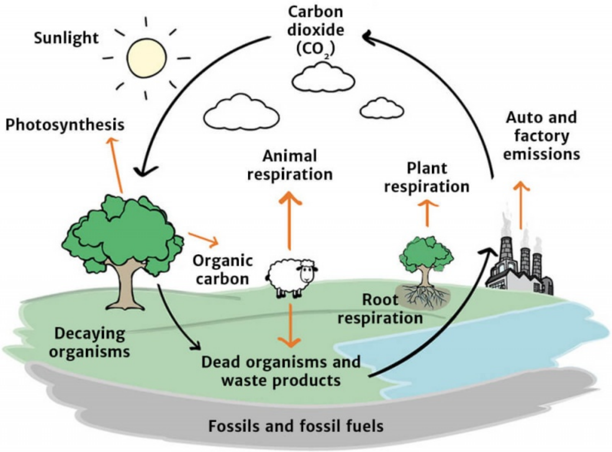
Pratik Patil

Carnegie Mellon University

Data Analysis Talk 2019

Based on joint work with Mikael Kuusela and Jonathan Hobbs

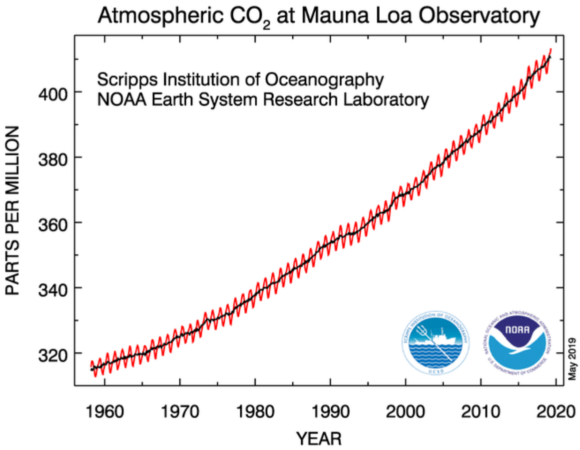
Carbon cycle



Source: Scholar Schools

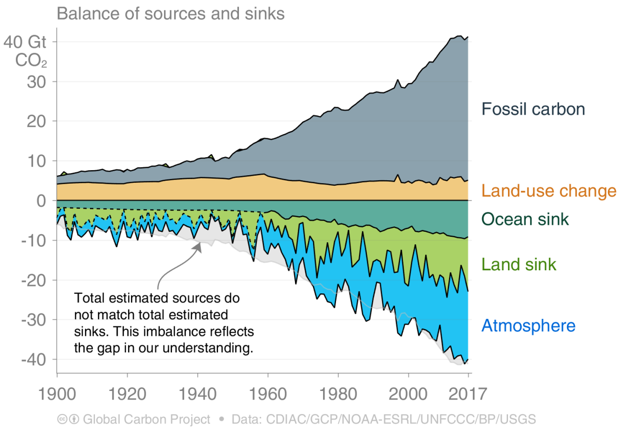
CO_2 is the main component of carbon cycle and greenhouse effect.
Balance of CO_2 is essential to sustaining life on Earth. But ...

CO_2 trend



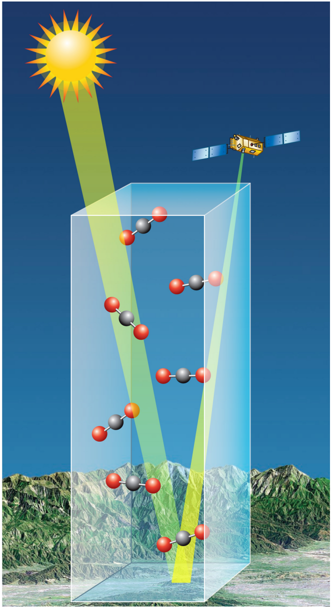
CO_2 levels were 280 ppm at start of industrial revolution. Now $45\% \uparrow$.
Present levels highest in last 800,000 and possibly last 20 million years.

