SKA Data Challenge 3a (CD/EoR)

Samir Choudhuri and Arnab Chakraborty

Suman Chatterjee, Srijita Pal, Santanu Das, Samit Pal, Narendra Nath Patra, Madhurima Choudhury, Anshuman Tripathy, Chandrashekhar Murmu, Rajesh Mondal, Abinash Kumar Shaw, Rahul Shah, Gurmeet Singh, Soumadeep Saha, Utpal Garain

SDC3 Foregrounds
Foreground Subtraction + 21cm Power Spectrum Extraction

Input Data: Calibrated Visibilities and High Fidelity Image

Challenge will be based on:

a) ability to remove the point source + diffuse foregrounds from the data-set

b) ability to extract the cylindrical power spectrum

Verification of the results from participants

c) Comparison with the original input signal PS

Credit: A. Bolandi

website

sdc3.skao.int

SKA SDC3

Overview Challenges Computational resources Challenge registration Discussion forum FAQs

Science Data Challenge 3

Overview

Purpose

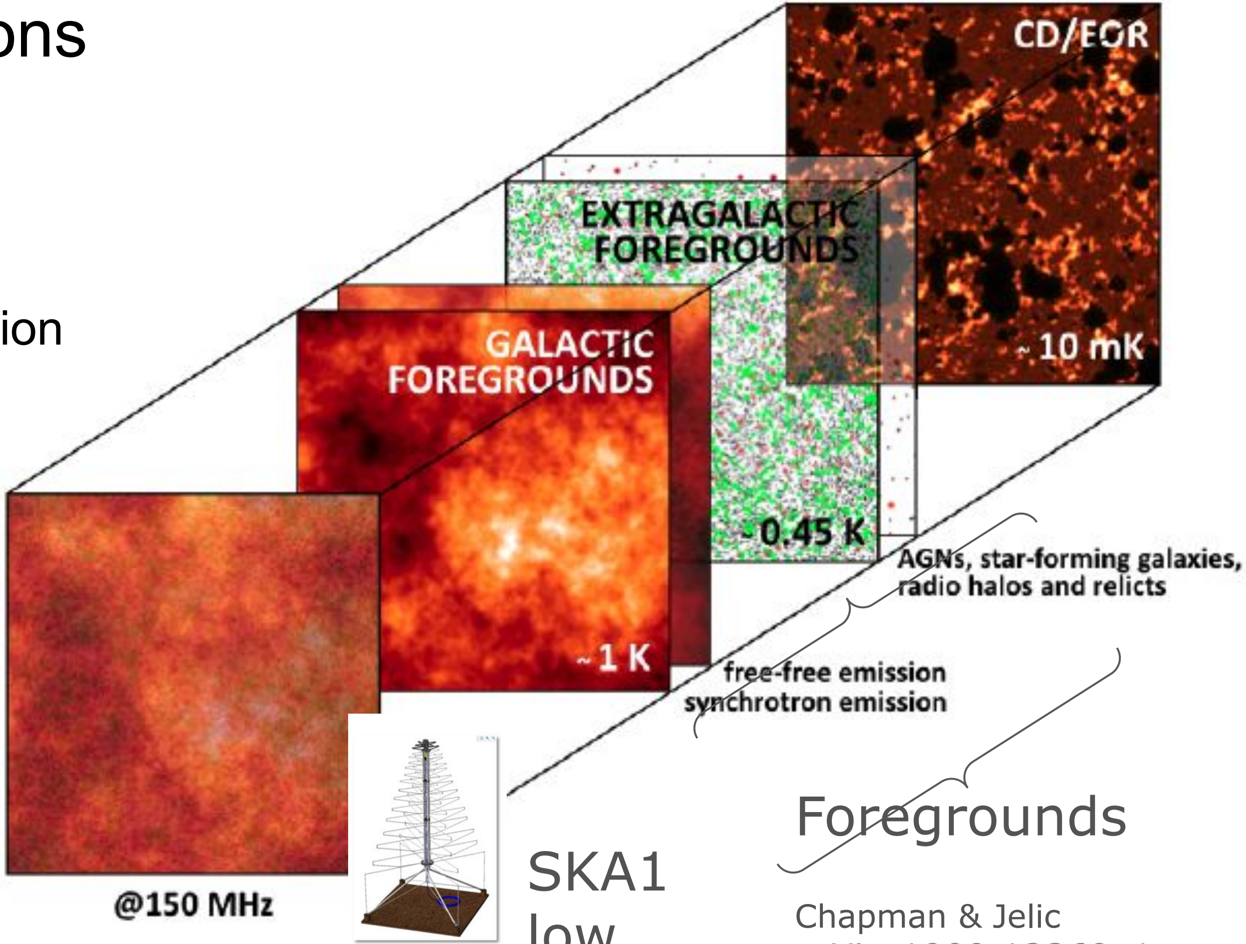
As with our previous two data challenges (SDC1 and SDC2), our goal is to prepare the radio-astronomical community for the novel nature of the data expected from the Square Kilometre Array. Given the order-of-magnitude improvement in sensitivity, new analysis methods are required for both the challenging nature of resulting data, but also for the previously untouched science. Thus, realistic, synthetic datasets emulating the telescope's capabilities will be disseminated to the community to test the suitability of existing methods and foster the development of new ones on these next-generation, scientific datasets. Ultimately, results of each of the competing teams' approaches will be compared via a standard figure-of-merit, instigating a competitive nature to our challenges.

Credit: A. Bolandi

Synthetic Observations

- Signal Cube
- Foreground model
- Telescope Configuration
- Gain Errors (DI +DD)

Synthetic
Data Cubes



SKA1
low

Foregrounds

Chapman & Jelic
arXiv:1909.12369v1

Credit: A. Bolandi

Synthetic Observations

Signal Cube

21cmFAST simulation (corresponding to a specific ionisation history)

512x512 pixel

grid covering 8x8 degrees.

Synthetic Observations

Foreground model

“outer” component is defined over the full 2π steradians above the horizon

—-> A-Team sources that are brighter than a few 100 Jy at 200 MHz as well as the GLEAM catalogue. All sources brighter than 5 Jy at 150 MHz (about 1200 in number) were included in the “outer” Sky Model.

Synthetic Observations

Foreground model

“inner” sky model, defined within the first null of the station beam pattern at the lowest observing frequency

GLEAM and LoBES catalogue with a 150 MHz flux density greater than 100 mJy (some 1900 in number)

flux densities less than 100 mJy (at 150 MHz) down to 1 microJy was modelled with the T-RECS code.

8x8 degrees and was gridded with a 5x5 arcsec pixel sampling

Synthetic Observations

Foreground model

GSM2016 model is severely limited in its effective angular resolution at the low radio frequencies of relevance (about 1 degree),

Galactic foreground emission is supplemented by including simulated emission at the relevant radio frequencies from an MHD simulation of a small Galactic volume sampled with 512x512 pixels.

Synthetic Observations

Error model

partially successful modelling and subtraction of the bright all-sky source population,

an artificial attenuation in the “outer” sky model beyond the central 8x8 degrees. The magnitude of that attenuation is a factor of 1e-3

Synthetic Observations

DD Error model

ARatmospy code - Several ionospheric layers were simulated

The code is used to construct a time evolving phase screen above the telescope site that introduces Direction Dependent (DD) calibration errors into the visibilities via OSKAR.

Synthetic Observations

DI error model

Random values from a Gaussian distribution with a specified standard deviation 0.02 degrees in phase and 0.02% in amplitude for each of the time and the frequency domains.

