Ultra-high dimensional Bayesian analysis techniques for 21cm surveys

Phil Bull

Jodrell Bank Centre for Astrophysics
and
Centre for Radio Cosmology, U. Western Cape

Overview

Statistical challenges in 21cm analysis

Flagging and ringing (power spectra)

Spherical harmonics (maps)

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 948764)

European Research Council
Established by the European Commission

Statistical challenges in 21cm analysis

Why is 21cm data analysis so tricky?

- There is a large dynamic range between 21cm signal and contaminants
- Instrument and sky models are incomplete and inaccurate
- These inaccuracies really matter because:

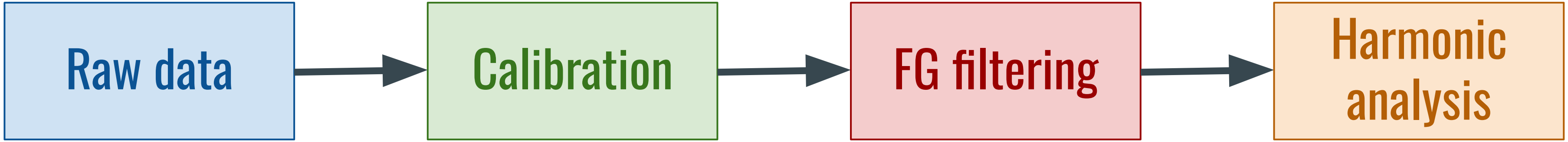
(small inaccuracy) × **(big contaminant)** $\gtrsim$ **(small 21cm signal)**

Challenge: How do we extract the extremely delicate 21cm signal:

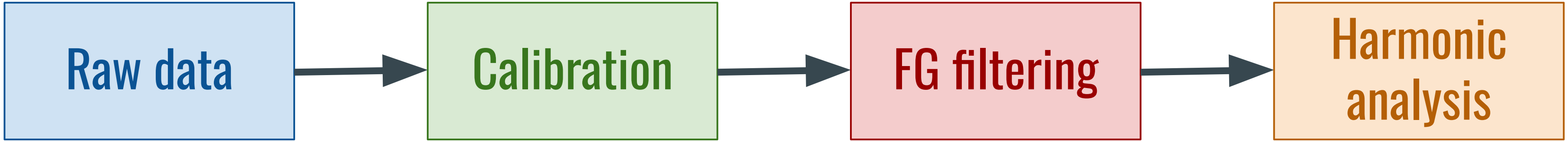
(a) without wrecking everything (over-subtraction/signal loss); or

(b) at least knowing if/when we've wrecked everything?

Example: “Traditional” pipeline



Example: “Traditional” pipeline

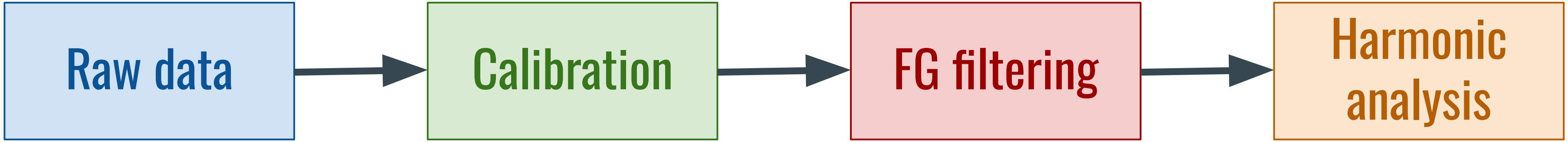


Incomplete sky model
+
Simplifying assumptions about instrument model
↓
Frequency-dependent **modulation** of true signal

Overlap between 21cm signal and foregrounds
+
Complex spectrum due to modulation/instrument
↓
Over-subtract/over-fit FGs, leading to **signal loss**

Gaps in data due to RFI flagging etc.
+
Imperfect in-painting/ deconvolution of flags
↓
Ringing and **mode coupling**

Example: “Traditional” pipeline



Can we make improved models… or at least rigorously estimate the **uncertainty** and **model error** to better interpret our results?