CO_2 sources and sinks

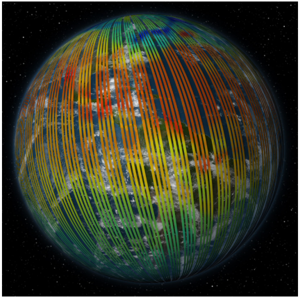


Only half of CO_2 emitted is getting absorbed. Sink processes fluctuating. What is the spatial and temporal distribution of sources and sinks?

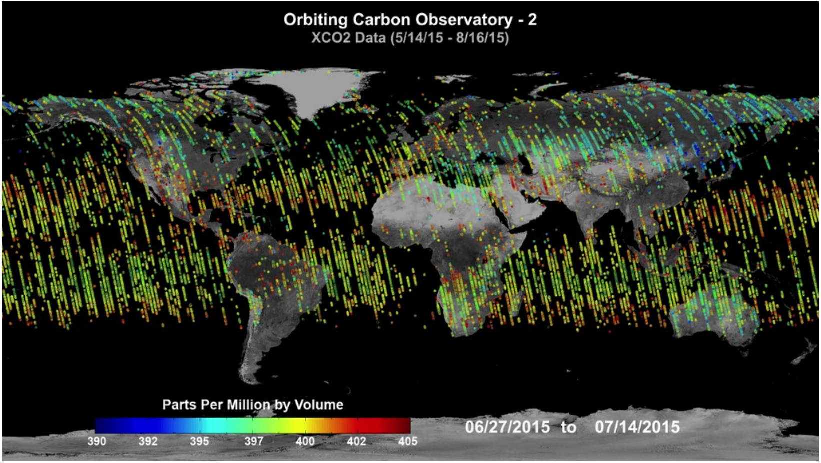
Orbiting Carbon Observatory - 2



Source: NASA



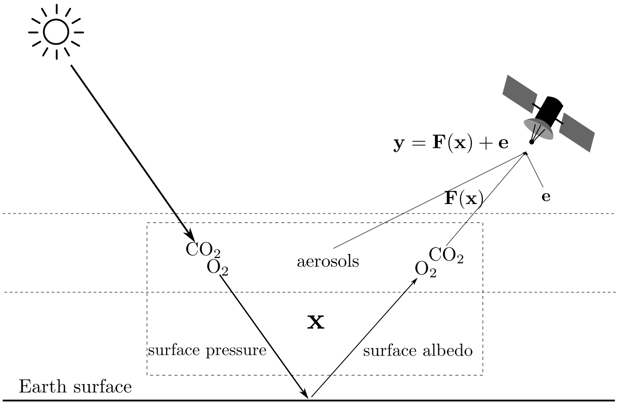
CO_2 map



Source: NASA

This project investigates how *reliable* these estimates are.

Observation system: physical model



$\mathbf{x} \in \mathbb{R}^p$: state vector, $\mathbf{F}$: forward model, $\mathbf{e}$: noise, $\mathbf{y} \in \mathbb{R}^n$: observations
The quantity of interest is a functional of state vector $\theta(\mathbf{x}) \in \mathbb{R}$

Observation system: approximated model [1]

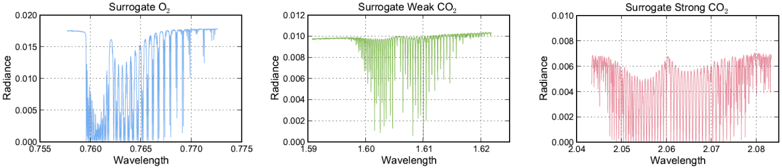
- state vector $\mathbf{x}$:
  - CO_2 profile (layer 1 to layer 20) [20 elements]
  - surface pressure [1 elements]
  - surface albedo [6 elements]
  - aerosols [12 elements]
- forward model $\mathbf{F}$:
  linearized with forward model Jacobian $\mathbf{K}(\mathbf{x}) = \frac{\partial \mathbf{F}(\mathbf{x})}{\partial \mathbf{x}}$
- noise $\mathbf{e}$: normal approximation
- observations $\mathbf{y}$:
  discretized radiances in 3 near-infrared bands [1024 in each band]
  - O_2 A-band (around 0.76 microns)
  - weak CO_2 band (around 1.61 microns)
  - strong CO_2 band (around 2.06 microns)