Telescope: SKA1 - Low

General

- Observation track length HA = -2 to +2 hours
- Thermal noise equivalent 1000 [h]
- Field of View: one SKA1-Low pointing at RA, Dec = 0h, -30deg

Measurement sets

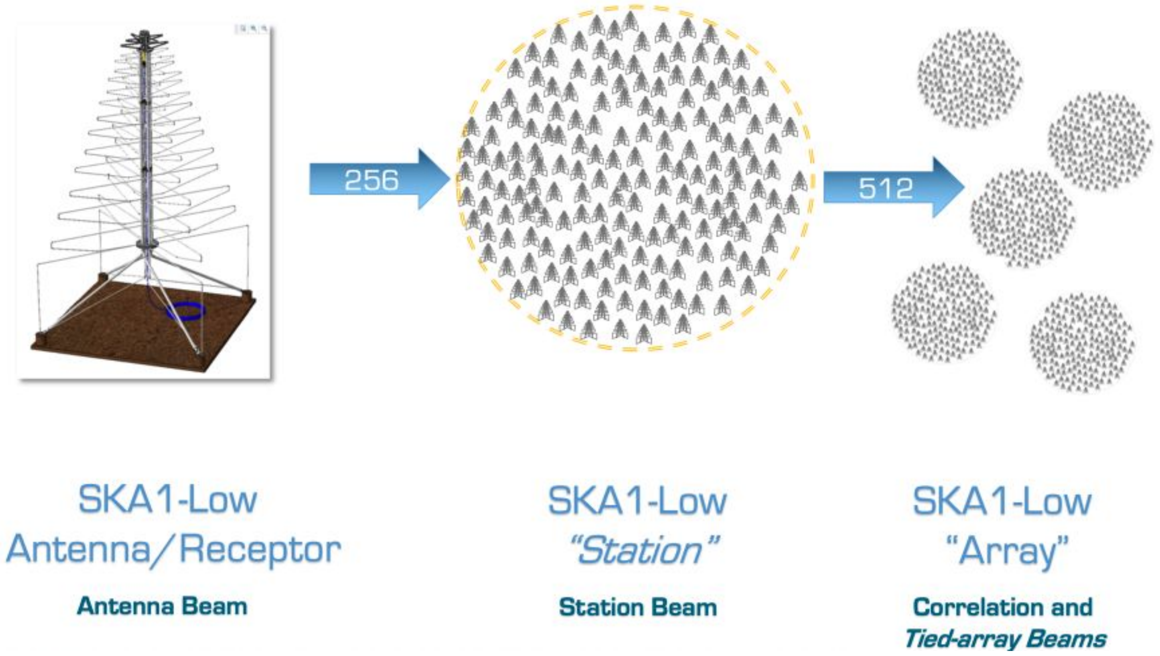
- Integration time 10 [s]
- Channel width 100 [kHz]
- Frequency coverage 106 - 196 [MHz]

Image cube

- Weighting: Natural
- Pixel size [arcsec]: 16x16 arcsec
- Number of pixels in RA/Dec 2048x2048

Ancillary data

- Synthesised beam and primary beam at each frequency



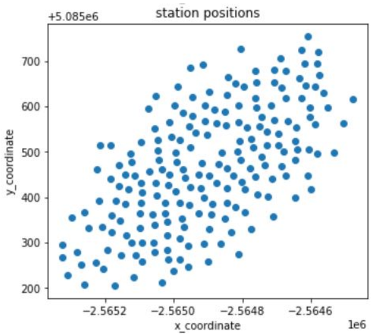
Credit: A. Bolandi

SKA data specification:

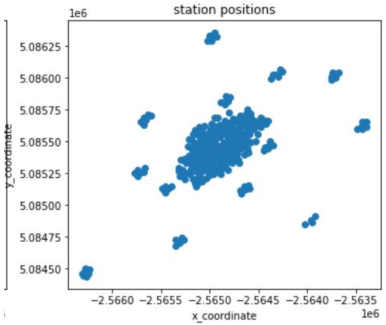
- **Test data:**
- 150 uvfits test files was given
- This data was given to check that the estimator we are using is giving correct power-spectrum.

- **Main data**:
- 900 uvfits files
- 7.5TB
- Station beam image file(for 900 channels)
- Field of View: 5°×5°in sky at RA, Dec = 0h, -30deg

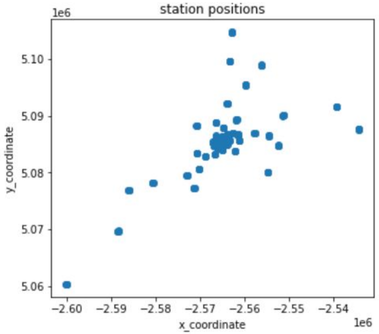
Antenna Layout:



200 stations



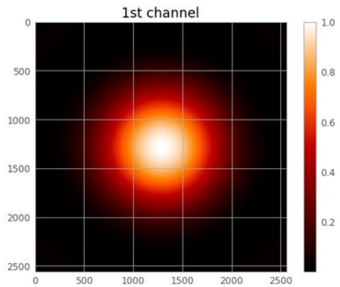
300 stations

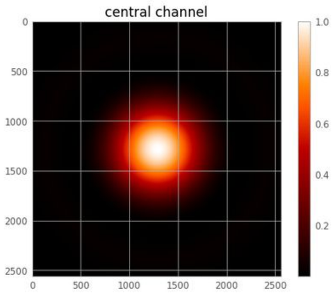


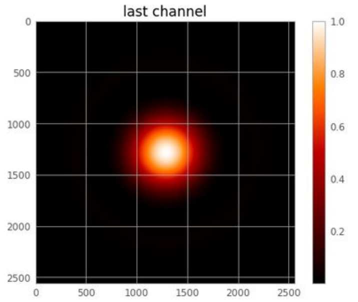
512 stations

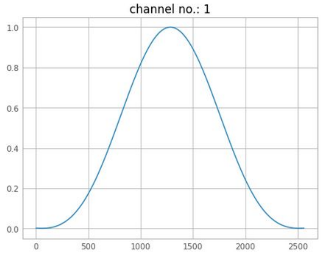
Credit: Santanu Das

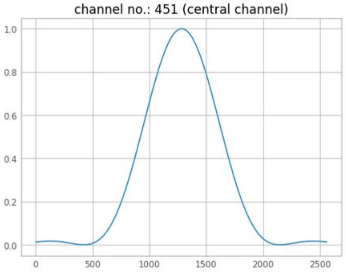
Some Plots of Station Beam pattern 5°×5° (from the Station Beam fits file)

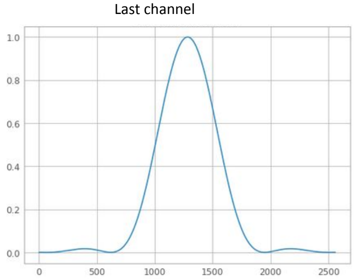












Credit: Santanu Das

Result Submission

Power spectrum 6 frequency bins, 9 bins in k // and 9 bins in k⊥, a total of 12 files (6 containing values, and 6 their corresponding errors)

Score computation

$$Prob(P'_j) = 1/[\sqrt{2\pi}\,\Delta P_j]\ exp[-\,(P'_j - P_j)^2/2\Delta P_j^2].$$

$$Prob(P') = \prod_j Prob(P'_j).$$

Results

FG filter and Power spectrum estimation : — divided into **Even** and **Odd** times-tamps, to avoid noise bias

1. **Make a high resolution map from the vis data** →
2. **Subtract the bright point source model from the data** →
3. **Make a dirty map from the residual vis data with baselines <2.5kl** →
4. **We generate two image-cubes for even and odd time samples - p0 and p1** ↓
5. **Apply Gaussian Process Regression (GPR) to each image cubes to remove residual FG.** →
6. Fourier transform along freq axis :