Bayesian approach:

- **Propagate** uncertainties so biases *from all steps* included within errorbar
- Permit flexible “nuisance” parametrisations to **absorb** model errors
- Use priors to **prevent** weird/unphysical results
- **Inspect** posterior distribution to understand the results

Example: "Traditional" pipeline

Bayesian approach:

- **Propagate** uncertainties so biases *from all steps* included within errorbar
- Permit flexible "nuisance" parametrisations to **absorb** model errors
- Use priors to **prevent** weird/unphysical results
- **Inspect** posterior distribution to understand the results

Challenges:

- Can we make models that are accurate without being overly flexible?
- How can we handle the gigantic number of parameters?

Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
    - Missing data causes ringing (very bad for 21cm due to dynamic range)

Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
  - Missing data causes ringing (very bad for 21cm due to dynamic range)
- What to do?
  - Lomb-Scargle / Least-Squares Spectral Analysis
  - In-painting (fill-in missing data)
  - Deconvolve the mask

All involve implicit models of missing data

Flagging and ringing

- Flagging of RFI-affected channels is unavoidable
- This is a major headache for harmonic analysis (e.g. power spectra)!
  - Missing data causes ringing (very bad for 21cm due to dynamic range)
- What to do?
  - Lomb-Scargle / Least-Squares Spectral Analysis
  - In-painting (fill-in missing data)
  - Deconvolve the mask
  - **Infer the masked data** → *Gaussian constrained realisations*

All involve implicit models of missing data

GCR and Gibbs sampling

Kennedy et al. [2211.05088]

- **GCR:** Draw samples of the 21cm signal + foregrounds given **observed data**, **foreground basis functions**, **noise** and **21cm signal covariance** estimates

$$p(\mathbf{e}, \mathbf{a}_{\rm fg} | \mathbf{E}, \mathbf{g}_j, \mathbf{N}, \mathbf{d}) \propto p(\mathbf{d} | \mathbf{e}, \mathbf{a}_{\rm fg}, \mathbf{g}_j, \mathbf{N}) p(\mathbf{e} | \mathbf{E})$$

