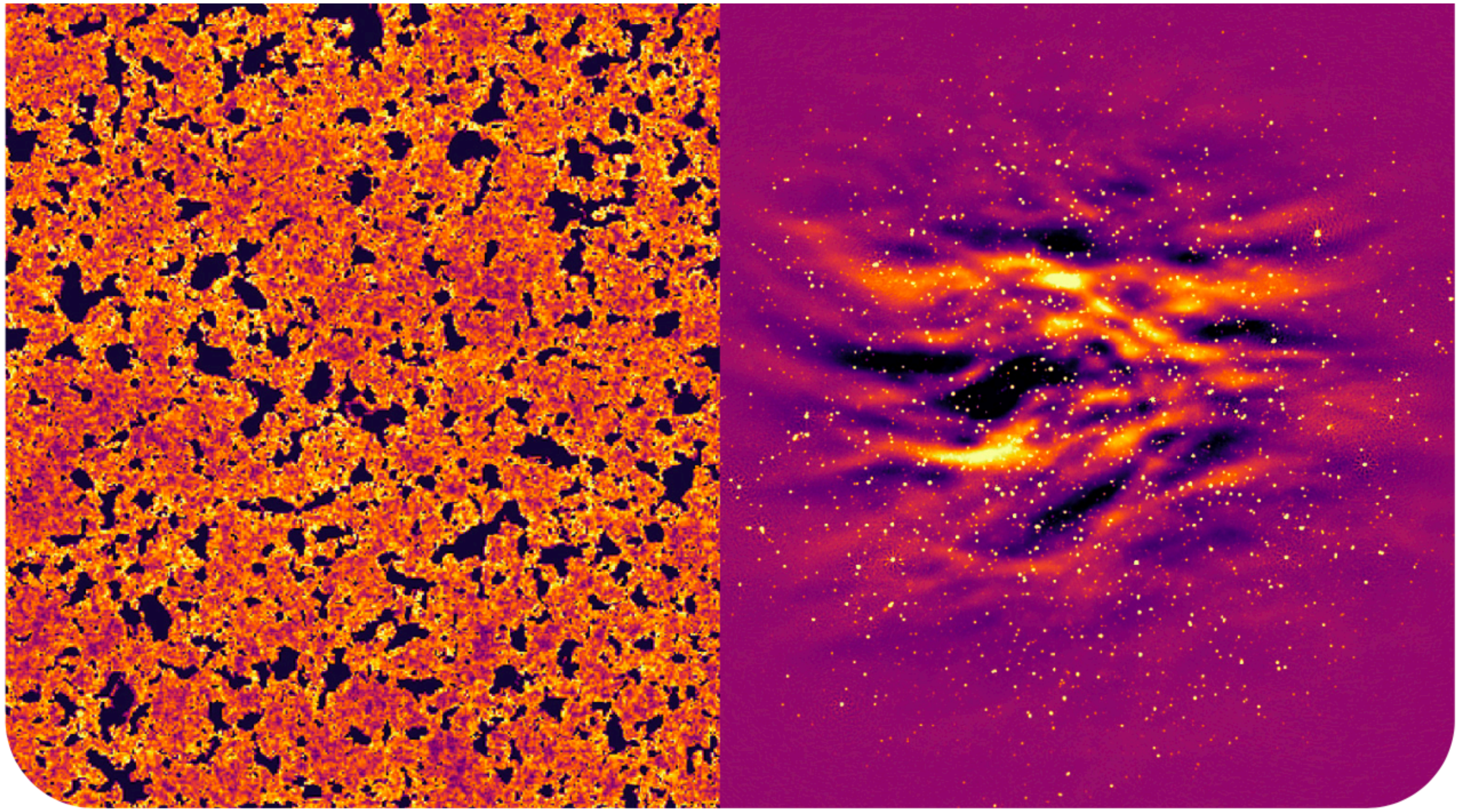
SDC3a discussion

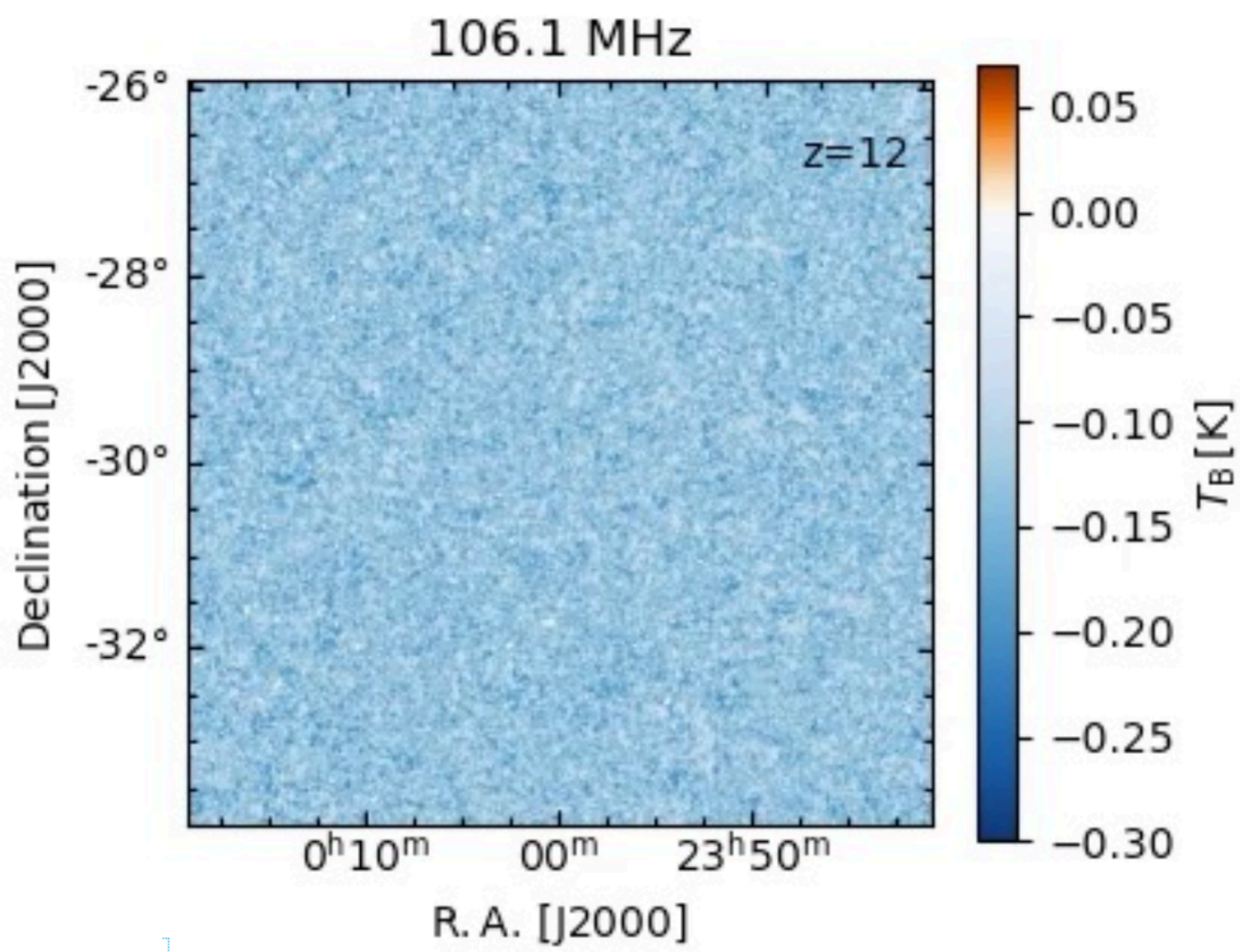


2024.7.19 Beijing

Science Data Challenge 3

Developed in collaboration with SKA EoR SWG members

- SDC3a "**Foregrounds**" (SDC3a; SWG Coordinators: C. Trott, V. Jelic)
  - **Foreground removal** exercise
  - SDC3a started 1 March 2023, closes 30 September 2023 **two-month delayed**
- SDC3b "**Inference**" (SDC3b; SWG Coordinators: A. Mesinger, G. Melema)
  - Extraction of **cosmological parameters**
  - SDC3b launching Q1 2024



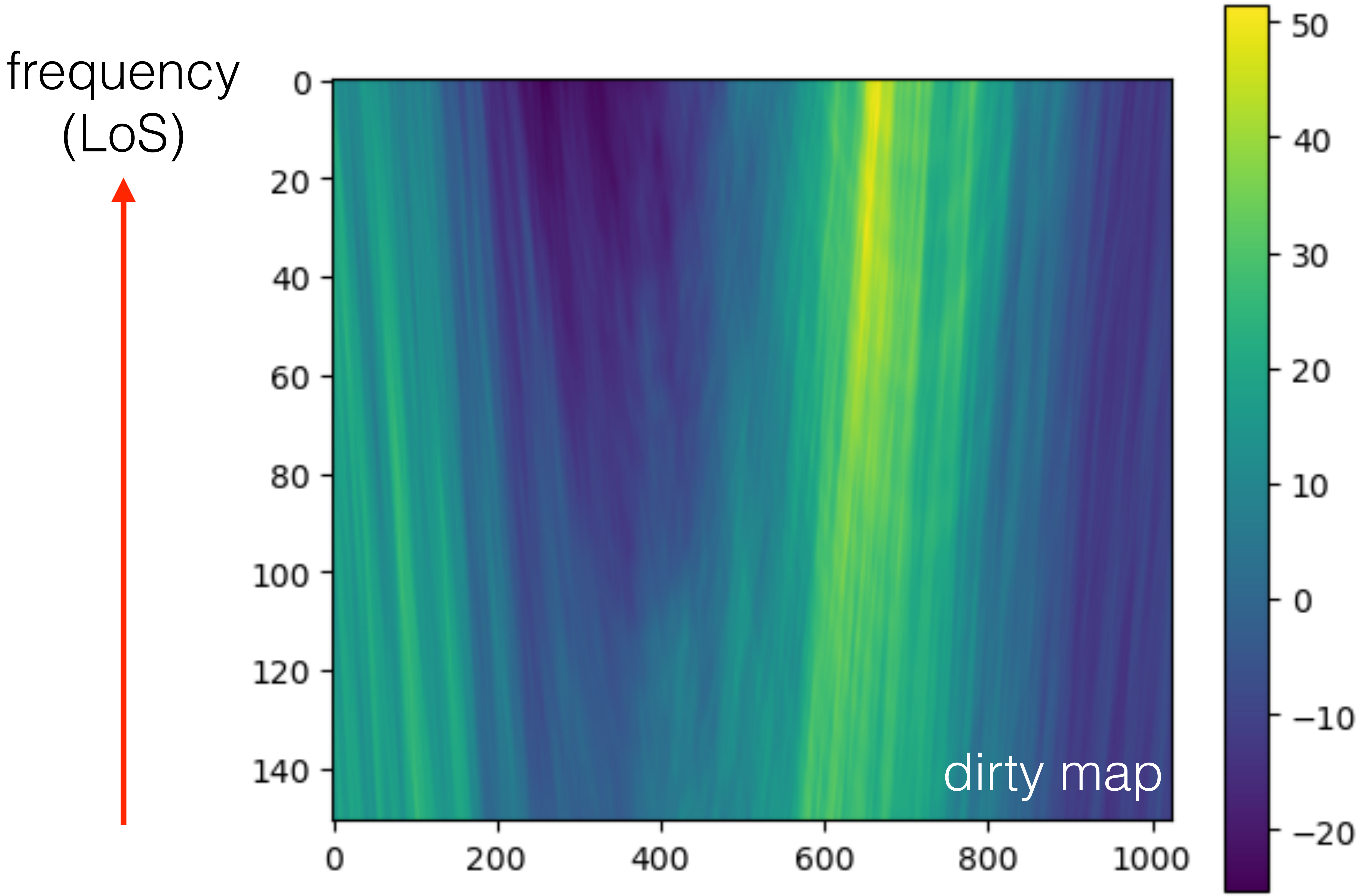
- The Challenge
  - Determine EoR power spectrum as function of scale and frequency
  - Cleaning foreground (10^5 times brighter)
  - Presence of various residual calibration errors: DI, DD & bandpass

Science Data Challenge 3a – Datasets

- General
  - Four-hour duration tracking observation
  - Thermal Noise Equivalent: 1000 hours
  - Field of View: One SKA1-Low pointing at RA, Dec = 0h, -30 deg