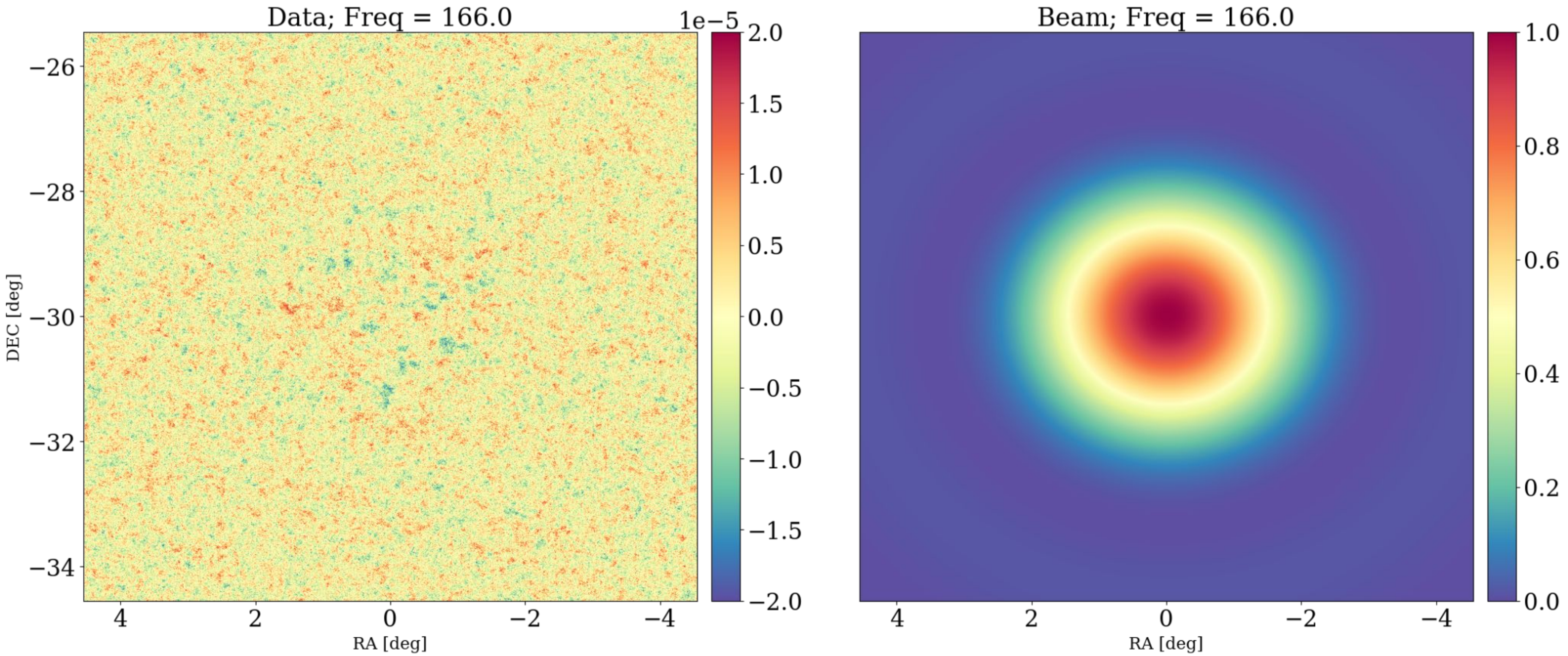
   (RA,DEC,nu) -> (RA,DEC,tau) →
7. Spatial 2D Fourier transform along RA and DEC

   (RA,DEC,tau) -> (kx,ky,tau) ↓
8. Correlate Even cube (p0) with Odd cube (p1) :

   3-dimensional power spectrum →
9. Cylindrically averaged the modes to get **2-dimensional power spectrum** →
10. **Correct for signal loss bias due to FG filtering, by applying a transfer function**

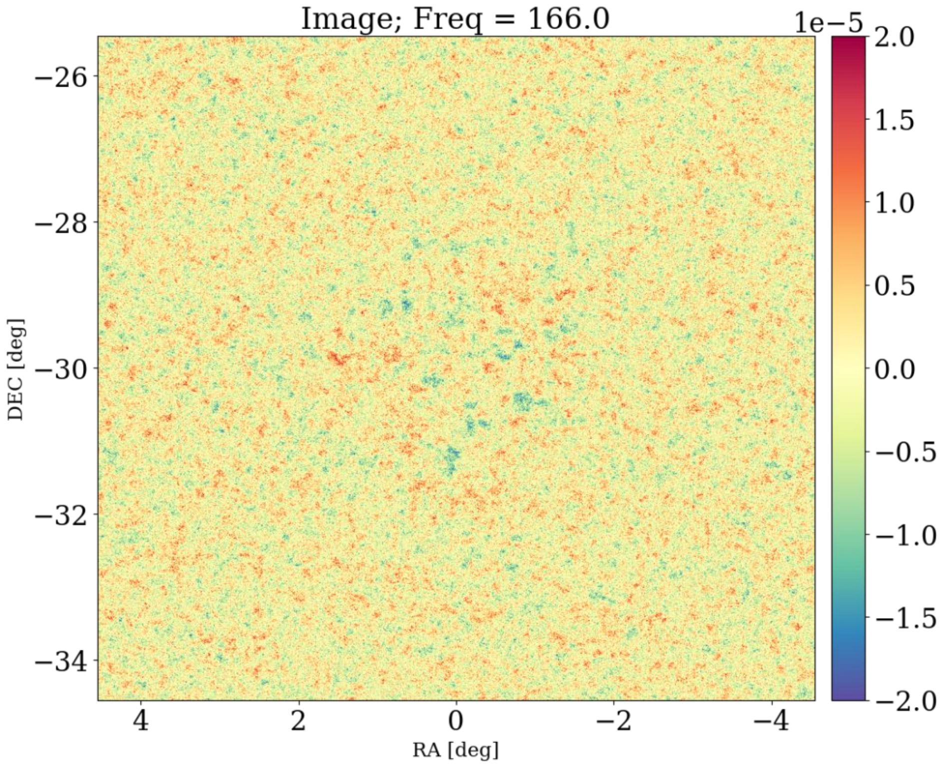
Pipeline development : **Arnab Chakraborty** + **Suman Chatterjee**

Check the Power Spectrum estimation with the test data: HI + Noise



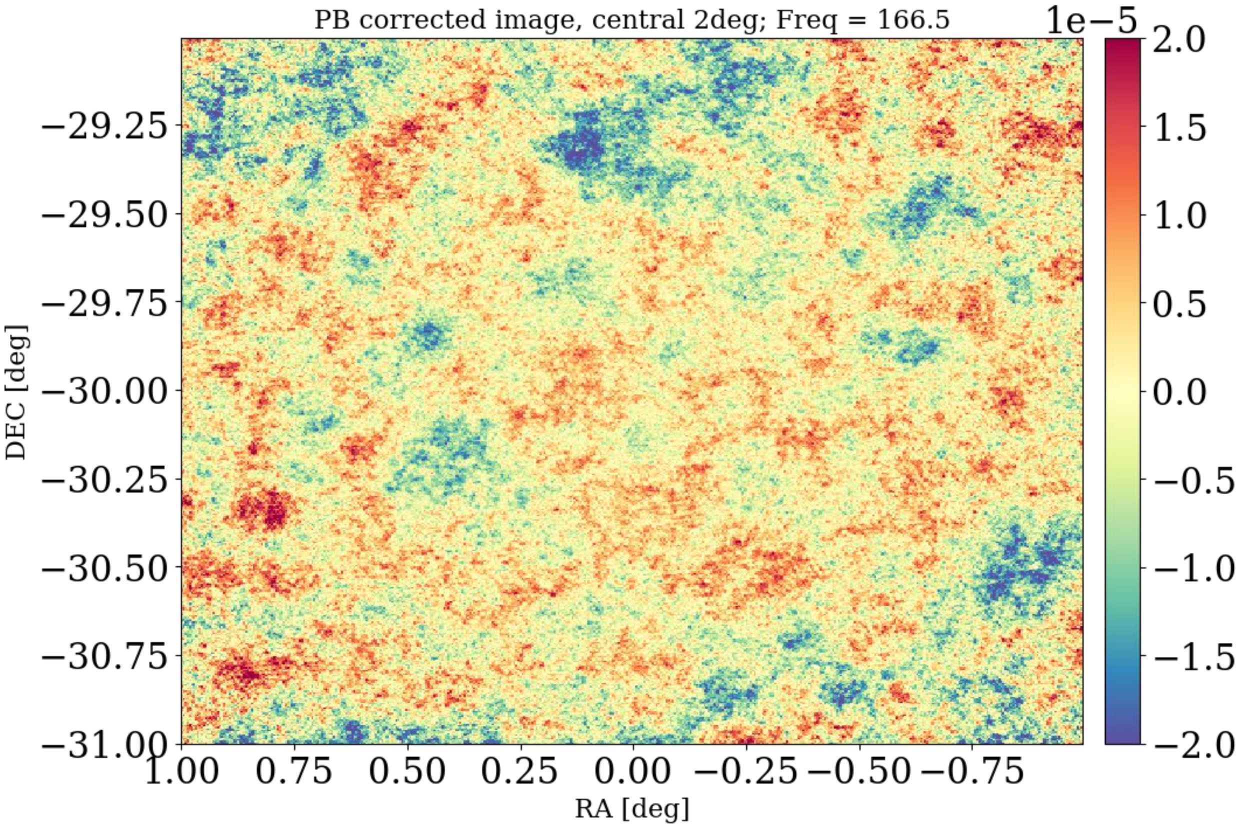
Check the Power Spectrum estimation with the test data: HI + Noise