[1] provided by Jon Hobbs [Hobbs et al., SIAM/ASA Journal on Uncertainty Quantification, 2017]

Question of interest

Input:

- radiance observations $\mathbf{y}$

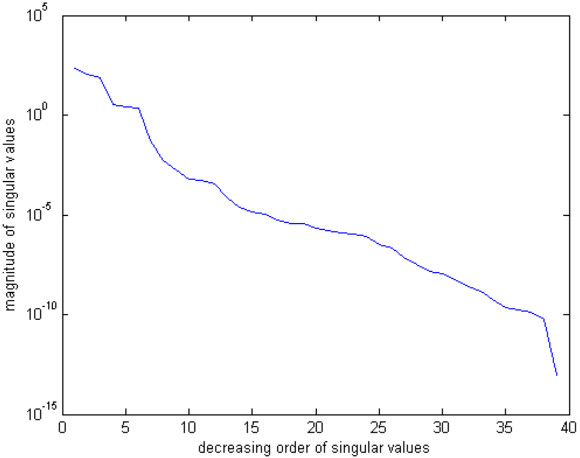


- an approximated model $\mathbf{y} \approx \mathbf{K}\mathbf{x} + \mathbf{e}$

Output:

- a functional $\theta(\mathbf{x})$ of the form $\mathbf{h}^T\mathbf{x}$ that measure column averaged CO_2 with corresponding confidence interval $\left[\underline{\theta}, \overline{\theta}\right]$ with the frequentest coverage guarantee $\mathbb{P}(\theta \in \left[\underline{\theta}, \overline{\theta}\right]) \approx 1 - \alpha$ for any $\mathbf{x}$.

Ill-posed inverse problem



Inverse problem is severely ill-posed. Exponential singular values decay.
Some eigenvalues are numerically zero leading to null space directions.

Operational retrieval

Key idea: let prior on $\mathbf{x}$ regularize the problem (Bayesian procedure)

- Assume prior distribution on $p(\mathbf{x})$.
- Combine prior with likelihood from forward model $\mathbf{F}(\mathbf{x})$ using observations $\mathbf{y}$ to get posterior $p(\mathbf{x}|\mathbf{y})$
- Compute MAP estimator $\hat{\mathbf{x}}$ maximizing $p(\mathbf{x}|\mathbf{y})$.
- Use plug-in estimate as $\hat{\theta} = \theta(\hat{\mathbf{x}})$
- From the posterior distribution $p(\mathbf{x}|\mathbf{y})$, estimate covariance $\hat{\mathbf{\Sigma}}$ of $\hat{\mathbf{x}}$.
- Use plug-in estimate for variance $\hat{\sigma}$ as $\sigma(\hat{\mathbf{\Sigma}})$.
- Set the $(1-\alpha)$ credible interval as $\left[\hat{\theta} - z_{\alpha/2}\hat{\sigma}, \hat{\theta} + z_{\alpha/2}\hat{\sigma}\right]$

Potential issues: bias and undercoverage
The true uncertainty could be drastically underestimated!