- Visibilities/images
  - Size: 7.5 TB (Visibilities)
  - Integration Time: 10 seconds
  - Channel Width: 100 kHz
  - Frequency Coverage: 106-196 MHz (6 cubes × 151 channels)
  - Image Cube (60 GB): 2048×2048, 16 arcsec pixels, natural/uniform weighting

Mode-mixing: PSF varied with frequency, breaking down the smoothness



We we learned so far for signal recovery:

■ **1000 hrs integration, i.e., 250 repetitions of such a four-hour track**

- noise is not important, as $\sigma_{\mathrm{noise}} < \sigma_{\mathrm{HI}}$
- systematic effects (250 repetitions of such a four-hour track), including Direction Dependent (DD) calibration error & Direction Independent (DI) gain calibration error also seems not important

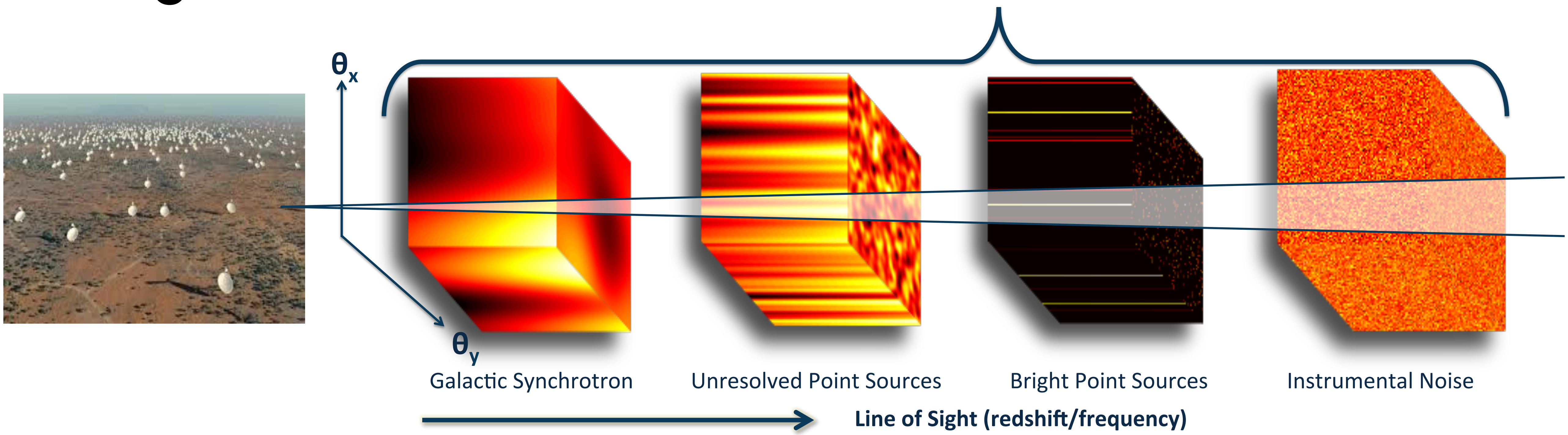
■ **mode-mixing is key challenge in foreground removal:**

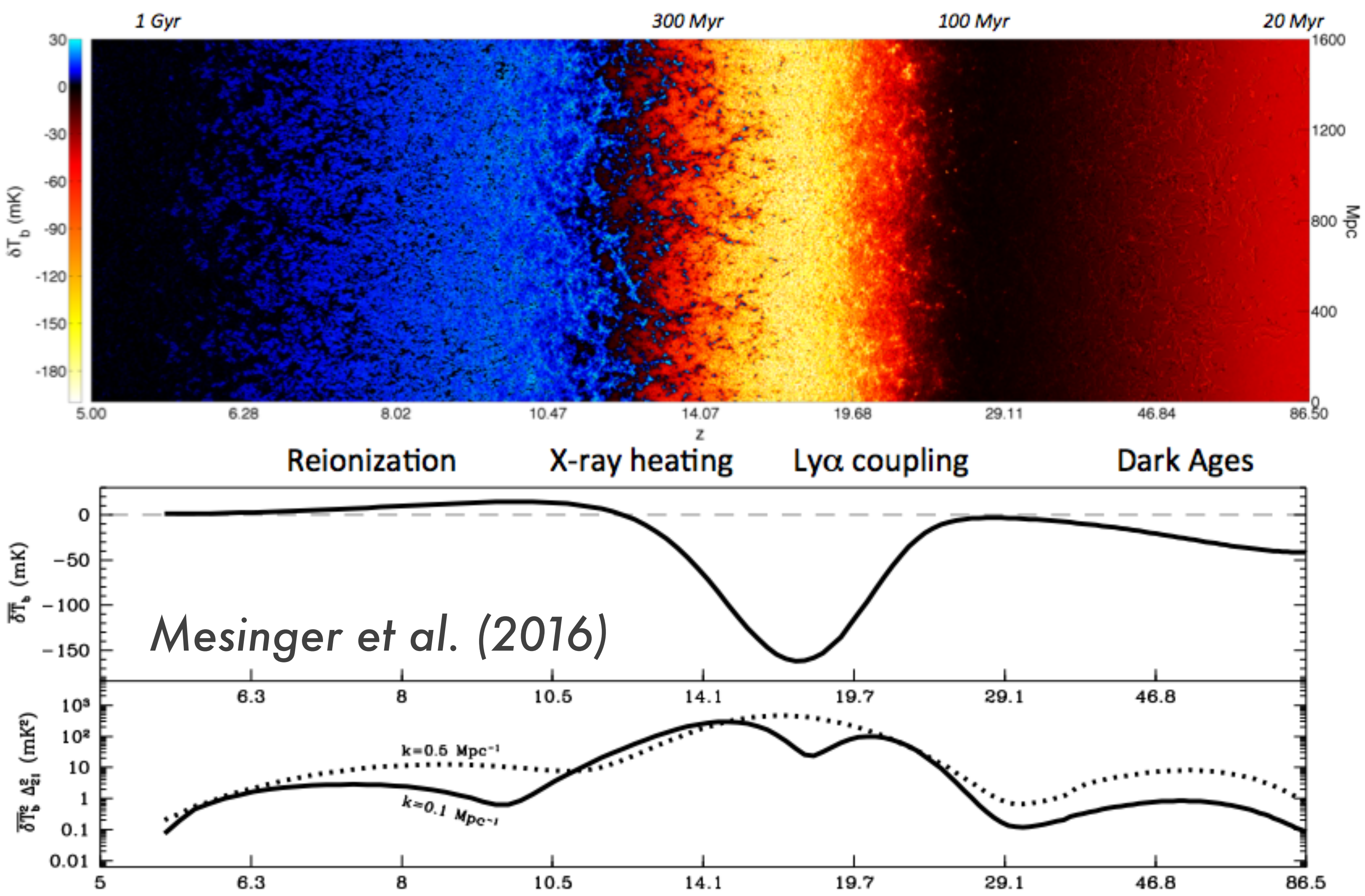
- if no instrumental effects, Fg removal can be done efficiently
- however, mode-mixing break down the smoothness

■ **beat down mode-mixing:**

- preprocessing, AI (deconvolution), ….
- using visibility or image? which one is better?

Foreground contamination





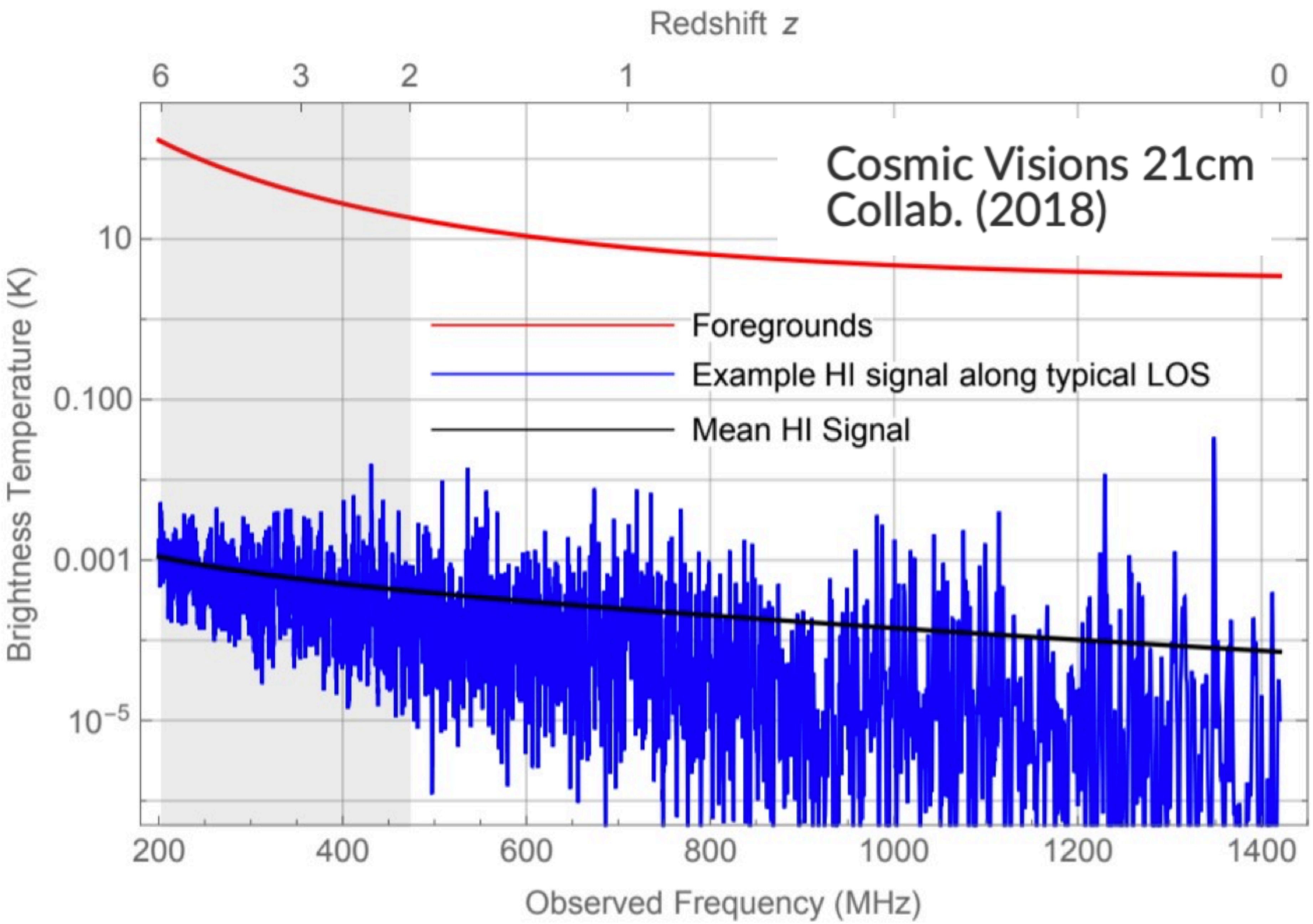
Strong foreground:

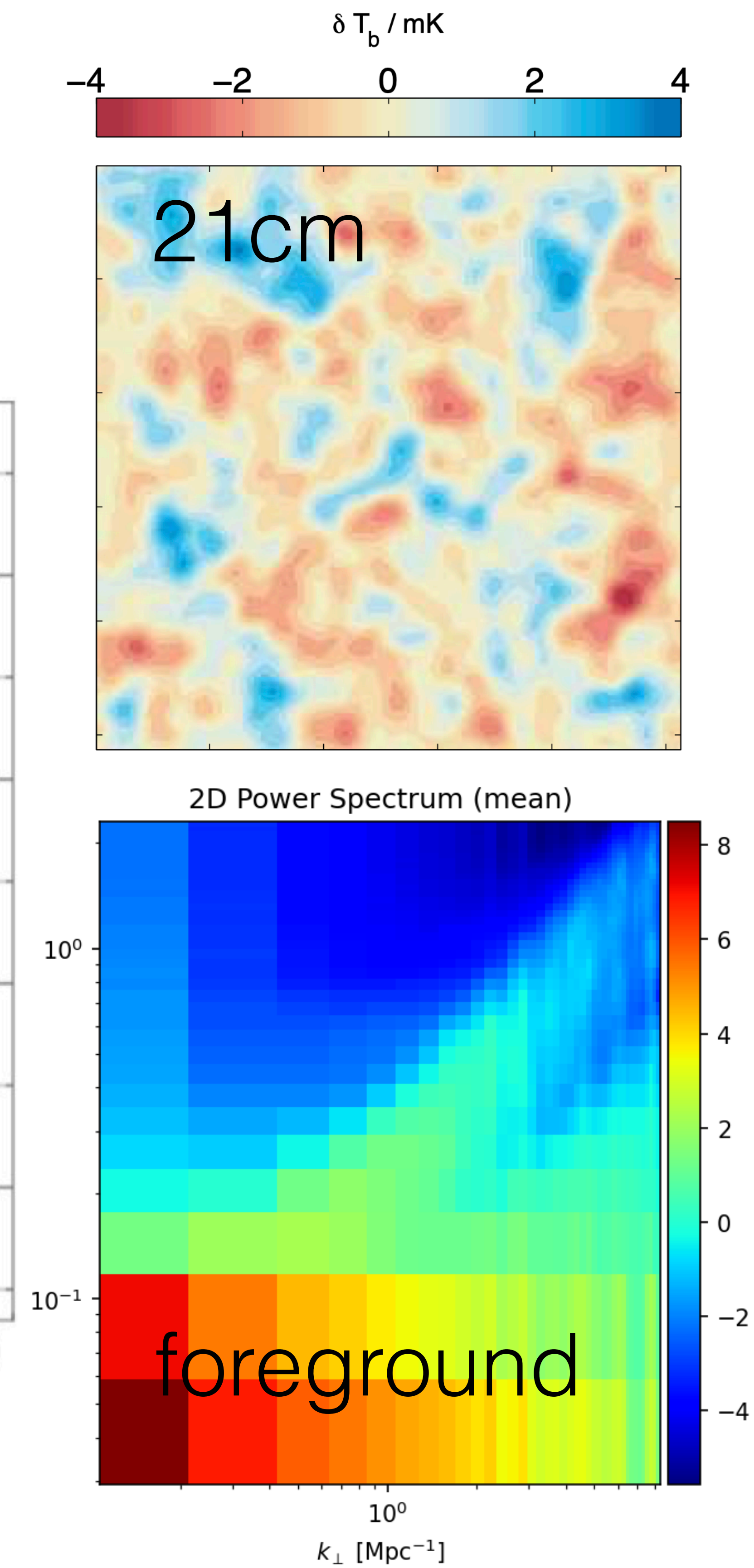
$$T_{\rm foreground} \gg 10 \text{ K}$$

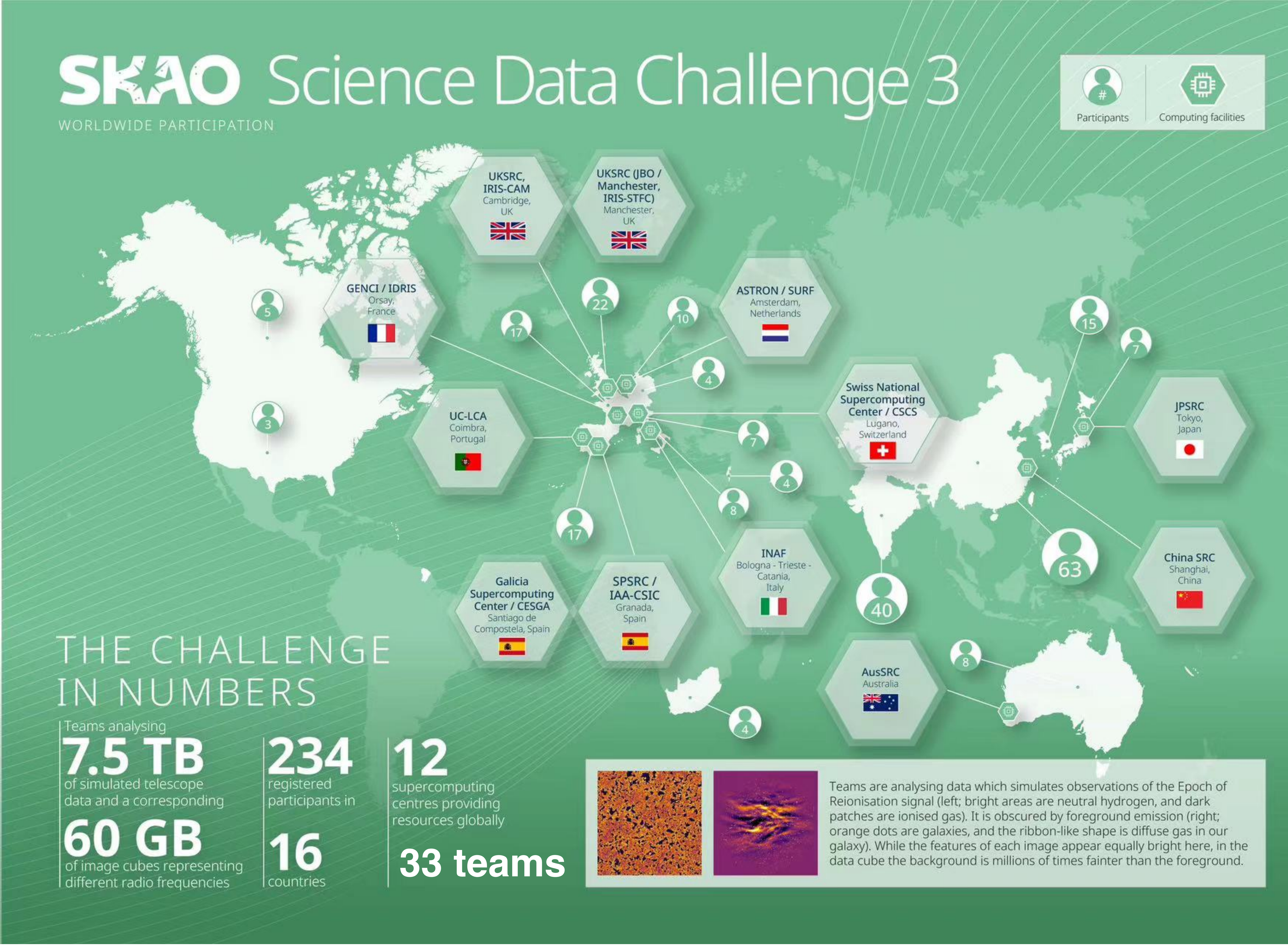
Faint signal:

$$T_{\rm 21cm} \sim 10 \text{ mK}$$

The key to separating out foregrounds: their spectral smoothness







SKAO Science Data Challenge 3
WORLDWIDE PARTICIPATION
Participants
Computing facilities
UKSRC, IRIS-CAM Cambridge, UK
UKSRC (JBO / Manchester, IRIS-STFC) Manchester, UK
GENCI / IDRIS Orsay, France
ASTRON / SURF Amsterdam, Netherlands
UC-LCA Coimbra, Portugal
Swiss National Supercomputing Center / CSCS Lugano, Switzerland
JPSRC Tokyo, Japan
Galicia Supercomputing Center / CESGA Santiago de Compostela, Spain
SPSRC / IAA-CSIC Granada, Spain
INAF Bologna - Trieste - Catania, Italy
China SRC Shanghai, China
AusSRC Australia
5
3
17
22
10
4
7
4
8
17
15
7
63
40
8
4
THE CHALLENGE IN NUMBERS
Teams analysing
7.5 TB
of simulated telescope data and a corresponding
60 GB
of image cubes representing different radio frequencies
234
registered participants in
16
countries
12
supercomputing centres providing resources globally
33 teams
Teams are analysing data which simulates observations of the Epoch of Reionisation signal (left; bright areas are neutral hydrogen, and dark patches are ionised gas). It is obscured by foreground emission (right; orange dots are galaxies, and the ribbon-like shape is diffuse gas in our galaxy). While the features of each image appear equally bright here, in the data cube the background is millions of times fainter than the foreground.

SDC3a teams and FG-cleaning approaches

10 top teams:

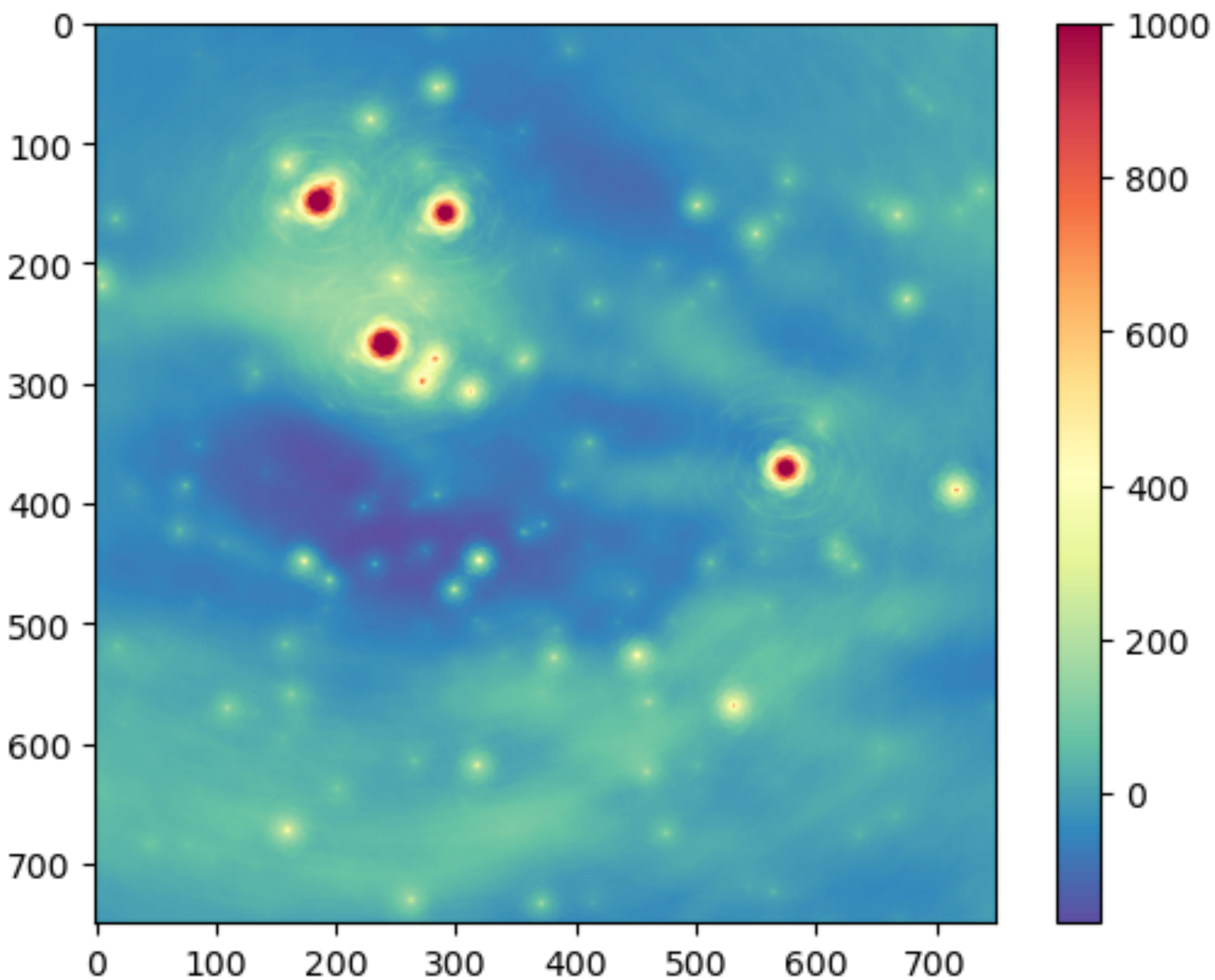
- **HIMALAYA** (SYSU): **reconvolution + transfer function**
- **DOTSS-21 (ML-GPR; Advanced_ML-GPR; Avoidance)**: Machine-learning+Gaussian Process Regression (GPR) to model the foregrounds to separate them from the 21-cm signal
- **ERWA**: neural network applied in image; China SRC supported
- **Shuimu-Tianlai** (Tsinghua-NAOC): oriented singular value decomposition (O-SVD)
- **Wizards of Oz 3D**: improving the sky model to improve the quality of the sky-model-subtracted visibilities; 4th-order polynomial fitting; Pawsey Supercomputer supported
- **Akashganga**: 2nd-order polynomial fitting+GPR
- **REACTOR**: PI-AstroDeconv +PCA
- **SKACH**: U-shaped 3D convolutional neural network to remove residual foreground; polynomial fitting in uv space