GCR and Gibbs sampling

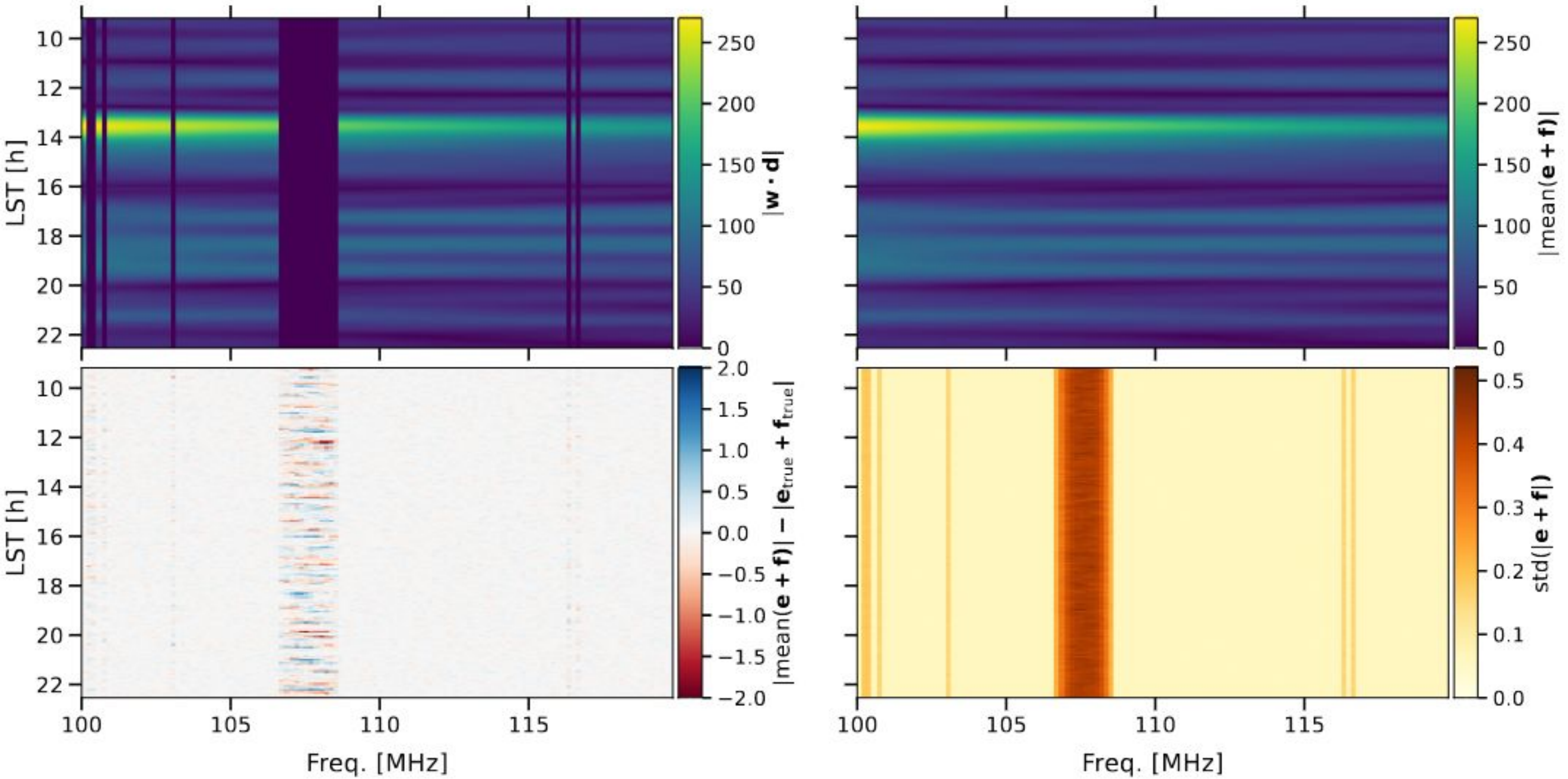
Kennedy et al. [2211.05088]
Burba et al. [in prep.]

- **GCR:** Draw samples of the 21cm signal + foregrounds given **observed data**, **foreground basis functions**, **noise** and **21cm signal covariance** estimates

$$p(\mathbf{e}, \mathbf{a}_{\rm fg} | \underline{\mathbf{E}}, \underline{\mathbf{g}_j}, \underline{\mathbf{N}}, \mathbf{d}) \propto p(\mathbf{d} | \mathbf{e}, \mathbf{a}_{\rm fg}, \underline{\mathbf{g}_j}, \underline{\mathbf{N}}) p(\mathbf{e} | \underline{\mathbf{E}})$$