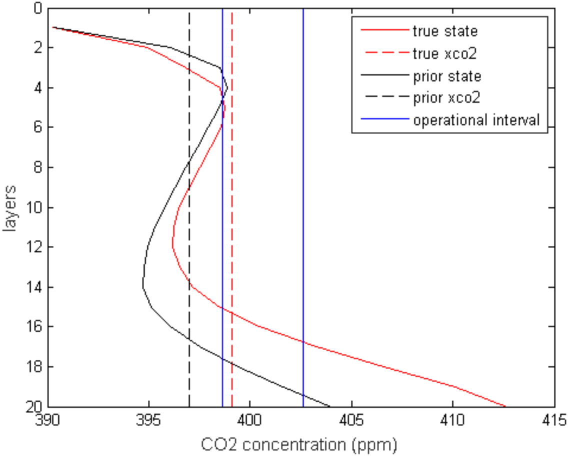
Issues with operational retrieval: single sounding

Coverage for some single soundings at Lamont, OK

state instance	operational coverage
1	0.777
2	0.800
3	0.780
4	0.787
5	0.764
6	0.830
7	0.830
8	0.729
9	0.735
10	0.787

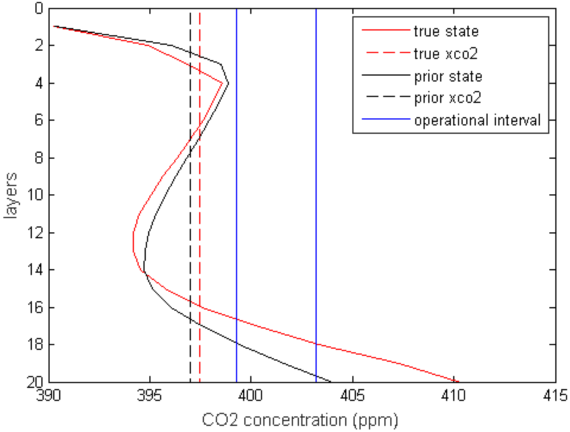
The lowest coverage sometimes drops even below 50%.

Issues with operational retrieval: single sounding



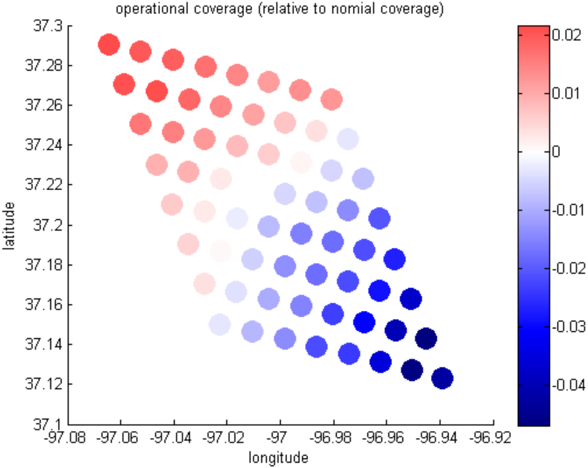
State instance: 5, operational coverage: 0.764

Issues with operational retrieval: single sounding



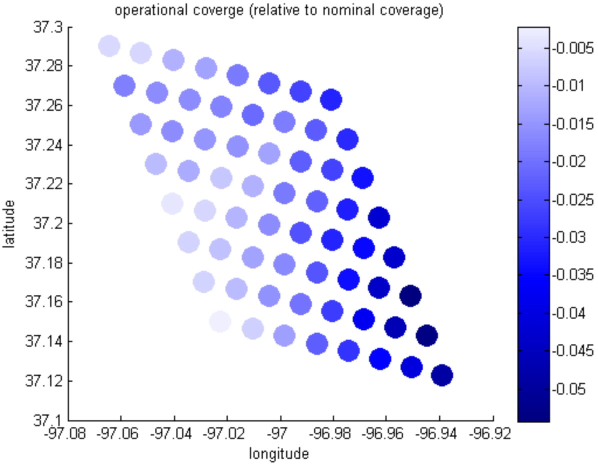
State instance: 4, operational coverage: 0.787