- **KUSANAGI**: PCA+neural network; **Cantabrigians**: Bayesian GPR model; **Hausos**: CNN-based approach; **Nottingham-Imperial**: point-source model+ FastICA; **Pisano Galaxy Moppers**: foreground avoidance; **Foregrounds-FRIENDS**: PCA+neural network; **HAMSTER**: delay spectrum approach; **KORSDC**: PSF deconvolution+ICA **SROT**: GPR …

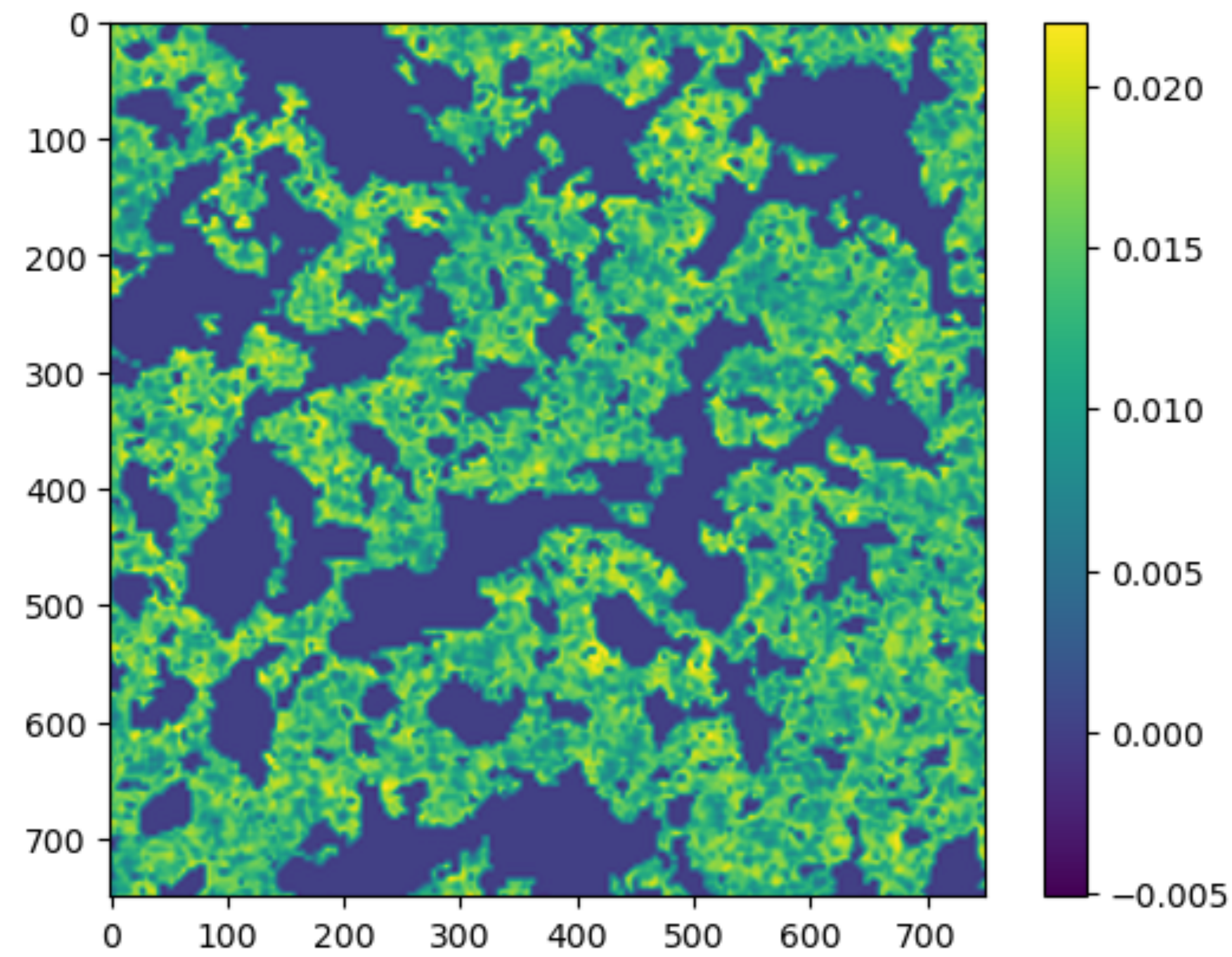
Challenge: FG contamination

five orders of magnitude brighter; FG removal accuracy of at least 1 in 10,000 required !!!

Raw SDC3a image
(natural weighting)



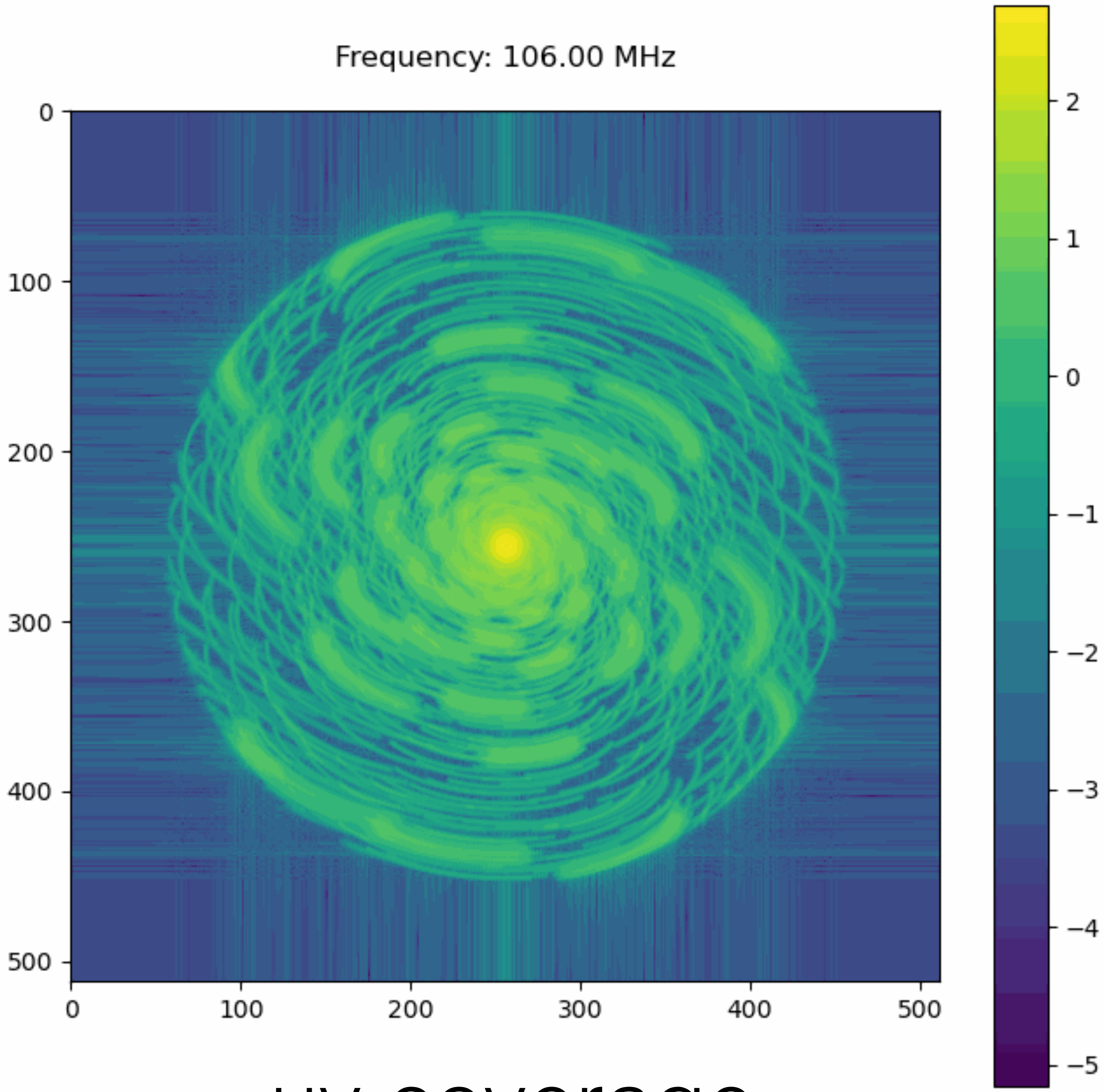
True EOR



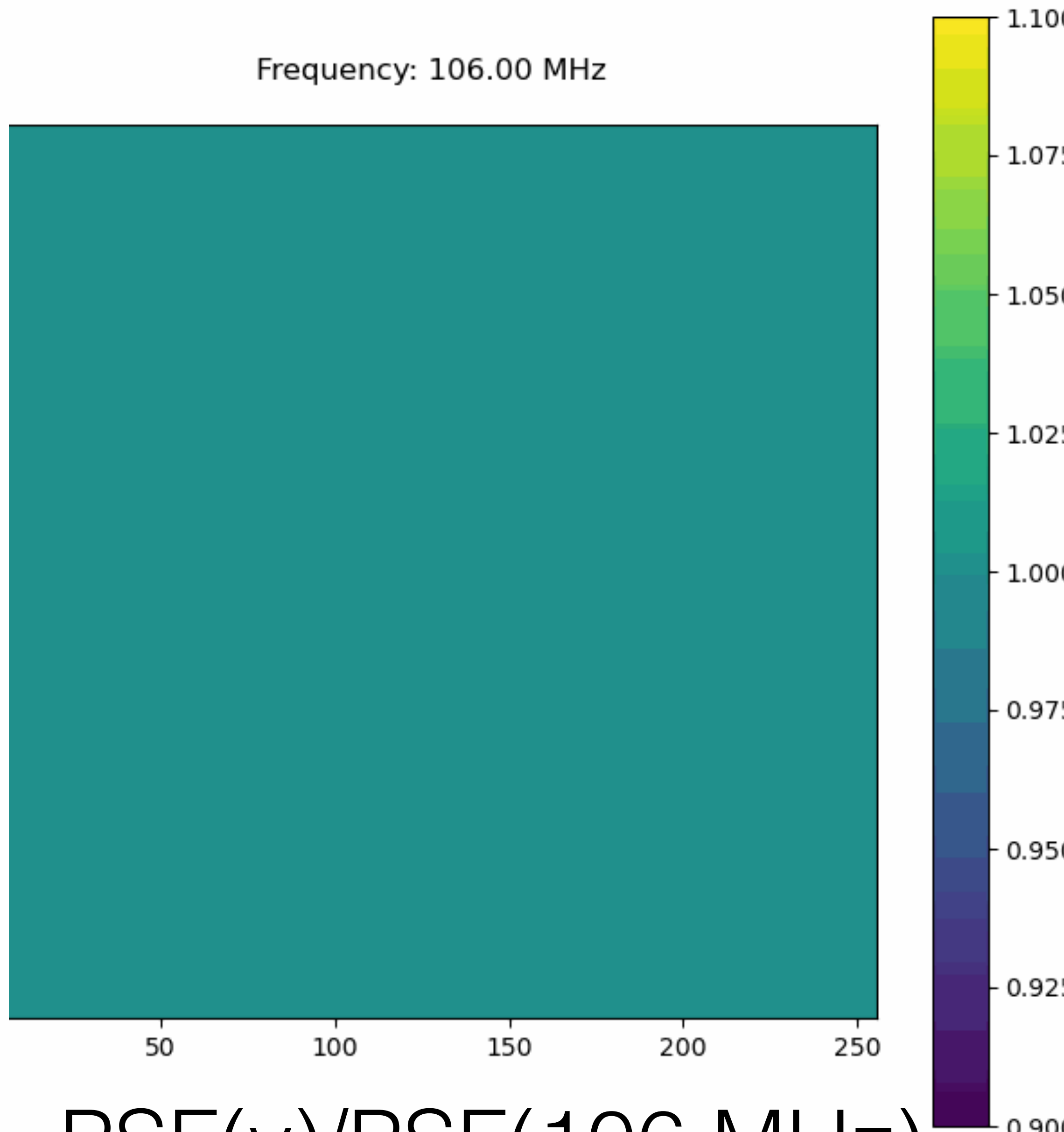
Key Challenge: frequency-dependent PSF

→ uv coverage changed with frequency

→ notable change for PSF (esp. its sidelobe) with frequency



uv coverage

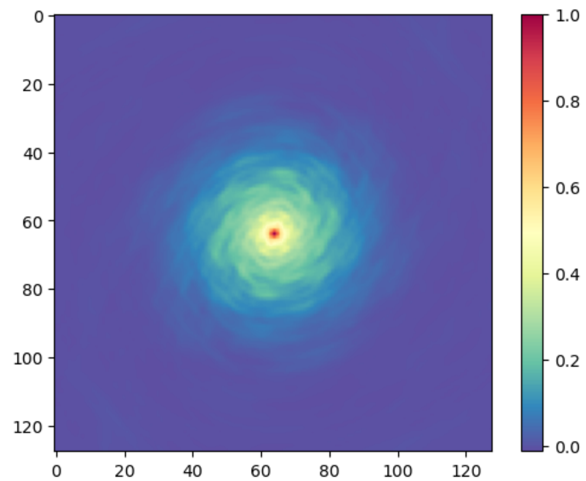


PSF(v)/PSF(106 MHz)