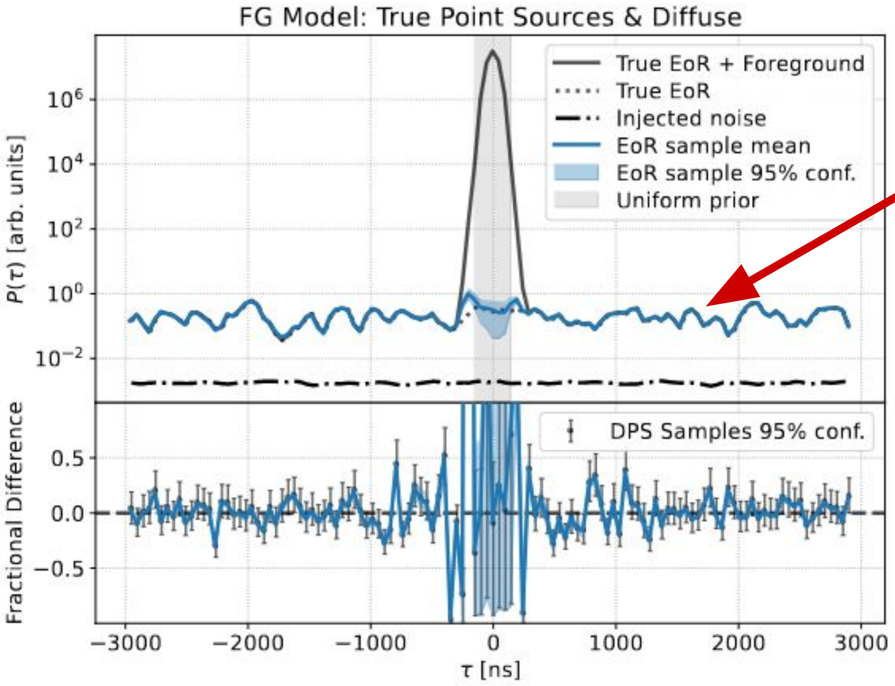
- Each sample has **no gaps**, so Fourier analysis can be applied exactly (no ringing). Repeat many times to build up statistical distribution.
- What if the 21cm signal covariance is poorly known? → **Gibbs sampling method**
  - Iteratively sample 21cm signal (+ foregrounds), then 21cm covariance

$$\mathbf{s}_{i+1} \leftarrow p(\mathbf{s}_i | \mathbf{S}_i, \mathbf{N}, \mathbf{d})$$
$$\mathbf{S}_{i+1} \leftarrow p(\mathbf{S}_i | \mathbf{s}_{i+1}).$$



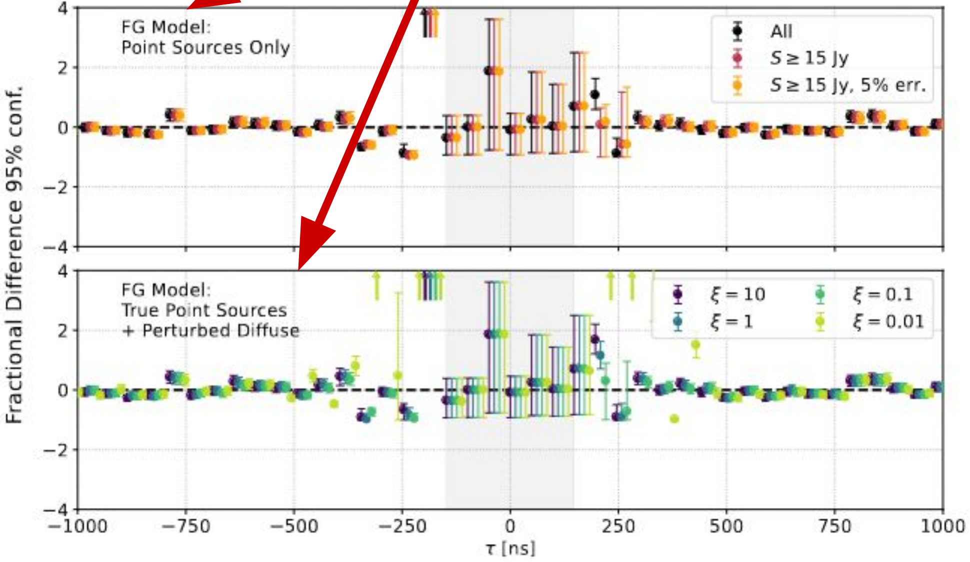
Kennedy et al. [2211.05088]

Validation



Recovers 21cm power spectrum in high SNR and low SNR regimes

Robust to **missing sources** and **spectral errors**



Caveat: Also need to model temporal correlations

Burba et al. [in prep.]

Hydra: A high-dimensional Gibbs sampler for 21cm data

github.com/HydraRadio

Primary beam
Mike Wilensky

21cm field/power
Jacob Burba

Fourier modes
Zheng Zhang

Spherical harmonics
Katrine Glasscock

HYDRA

Reflection systematics
Sohini Dutta

Gains + point sources
Hugh Garsden

MeerKAT
Geoff Murphy

Spherical harmonics (diffuse emission)

Glasscock et al. [in prep.]