Correct for Primary beam response

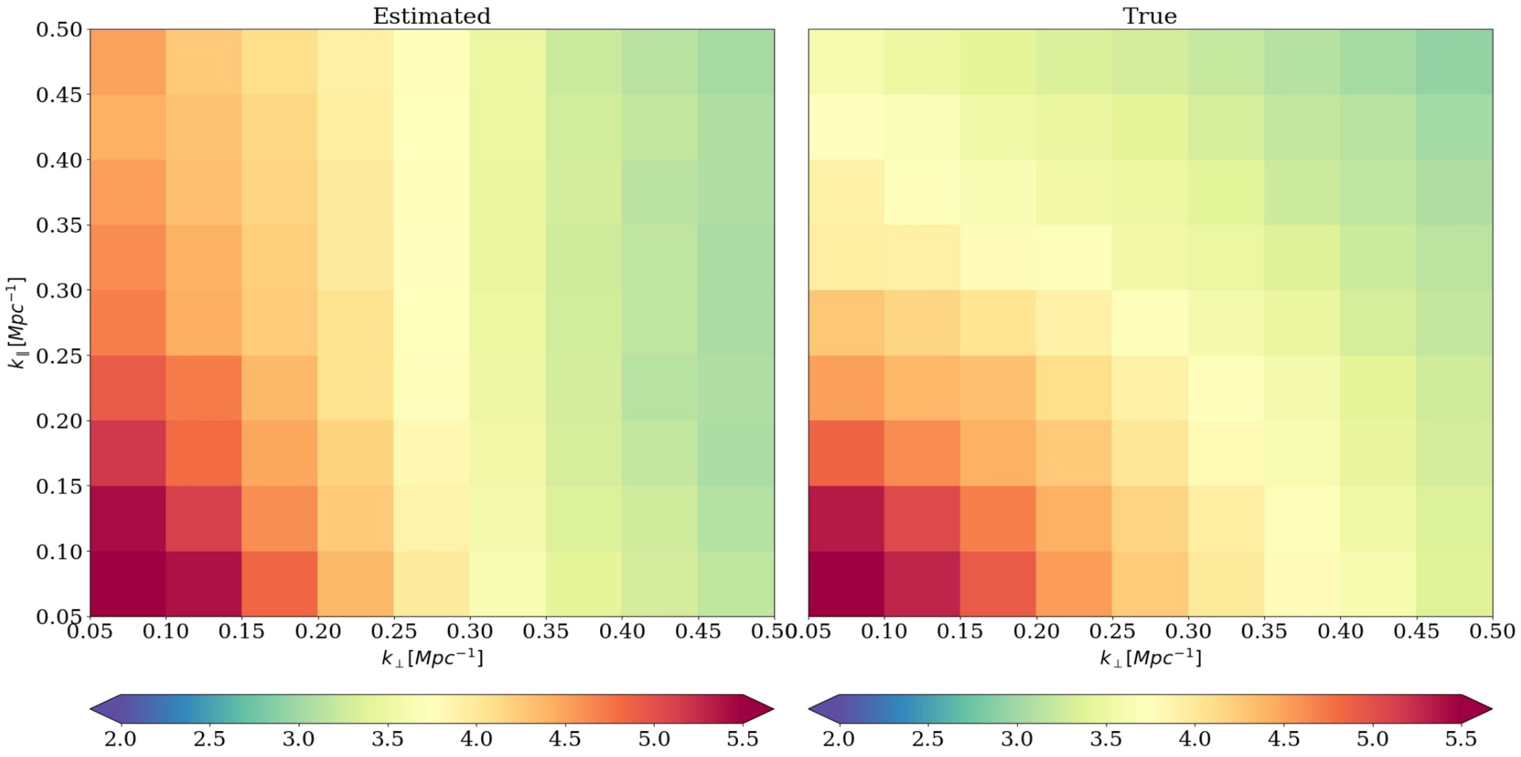


Check the Power Spectrum estimation with the test data: HI + Noise

Take the central 2deg x 2deg for which we have to estimate the power spectrum

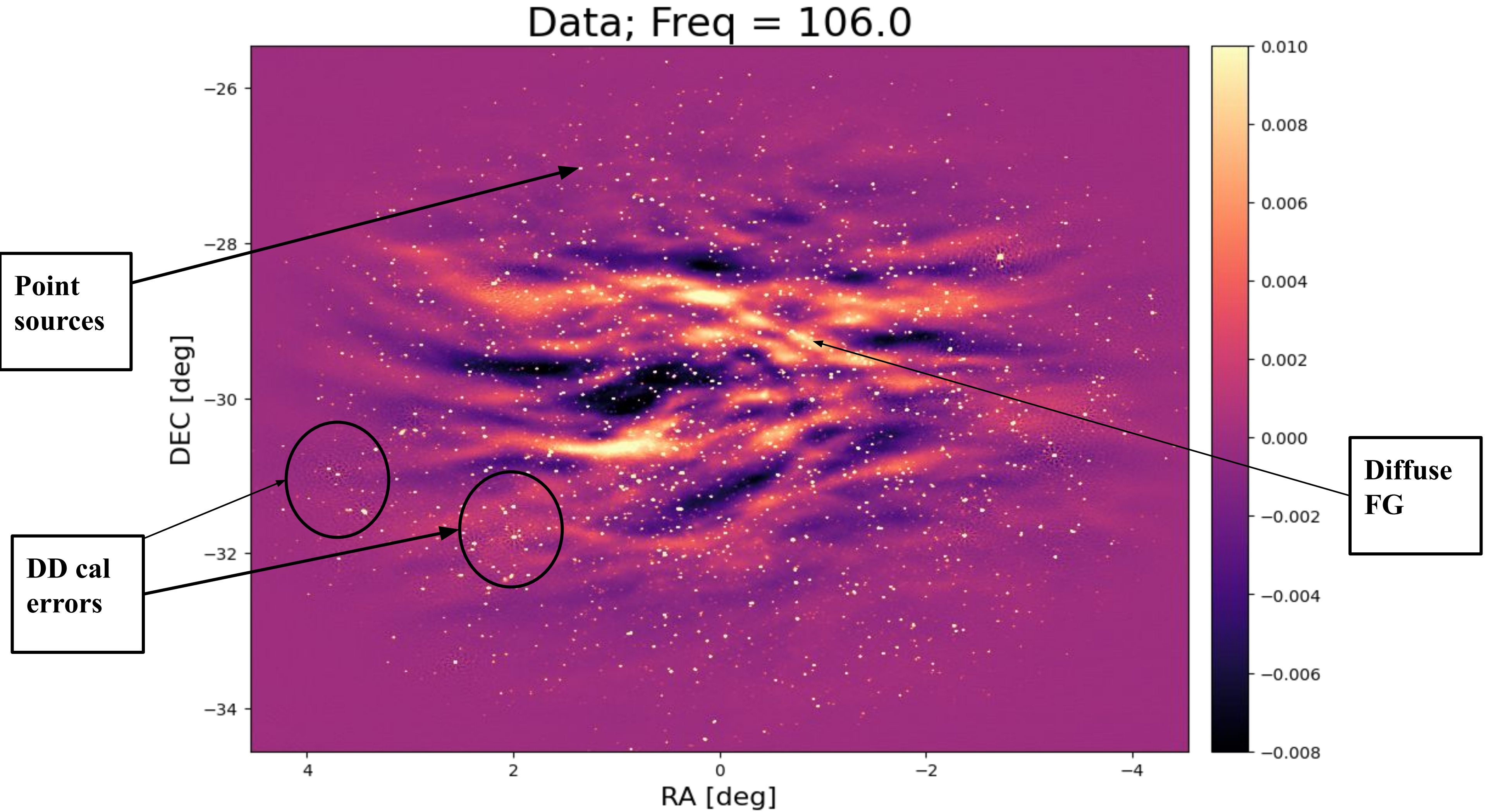


Check the Power Spectrum estimation with the test data: HI + Noise



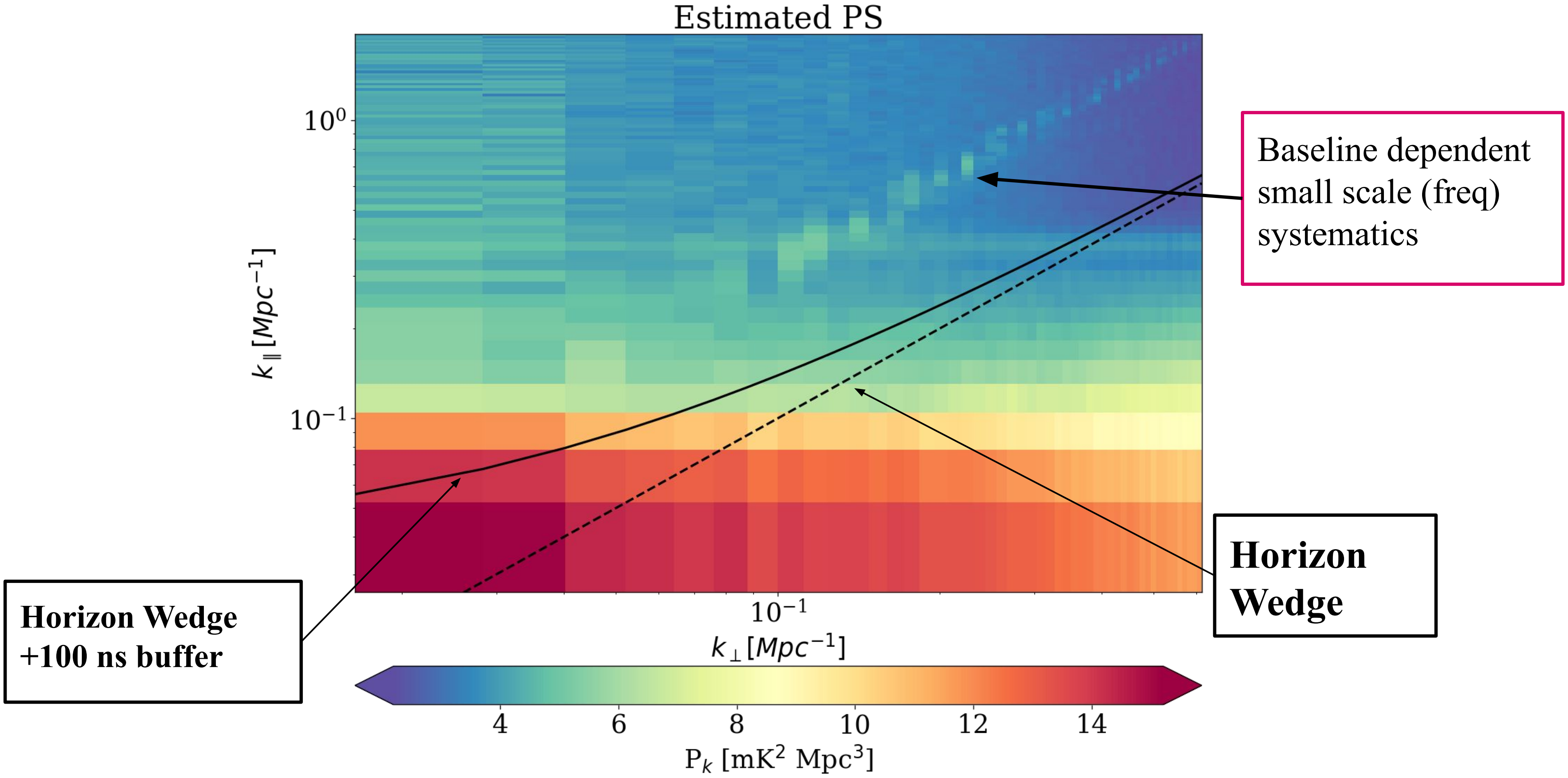
Left : Estimate PS from the HI+Noise data. Right : The true underlying HI PS provided by SKAO.