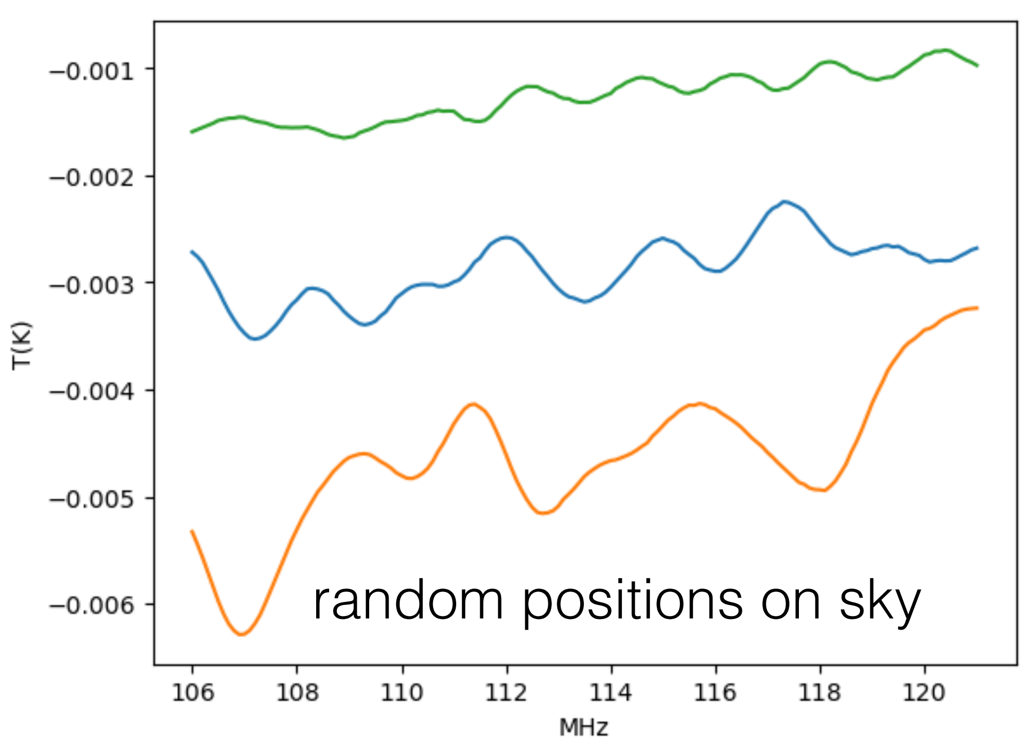
Key Challenge: frequency-dependent PSF

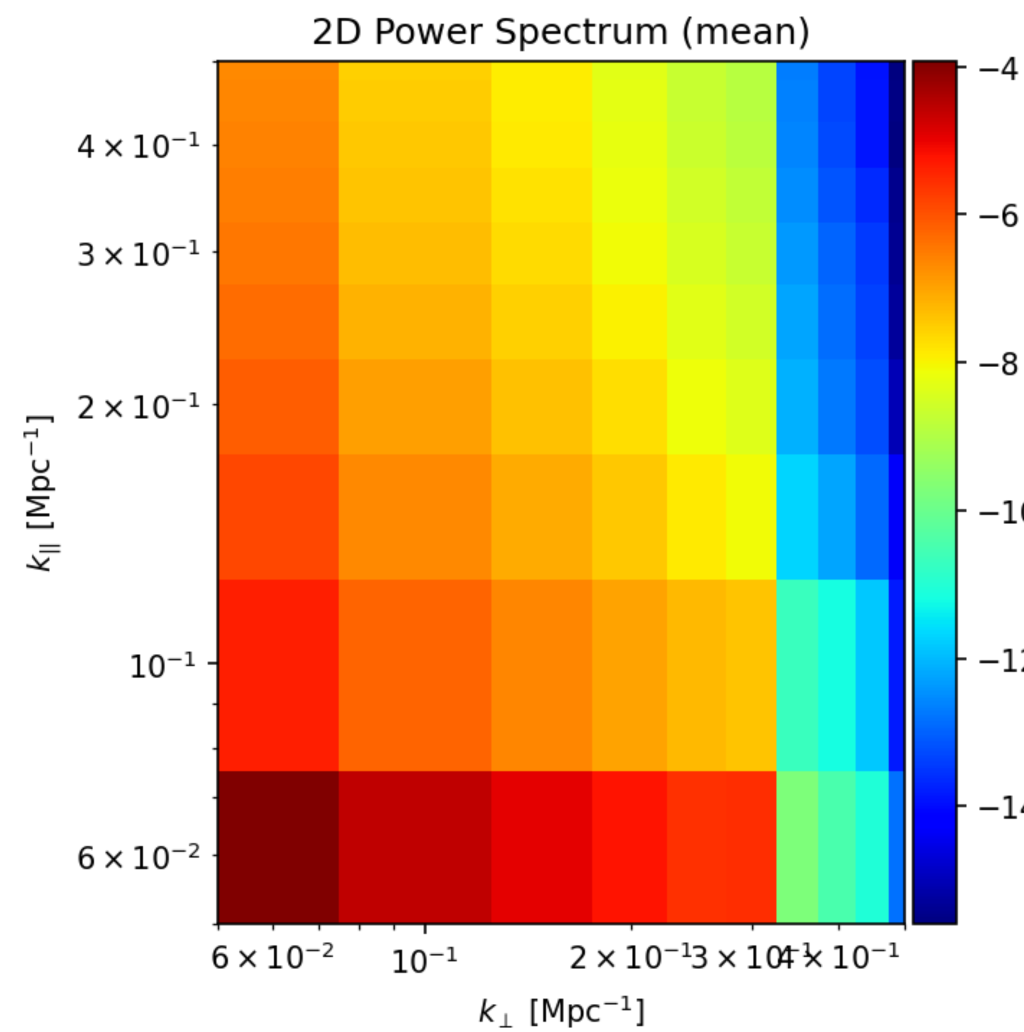
→ mode-mixing → producing non-smooth FG spectrum

- set up a point source on the sky without frequency dependence



- following interferometric observation, the spectrum undergoes an oscillation





- mode mixing: spatial modes mixed up with frequency modes

Challenges:

- "mode-mixing" breaks the smoothing and prevents foreground removal
- PSF deconvolution — an ill-posed inverse problem; achieving the desired precision of 1 in 10,000 is not feasible

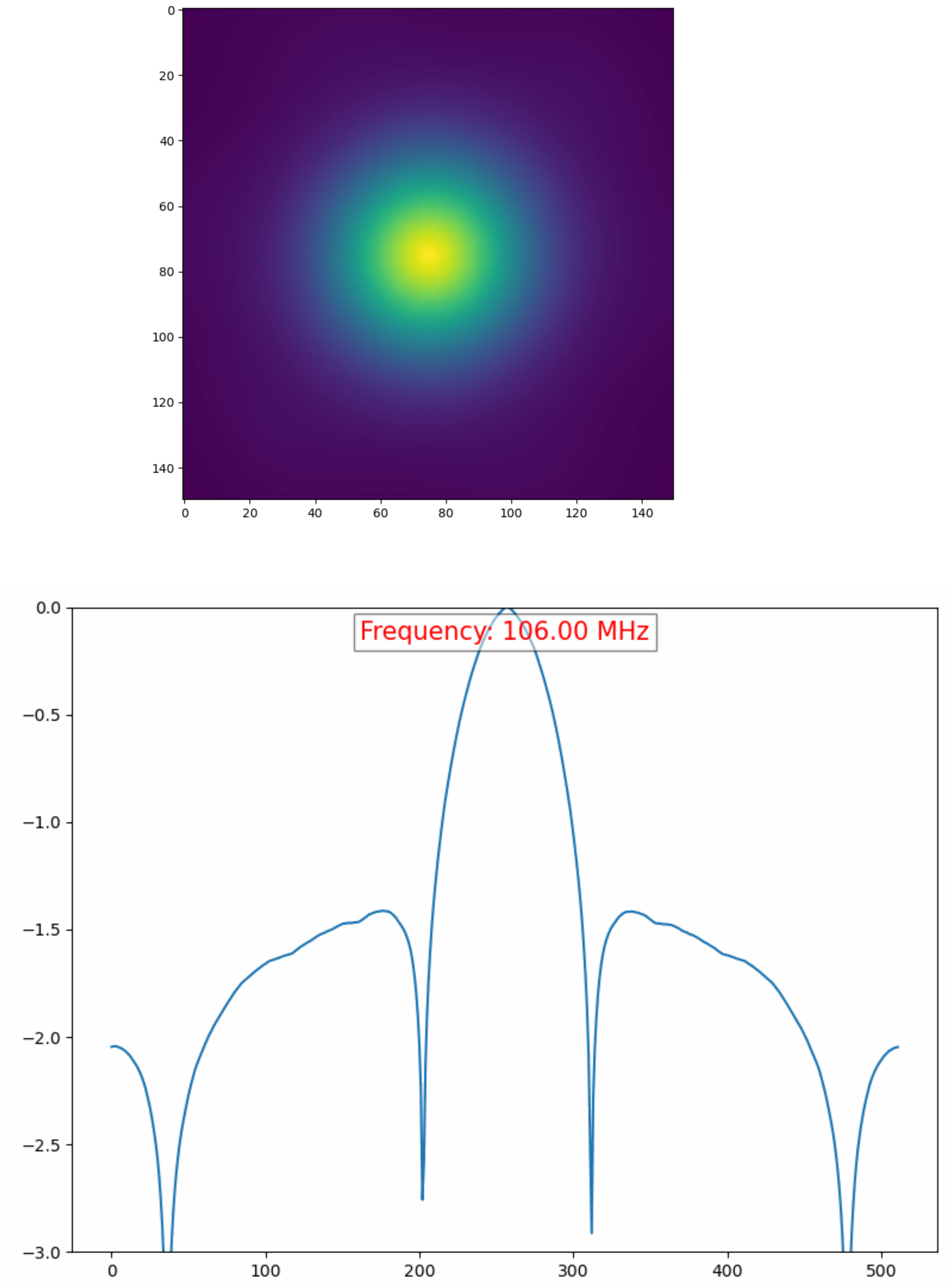
Solution: other way around?