Issues with operational retrieval: grid sounding



Fraction of soundings below nominal coverage: 0.55

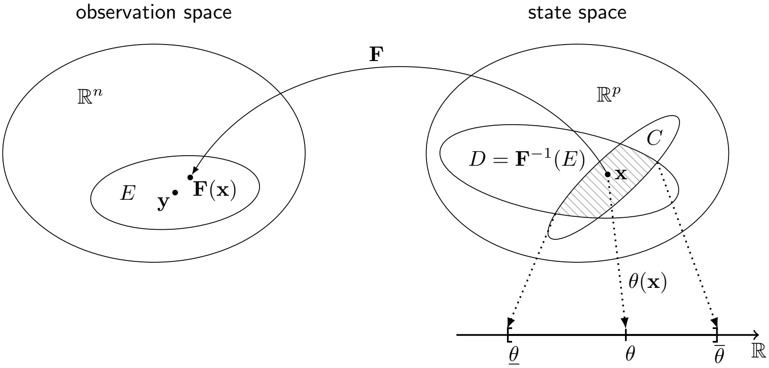
Issues with operational retrieval: grid sounding



Fraction of soundings below nominal coverage: 1

Proposed retrieval: version 1

Key idea 1: let actual physical constraints regularize the problem[2]



$$\theta = \theta(\mathbf{x}), \quad \underline{\theta} = \min_{\mathbf{x} \in C \cap D} \theta(\mathbf{x}), \quad \overline{\theta} = \max_{\mathbf{x} \in C \cap D} \theta(\mathbf{x})$$

[2] Stark, Journal of Geophysical Research, 1992; Kuusela and Stark, Annals of Applied Statistics, 2017

Proposed retrieval: version 2

- Version 1 is working harder than it needs to. The interval $\left[\underline{\theta}, \overline{\theta}\right]$ has correct finite-sample coverage for any functional θ. But we only care about a particular functional.
- Key idea 2: only require the procedure to satisfy *one-at-time coverage* rather than *simultaneous* coverage[3]
- One way is to restrict the set D in version 1 that still preserves the coverage guarantee for θ. For example, assume Gaussian white noise for simplicity. Then,
  - version 1 uses $D = \{\mathbf{x} : \|\mathbf{y} - \mathbf{F}(\mathbf{x})\|^2 \leq \chi^2_n(\alpha)\}$ which has $(1-\alpha)$ coverage set in the state space.
  - version 2 restricts it such that $D' = \{\mathbf{x} : \|\mathbf{y} - \mathbf{F}(\mathbf{x})\|^2 \leq z^2_{\alpha/2} + b^2\}$, where $b = \min_{\mathbf{x} \in C} \|\mathbf{y} - \mathbf{F}(\mathbf{x})\|$

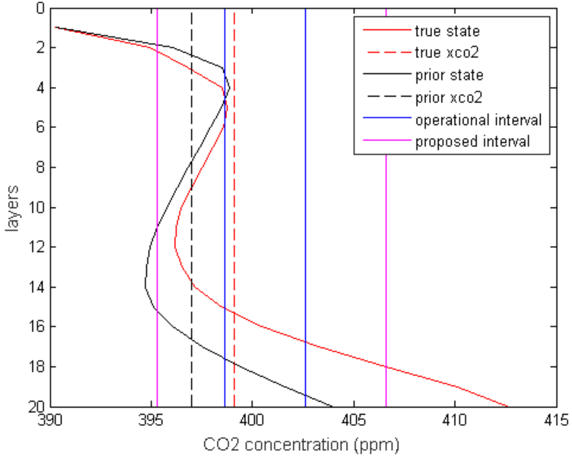
[3] inspired by Leary and Rust, SIAM Journal on Scientific and Statistical Computing, 1986

Improvements from proposed retrieval: single sounding

state instance	operational coverage	proposed coverage
1	0.777	0.952
2	0.800	0.955
3	0.780	0.952
4	0.787	0.956
5	0.764	0.953
6	0.830	0.950
7	0.830	0.960
8	0.729	0.952
9	0.735	0.955
10	0.787	0.950