Overall normalization and PS estimation is working. However, note in the left there is little noise bias, which we have not corrected for here.



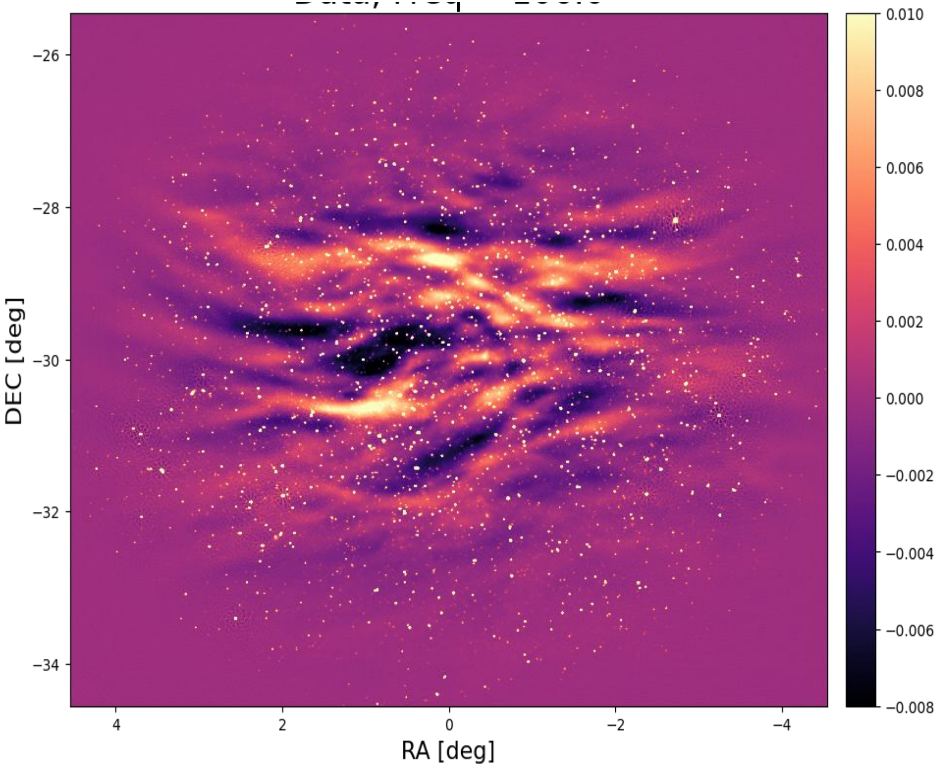
Data; Freq = 106.0
DEC [deg]
RA [deg]
−26
−28
−30
−32
−34
4
2
0
−2
−4
0.010
0.008
0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
−0.008
Point sources
DD cal errors
Diffuse FG

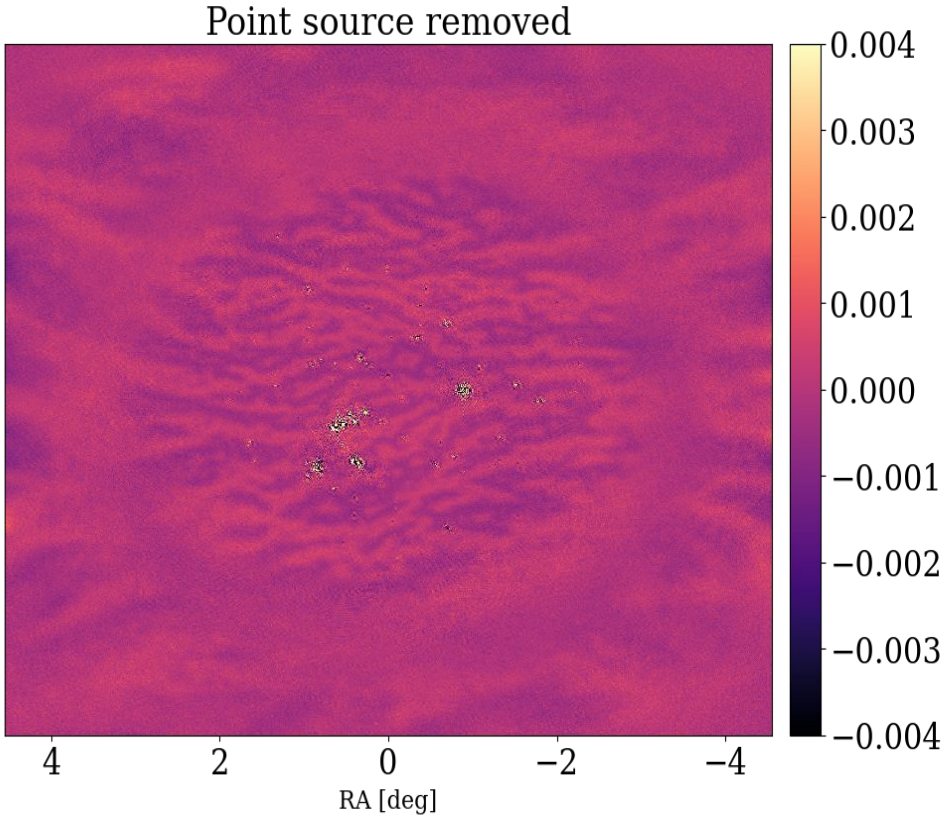
2D ps of the data (FG+Noise+HI) – No FG filtering and No noise bias subtraction