Counterintuitive approach:

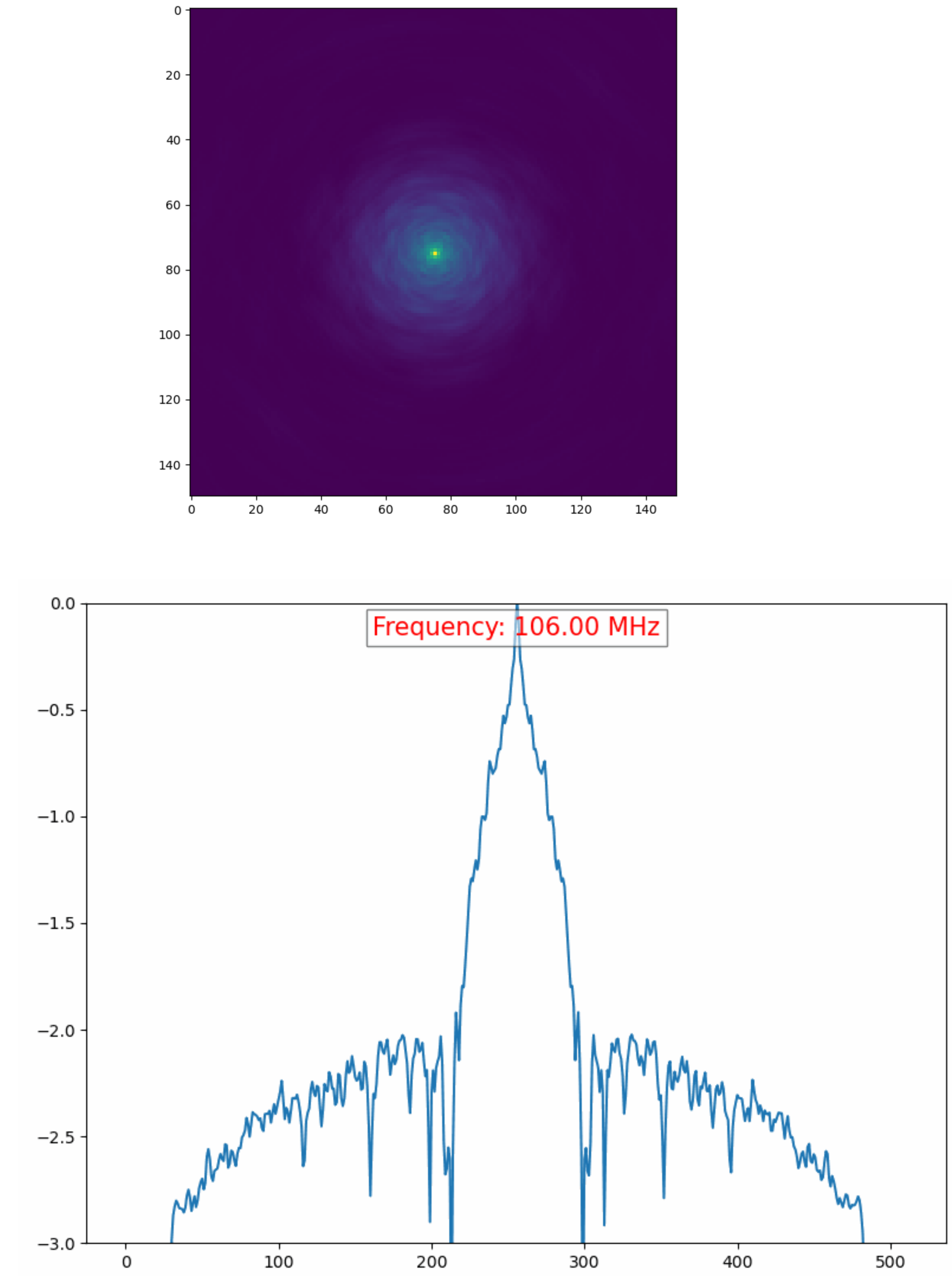
✓ reconvolution rather than deconvolution

Reduction of mode mixing: suppressing high-$k_\perp$ modes that vary significantly with frequency, which dominate the effect of mode mixing.