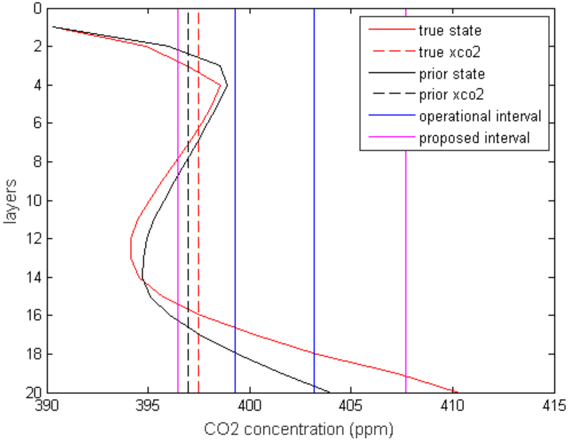
Length of operational interval about 4, proposed interval about 11.

Improvements from proposed retrieval: single sounding



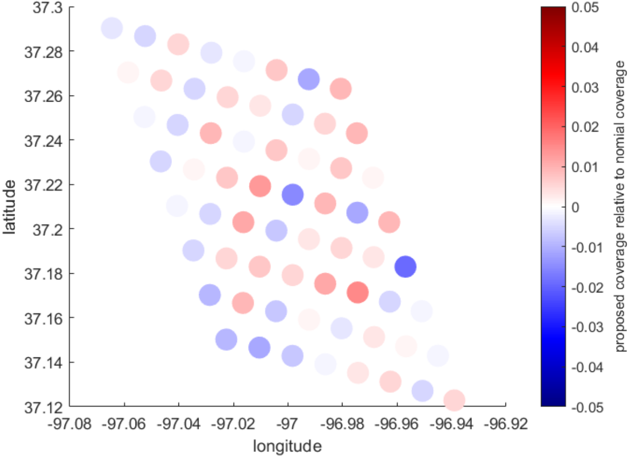
State instance: 5, proposed coverage: 0.953

Improvements from proposed retrieval: single sounding

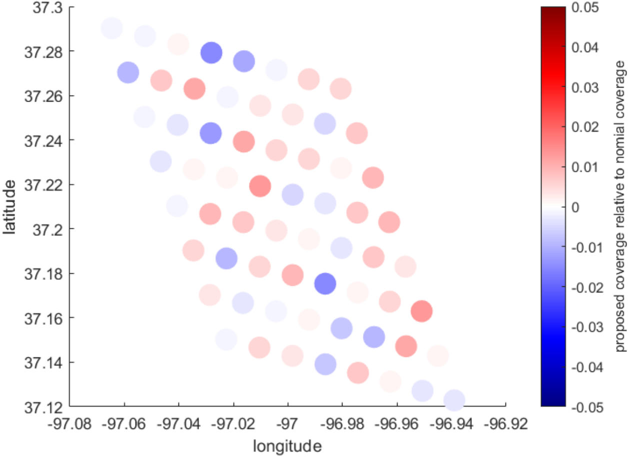


State instance: 4, proposed coverage: 0.956

Improvements from proposed retrieval: grid sounding



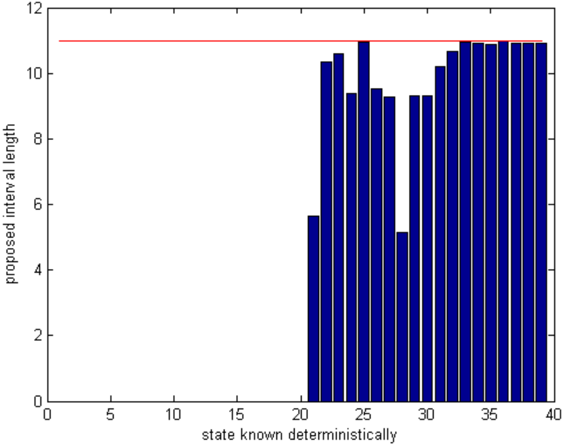
Improvements from proposed retrieval: grid sounding