- Need to connect a sky model (e.g. a map) to the observed visibilities
  - Large angular scales: spherical harmonics (curved sky, orthonormal etc.)
  - For max. angular mode ℓ_{max} we have $(\ell_{max} + 1)^2$ parameters (per frequency)
- Baseline of length d is sensitive to SH modes $\ell \sim \pi d/\lambda$ ~ d / [1 m] at 100 MHz
- We can write the visibilities as a linear model:

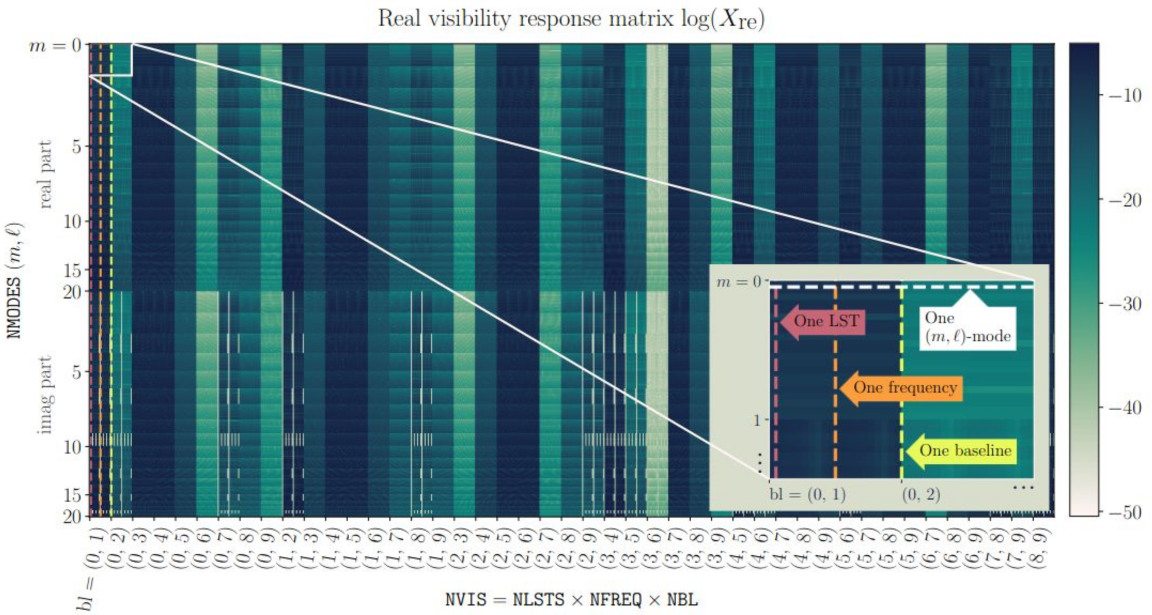
SH coefficients × **response function:**

$$V_{ij} = \sum_{\ell m} \underline{\delta V_{ij}^{\ell m}(\nu, t)}\, \underline{a_{\ell m}}$$

Spherical harmonics (diffuse emission)

Glasscock et al. [in prep.]

- Response operator can be very large; function of freq./time/baseline/SH mode
- Time dimension can be replaced by a rotation (in Equatorial coords)