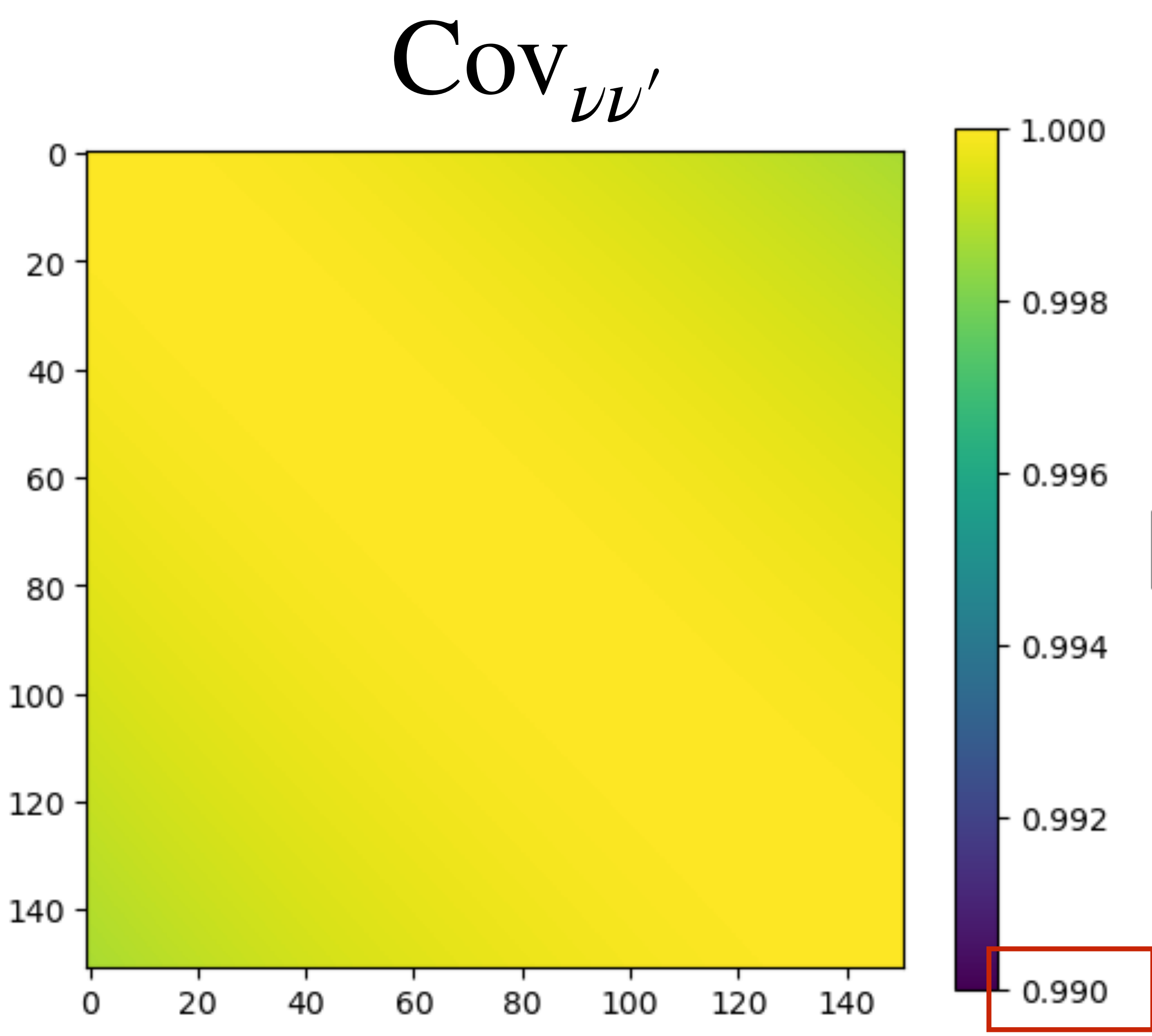
$PSF \otimes PSF$

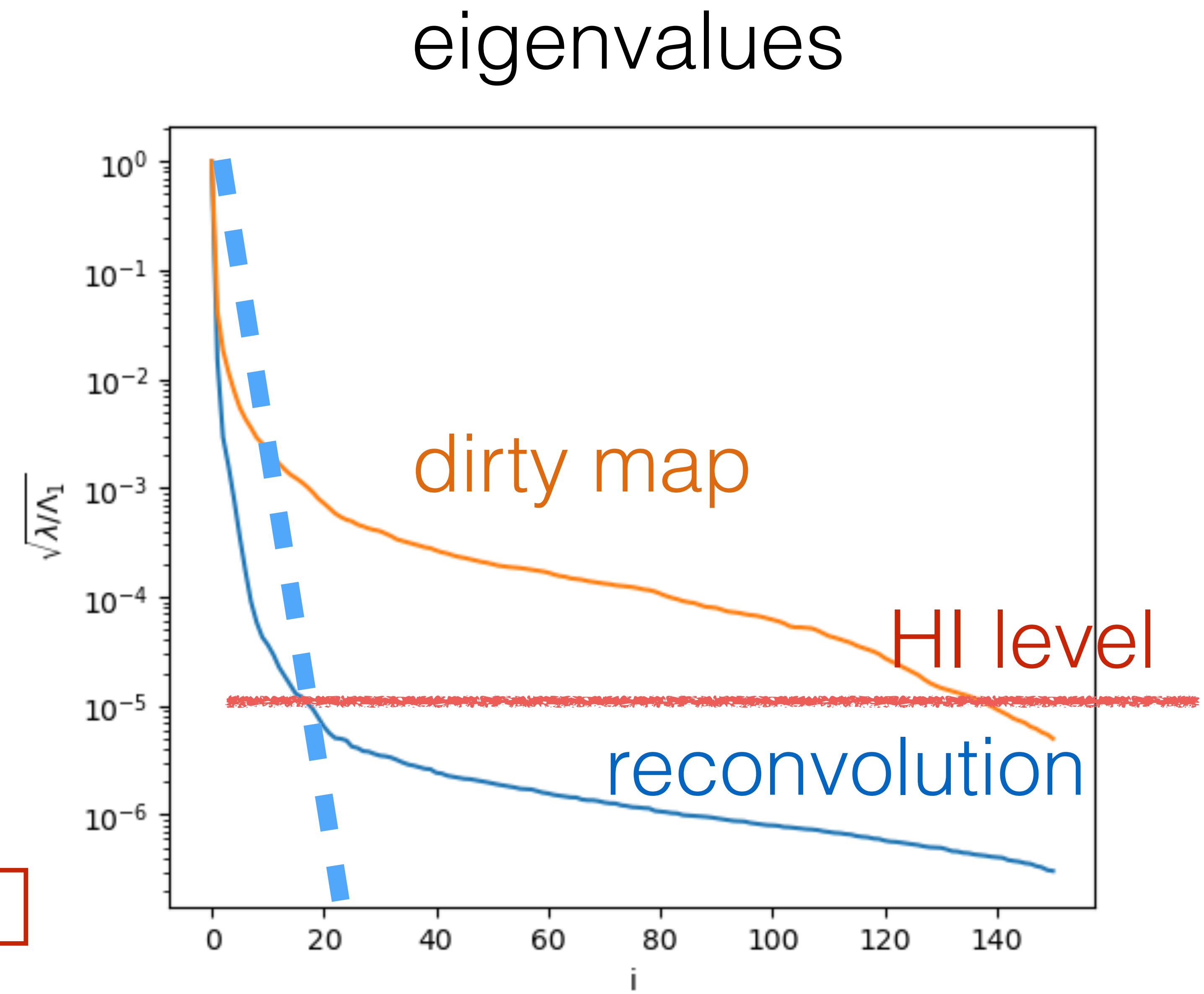


PSF



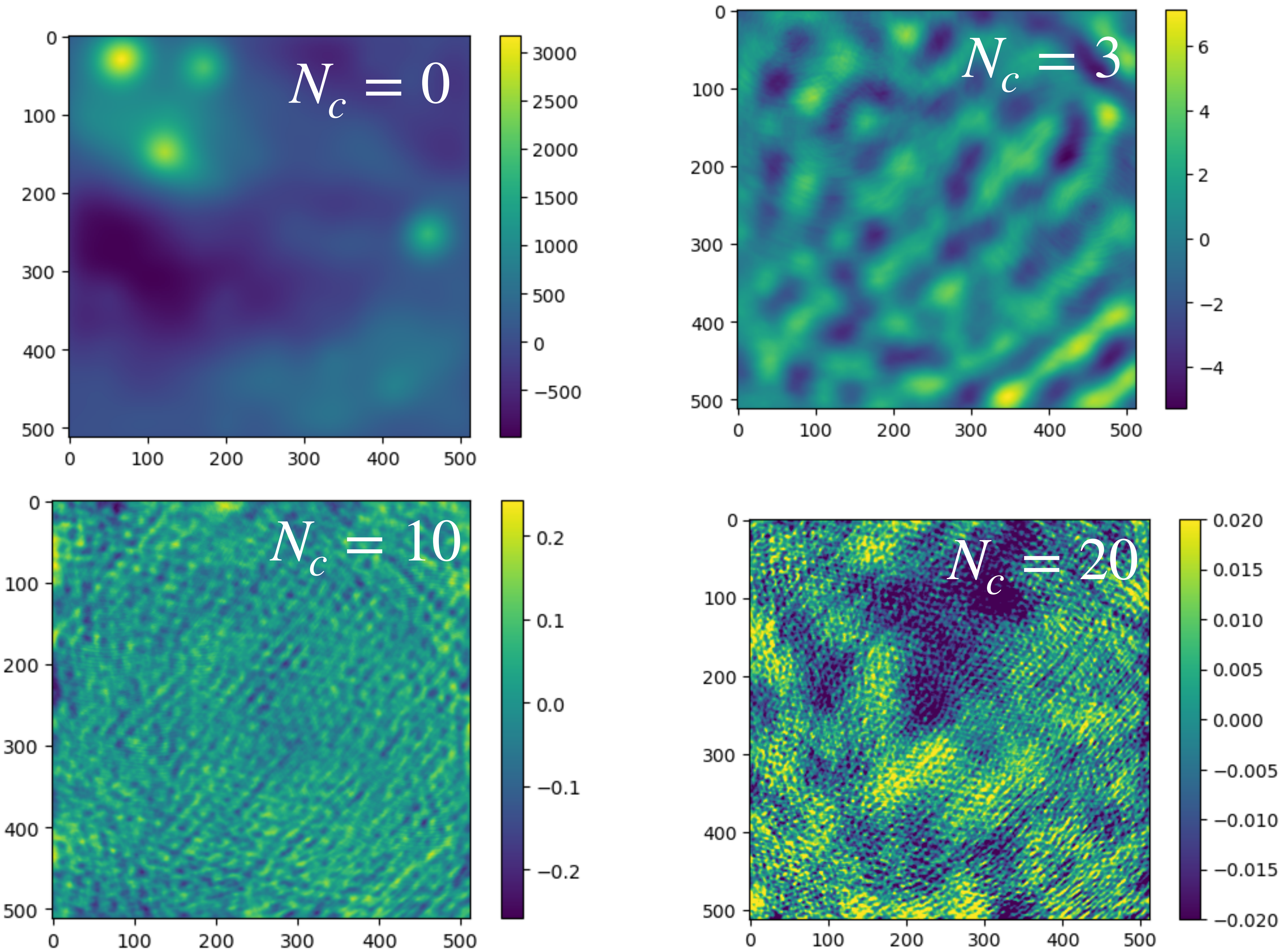
Smoothness recovered by "reconvolution"





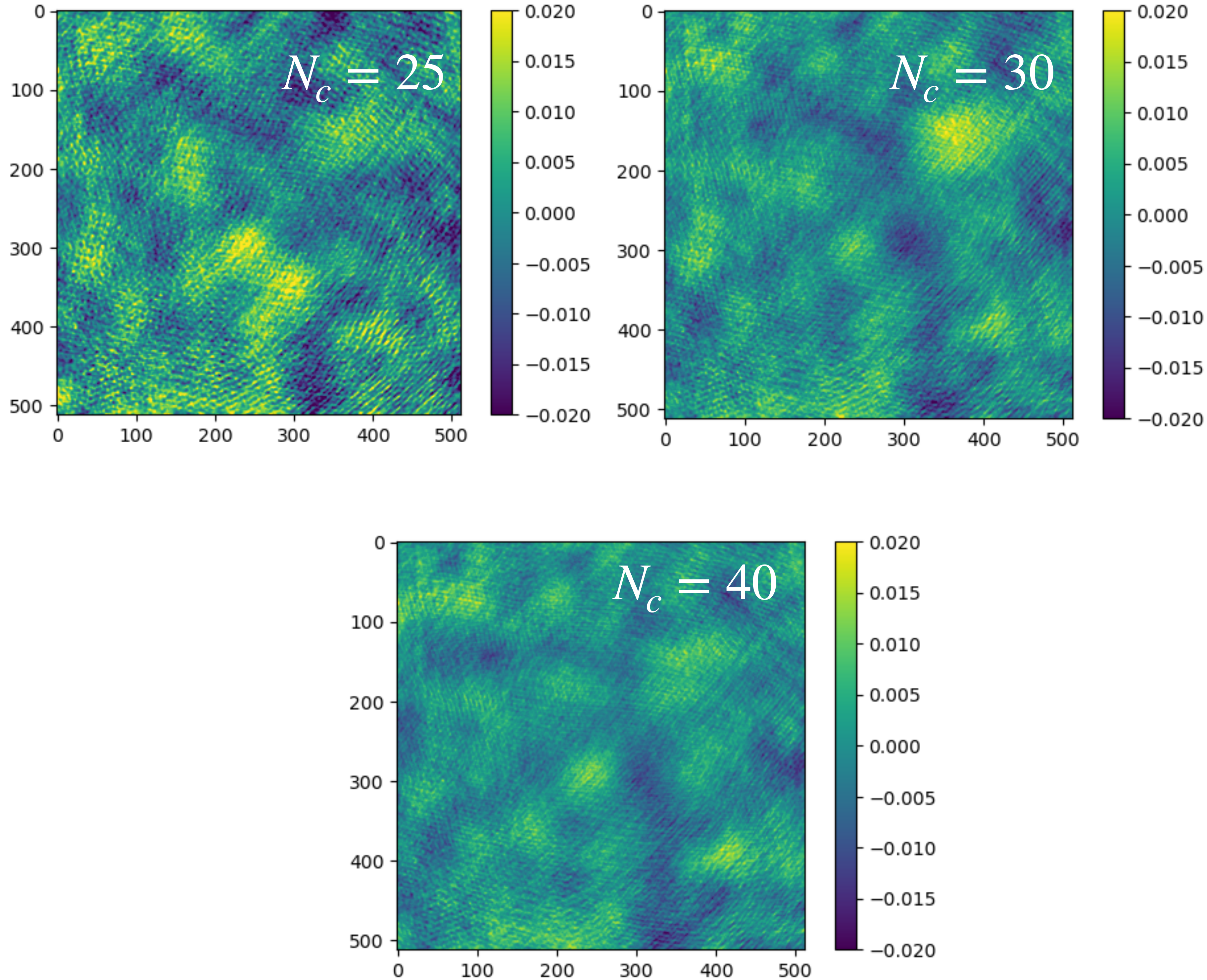
- project out the first ~20 modes - remove foreground perfectly!

Residuals after PCA projection

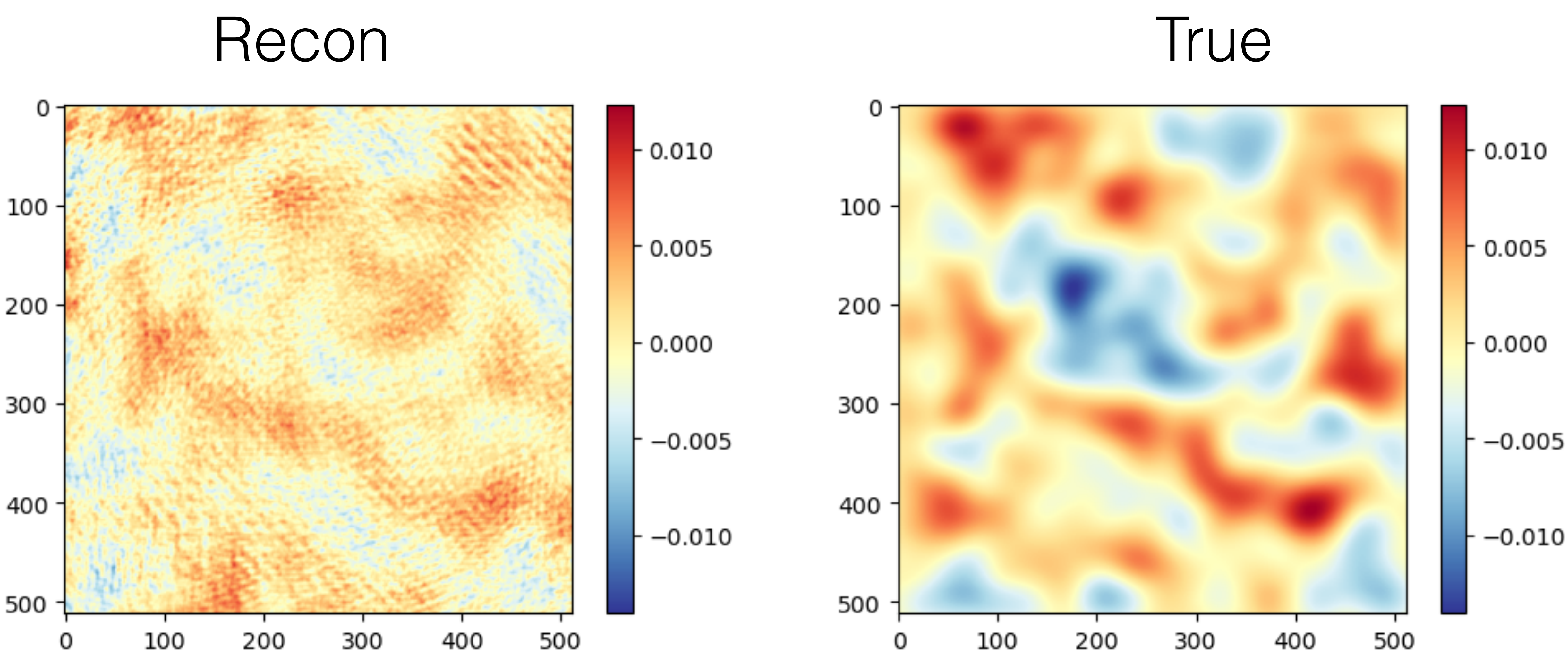


Convergence test:

- further modes projected out
- yet the residual patterns remained largely unaltered