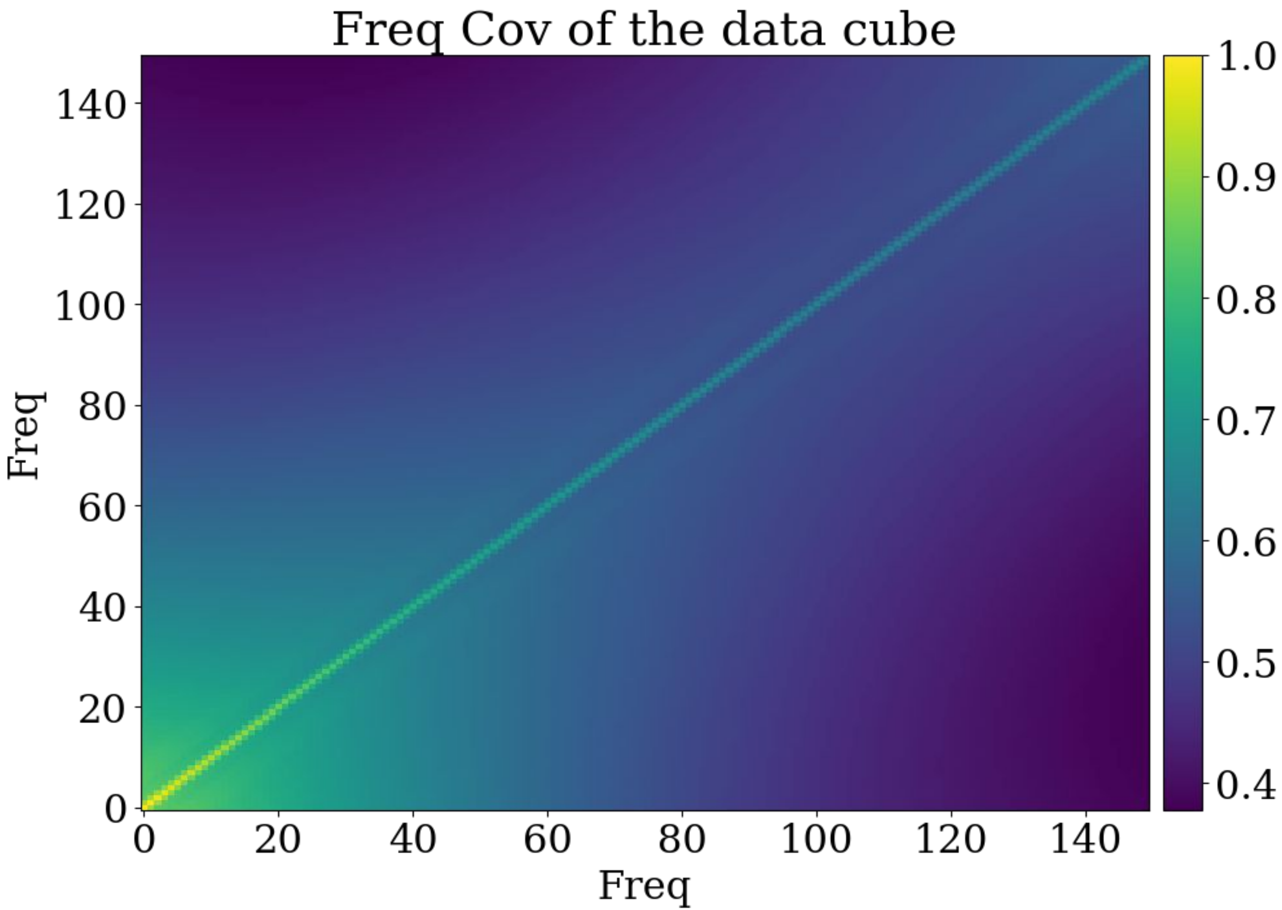
Point source removed map : Subtract a high resolution model from the data and make a naturally weighted dirty image cube with the residual