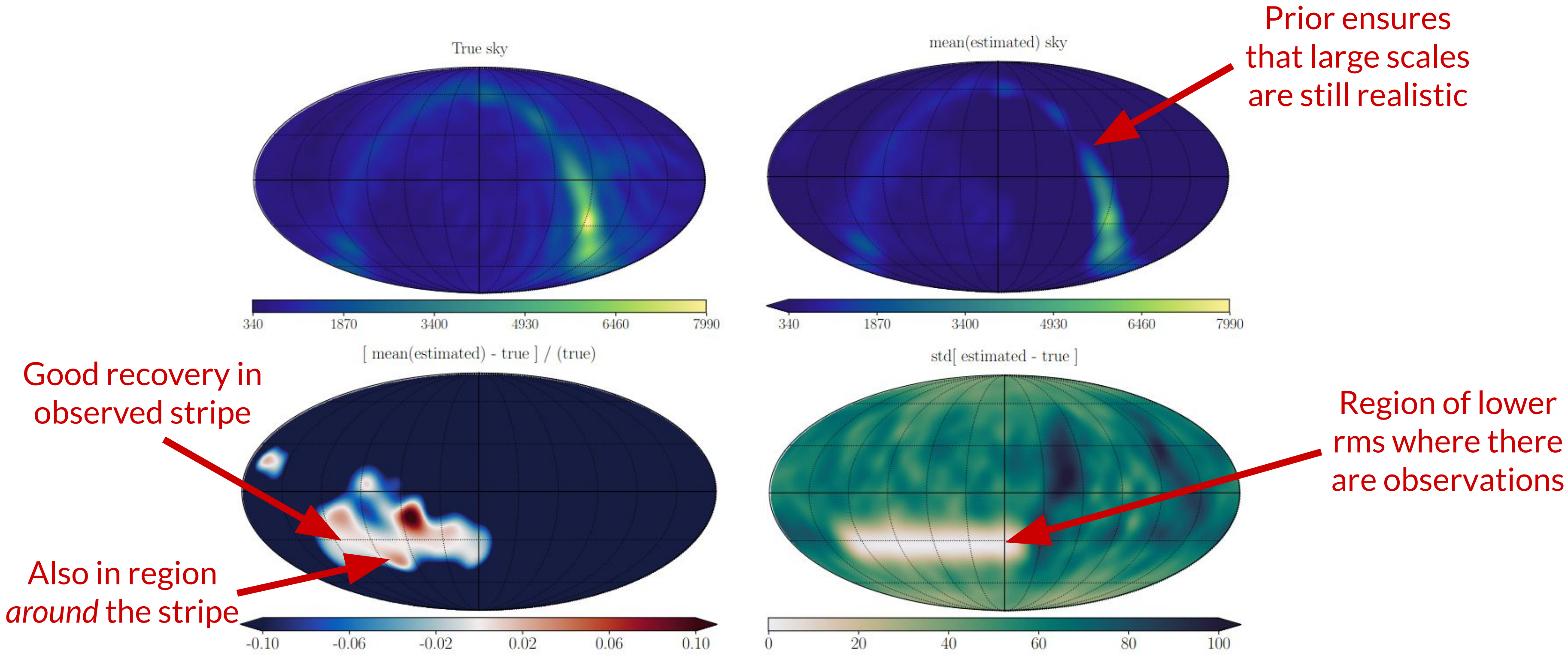
$$Y_\ell^m(\theta',\phi') = \sum_{m'=-\ell}^{\ell} Y_\ell^{m'}(\theta,\phi)\,\mathfrak{D}^\ell_{m'm}(\alpha,\beta,\gamma)$$

$$V_{ij}(\nu,t) = \sum_{\ell m} \delta V_{ij}^{\ell m}(\nu, t_{\rm ref}) \sum_{m'} \mathfrak{D}^\ell_{m m'}(t)\, a_{\ell m'}$$

Spherical harmonics (diffuse emission)

Glasscock et al. [in prep.]

- GCR: Draw samples of SH modes given data/priors $P(\mathbf{a} \mid \mathbf{d}, \mathbf{N}, \mathbf{a}_0, \mathbf{S})$



Summary

- Building a statistical model of the data allows us to treat sensitive analysis steps in a principled manner
- This will improve robustness / avoid signal loss
- **Hydra uses Gibbs sampling to make sampling tractable**

To get in touch: phil.bull@manchester.ac.uk

We are hiring!

- Postdoc in Radio Cosmology (deadline **22nd Jan**!) [jobs.manchester.ac.uk 27826]
- Senior Technical Specialist (Radio Antenna Dev.), apps online in ~1 week