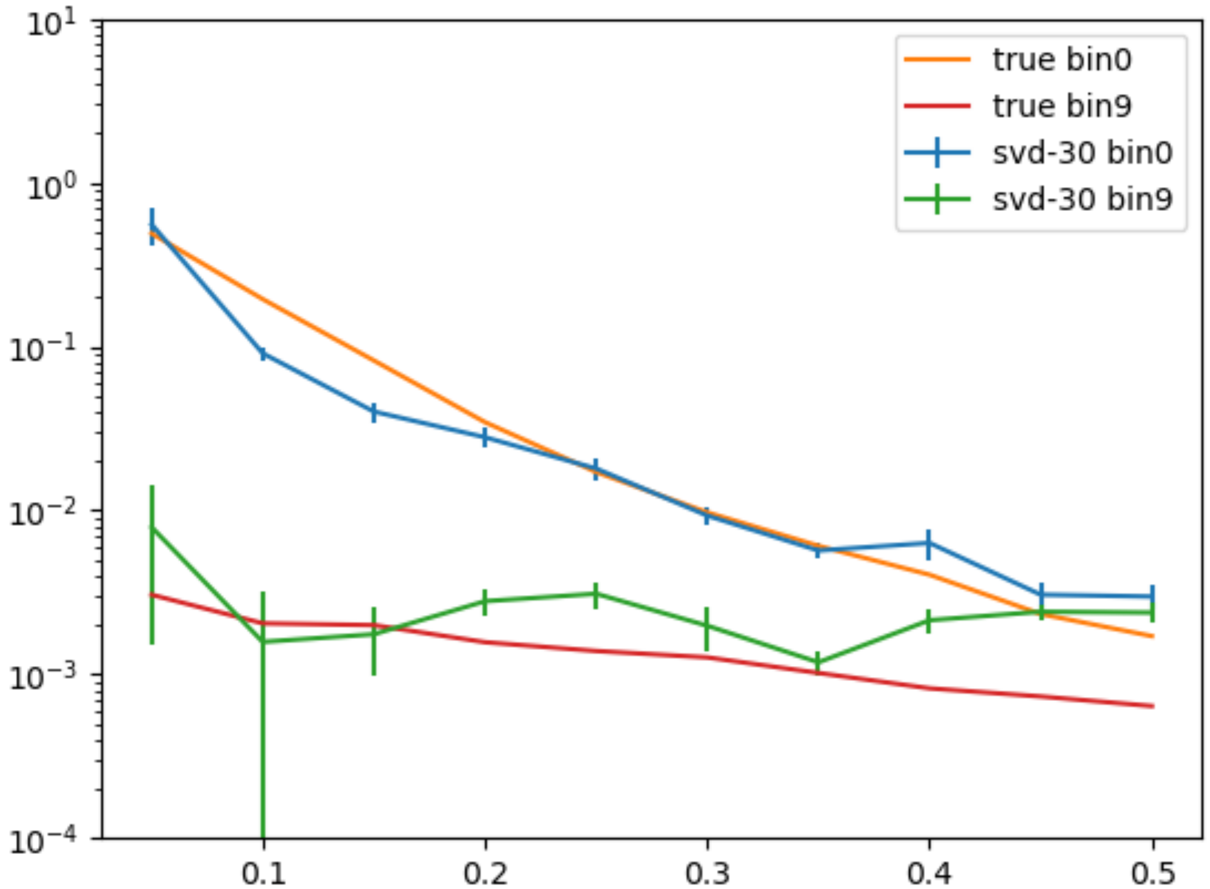
Compared with 21cm image@ 121 MHz

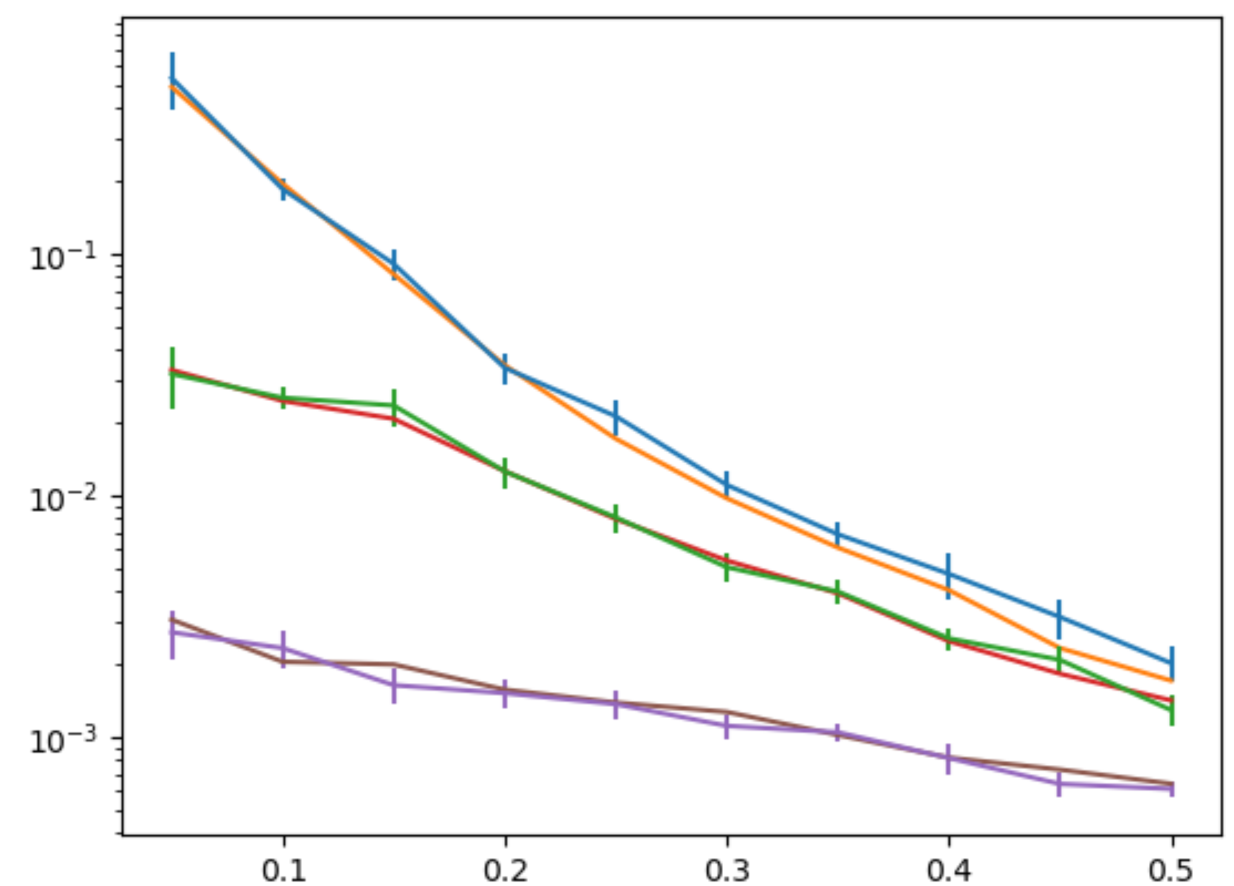


- Correlation coefficient of image ~ 0.5–0.6
- Direct imaging of EoR at large scales is promising!

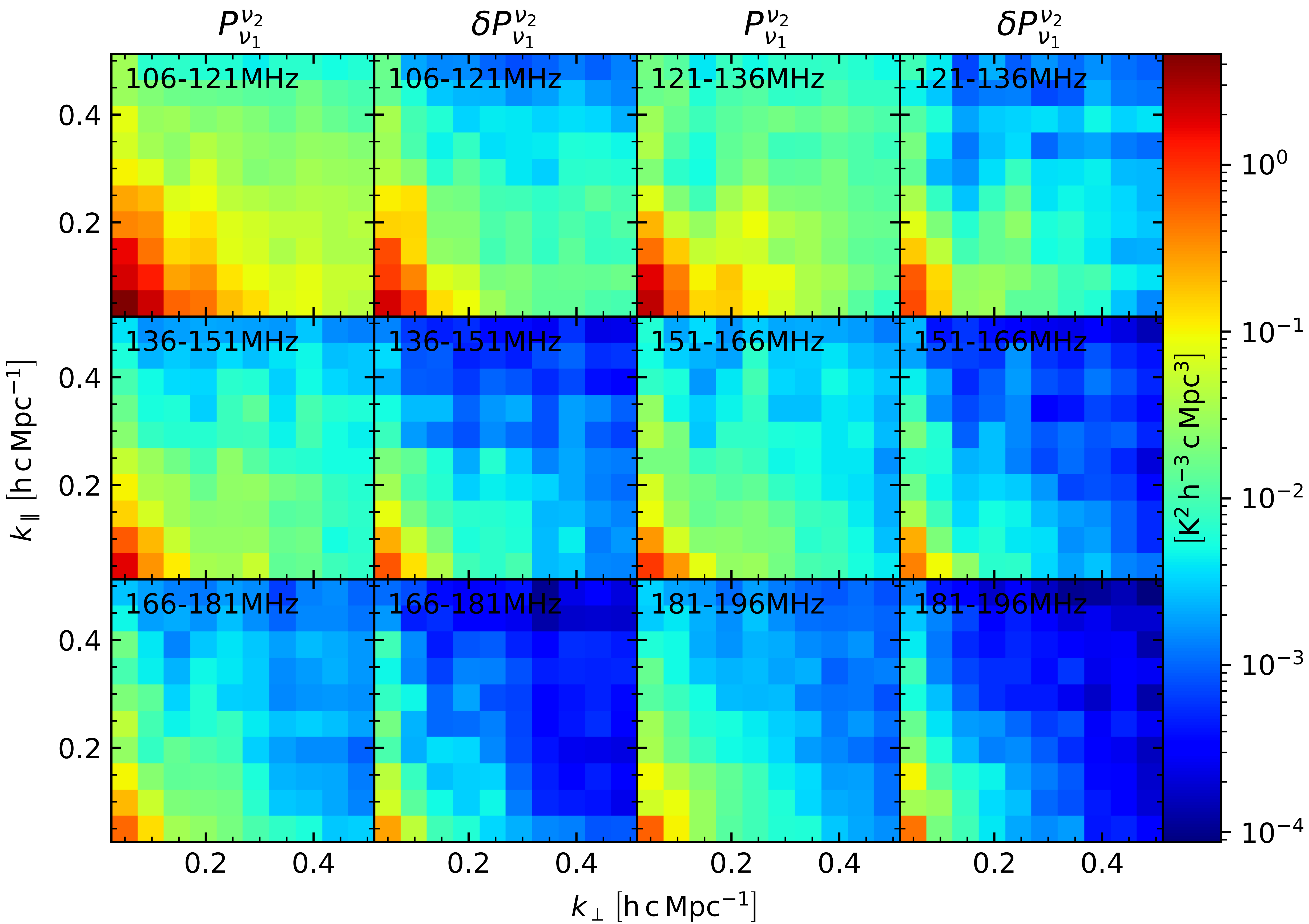
Transfer function: corrected for amplitude

- transfer function — $T(k_{\parallel}, k_{\perp}) \equiv P_{21}^{\mathrm{true}}(k_{\parallel}, k_{\perp}) / P_{21}^{\mathrm{rec}}(k_{\parallel}, k_{\perp})$
- correction for the amplitude of missing modes generated by PCA and various instrumental effects.
- use mock realizations to estimate its mean and std
- EoR reconstruction: $P^{\mathrm{reonc}} = T(k_{\parallel}, k_{\perp}) \times P^{\mathrm{svd}}$





Submitted results—HIMALAYA



Rank	Team	Team Affiliations	Score
1	HIMALAYA	China	74758.5
2	DOTSS-21cm_ML-GPR	NL, DE, FR, IT, USA	71573.2
3	DOTSS-21cm_Advanced_ML-GPR	NL, DE, FR, IT, USA	71135.0
4	ERWA	China	63670.3
5	DOTSS-21cm_Avoidance	NL, DE, FR, IT, USA	51888.8
6	Shuimu-Tianlai	China	43421.7
7	Wizards_of_Oz_3D	Australia	33295.4
8	Akashganga	India, Israel	31864.5
9	REACTOR	China	21888.3
10	SKACH	Switzerland, Italy	12103.4
	KUSANAGI	Japan, China, Australia	
	Cantabrigians	UK	
	Hausos	China, France, Italy	
	KUSANAGIb	Japan, China, Australia	
	Nottingham-Imperial	UK	
	Pisano_Galaxy_Moppers	Italy, USA	
	HAMSTER	UK, South Africa	
	Foregrounds-FRIENDS	Spain, France	
	KORSDC	South Korea	
	SROT	India	

Collaborative paper in preparation …

Square Kilometre Array Science Data Challenge 3a: foreground removal for an EoR experiment

TBC

8 April 2024

ABSTRACT

Key words:

1 INTRODUCTION

2 DATA SIMULATION

The SDC3a simulation was undertaken with the OSKAR (Dulwich et al. 2009, https://ska-telescope.gitlab.io/sim/oskar/) package. This makes use of a telescope model and a sky model to generate visibilities with a specified sampling in time and frequency. There is provision to include a variety of instrumental errors into the simulation, which we refer to as the error model.

2.1 Telescope Model

The basic telescope model makes use of the SKA-Low configuration of 512 stations (SKA-TEL-SKO-0000422_04_SKA1Low_Configuration_Coordinates). The station layout is the so-called *Vogel* layout, a one arm spiral configuration with a uniform areal density of antennas and a maximally diverse azimuthal sampling (Vogel 1979). Only a single station layout has been adopted to reduce computational expense, rather than a full set of 512 diverse station layouts. The implication is that the effective station beam for all visibilities is simply the auto-correlation beam of this one station layout. The most important consequence of this simplification is an elevated far side-lobe response. An attempt to compensate for this effect is made within the sky model definition.

2.2 Sky Model

within the far side-lobe regime of a tracking observation, they would normally be modelled and removed with a so-called "de-mixing" process within a calibration and imaging pipeline. We have consequently attenuated the amplitude of all such sources to simulate both the additional signal attenuation that a cross-correlation station beam would provide (relative to the simplified auto-correlation beam that has been used) as well as a partially successful de-mixing process. The assumed net attenuation due to both of these effects was taken to be a factor of 10^{-3}. This is in excess of the attenuation that is subsequently provided by OSKAR via the auto-correlation station beam defined in the Telescope Model. Each source in the model is represented by an elliptical Gaussian approximation spatially and with a frequency-dependent flux density determined from an amplitude and spectral index defined at some reference frequency.

2.2.2 *Inner sky model*

The inner sky model, defined within the first null of the station beam pattern at the lowest observing frequency, has been constructed from a number of components.

The first of these is the composite GLEAM and LoBES catalogue mentioned previously. All sources with a 150 MHz flux density greater than 100 mJy (the nominal completeness limit of that survey and some 1900 in number) were included with a Gaussian representation as noted previously. The extragalactic source population at flux densities less than 100 mJy and down to 1 μJy (at 150 MHz) over a spatial extent of 8×8 degrees was modelled with the T-RECS