Before Subtraction

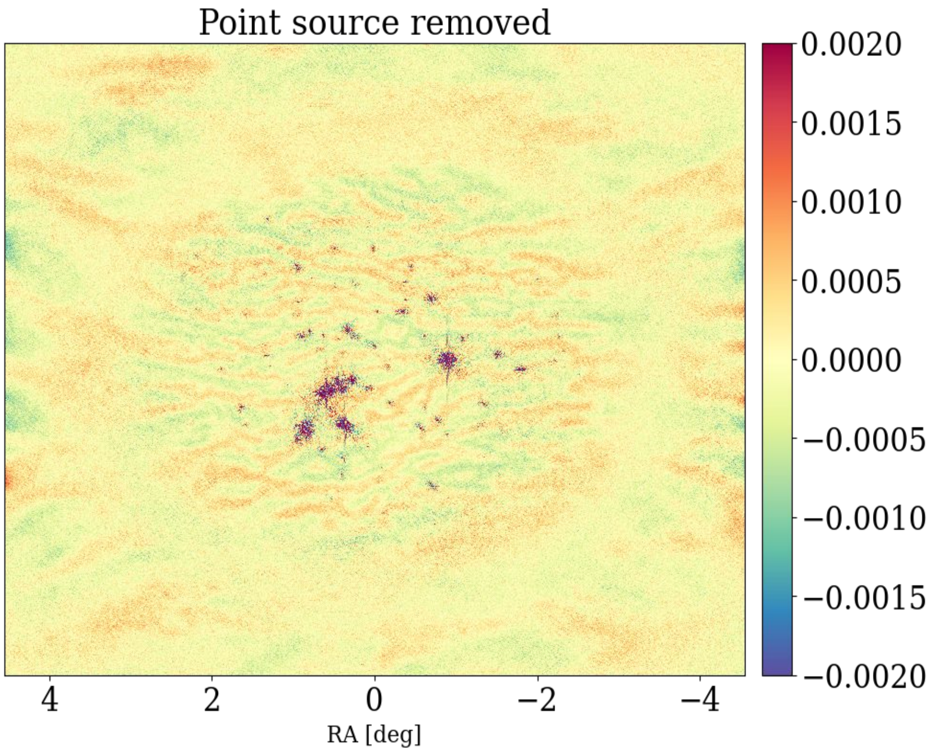


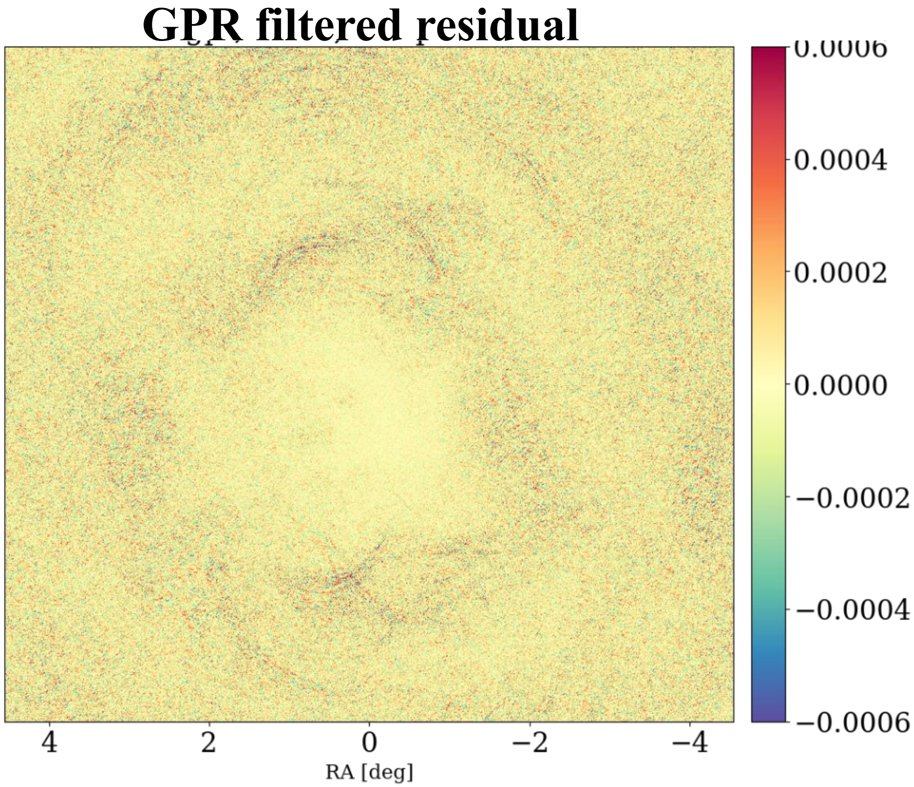


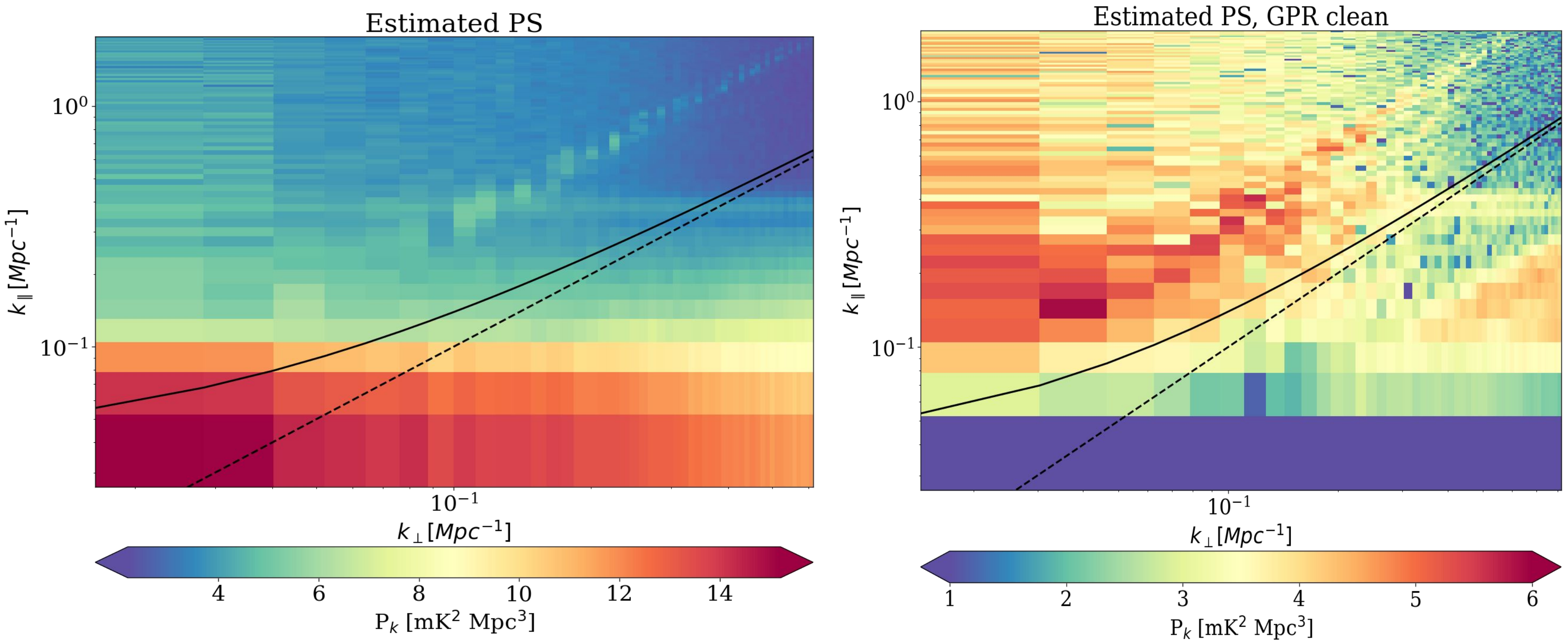
Freq-Freq Covariance of the residual data (after pts removal)