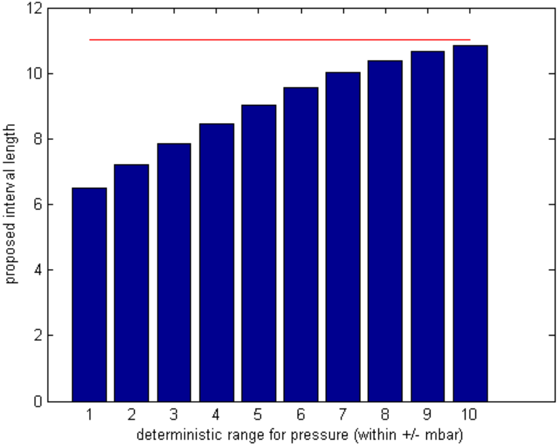
Proposed retrieval: version 3

- So far, we only used actual physical constraints on the state vector.
- But, what if we wanted to incorporate more information about state.
  - Certain ranges for some elements of state vector more likely.
  - Possibility of borrowing certainty from other sources.
- Version 3 provides a framework for incorporating additional probabilistic information and still maintaining finite-sample coverage guarantees.

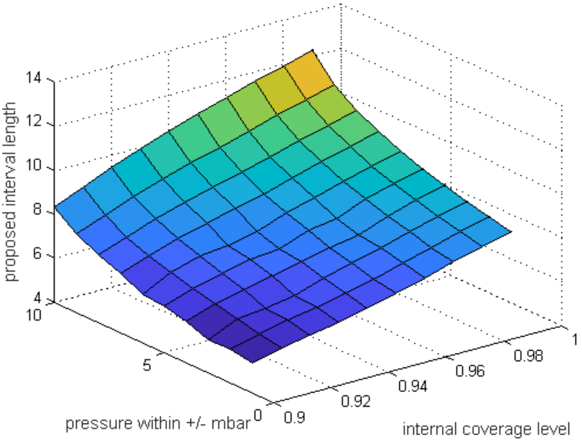
Deterministic exact information on individual elements



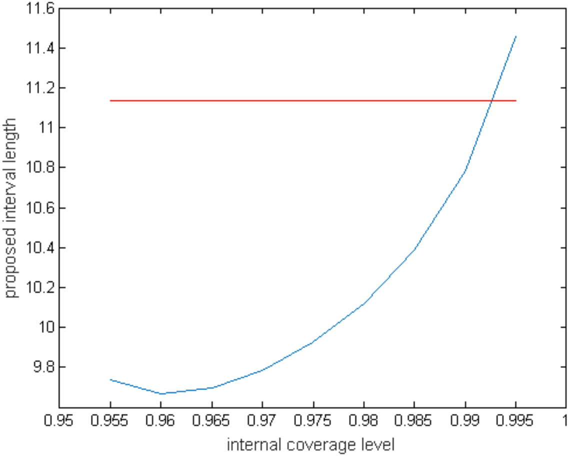
Deterministic range for pressure



Probabilistic range for pressure



Probabilistic range for pressure



Conclusions and extensions

- Uncertainties for CO2 estimates are important.
- Some evidence of potential bias and undercoverage for the operational retrieval.
- Proposed method can provide good coverage guarantees.
- Further improvements in the size of intervals from the proposed retrieval possible using additional information.
- Many extensions possible:
  - Different ways of restricting the sets for one-at-a-time intervals.
  - Optimality for the size of the intervals.
  - Combining information from different missions.
  - Different approaches for non-linear forward models.
  - Using intervals for downstream tasks instead of point estimates.