Foreground removal with GPR

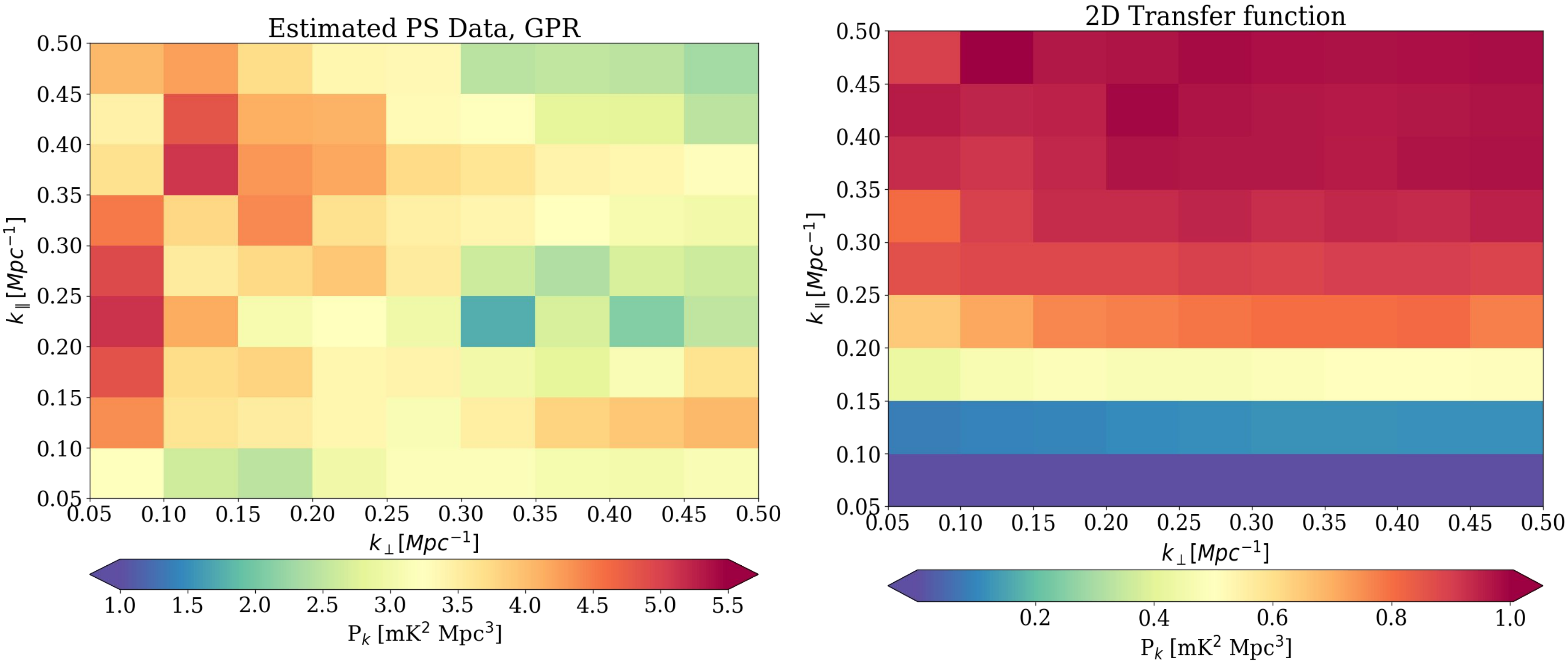




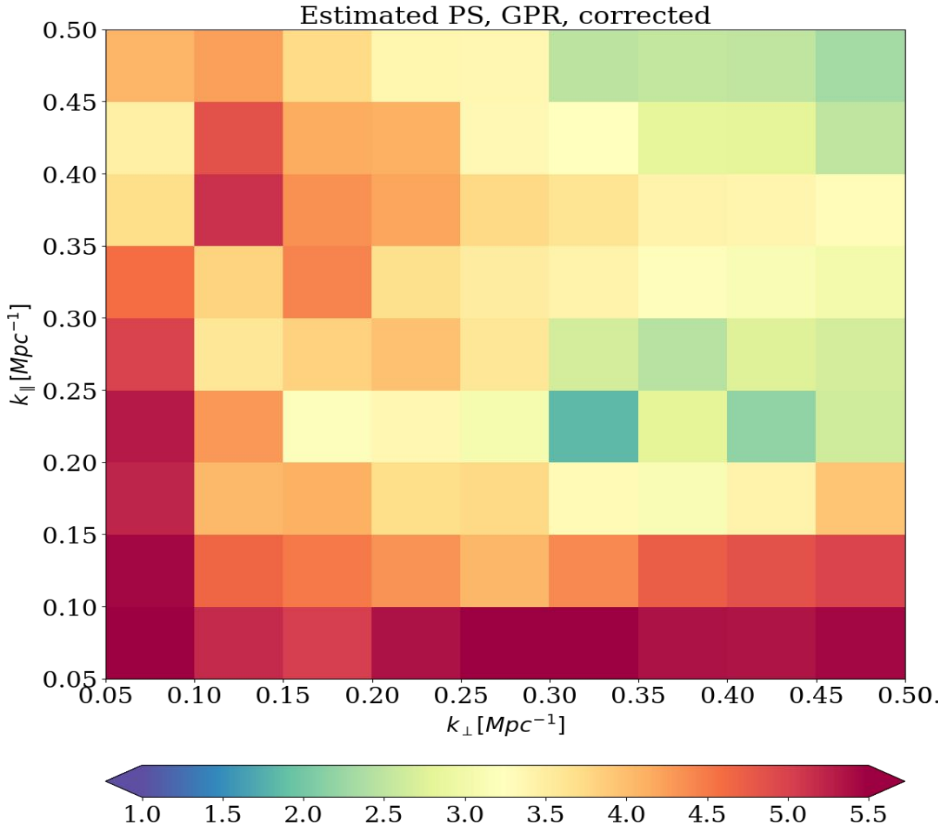


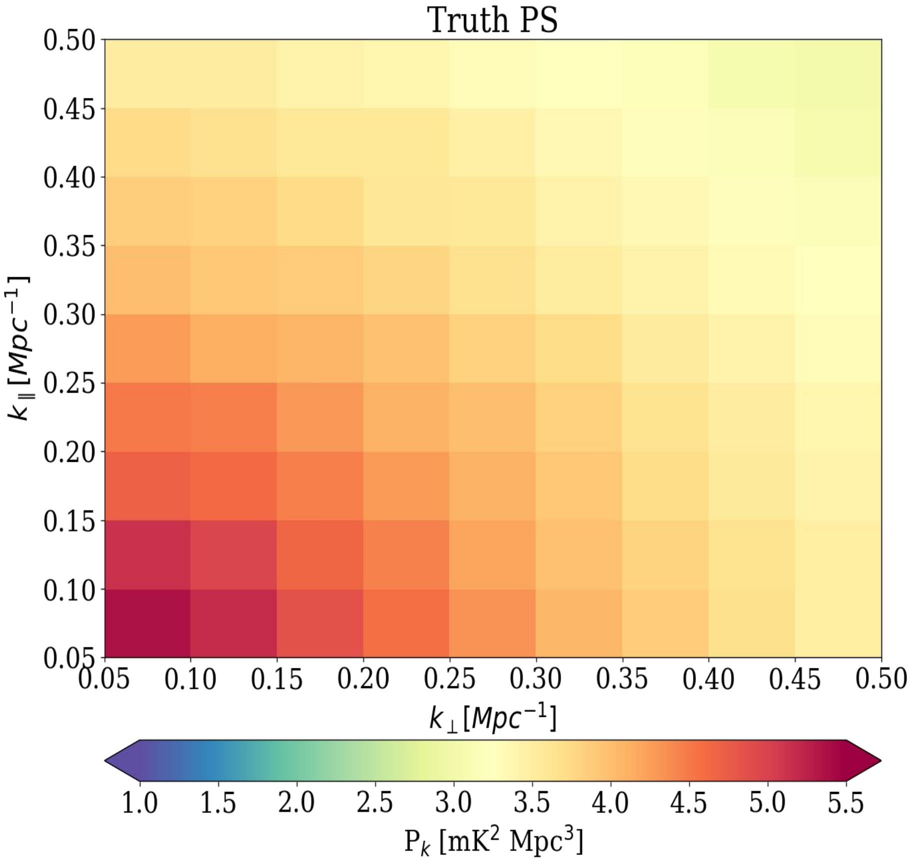
Although bulk of FG is removed, but there is a baseline dependent small scale systematics, which is diagonal and GPR is unable to capture that.

Signal loss correction through a transfer function



Signal loss corrected and compared with true HI

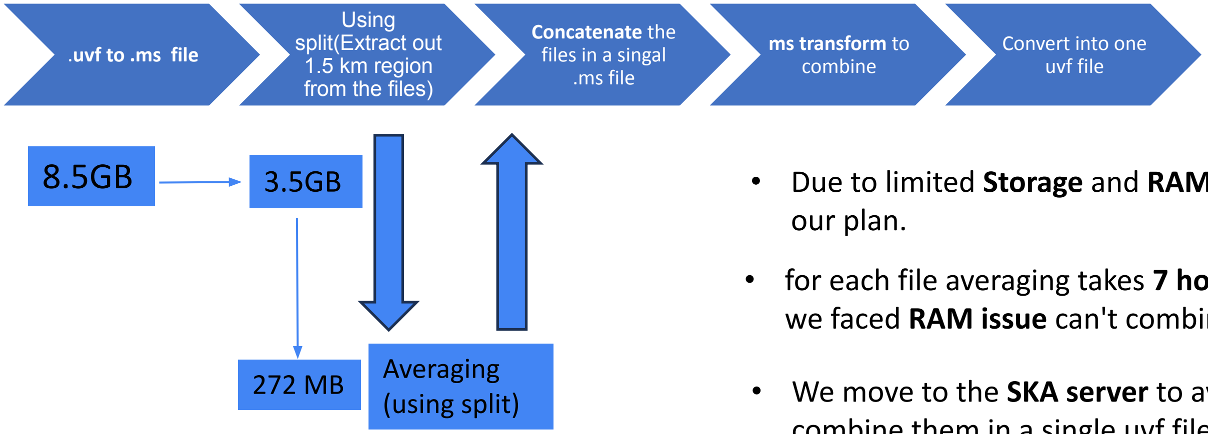




SKAO provided this recently

Visibility based TGE

The issues we are faced so far:

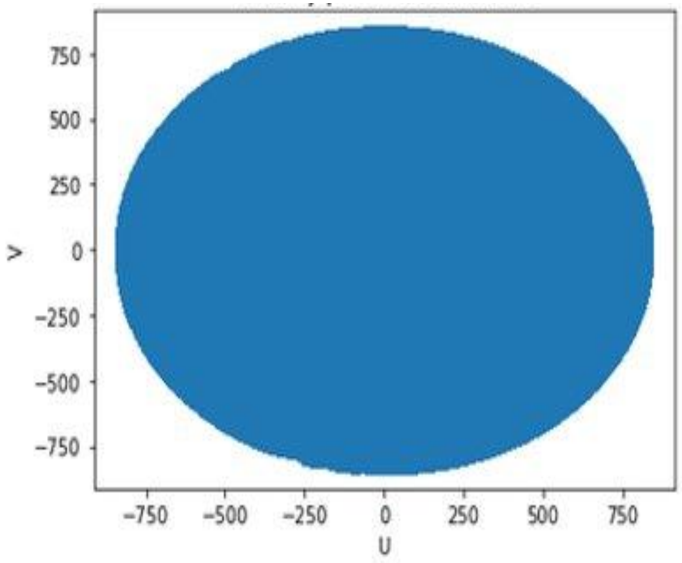


- Due to limited **Storage** and **RAM** in local server we had to change our plan.
- for each file averaging takes **7 hours 41 mins** in our local server but then we faced **RAM issue** can't combine the files**.**
- We move to the **SKA server** to average out 150 files and combine them in a single uvf file.
- But when we try to work main data the averaging of each file takes **35 hours 55 mins.**

- **Reduce down** the averaging time or use the **time very efficiently**(by running multiple script files) for averaging
- Process out the data of **900** uvfits files and combine them to make **6 bands** of frequencies.
- Apply the **TGE(Tapered Gridded Estimator)** to compute the power spectrum for each band.
- Analyse the output results.

Credit: Santanu Das

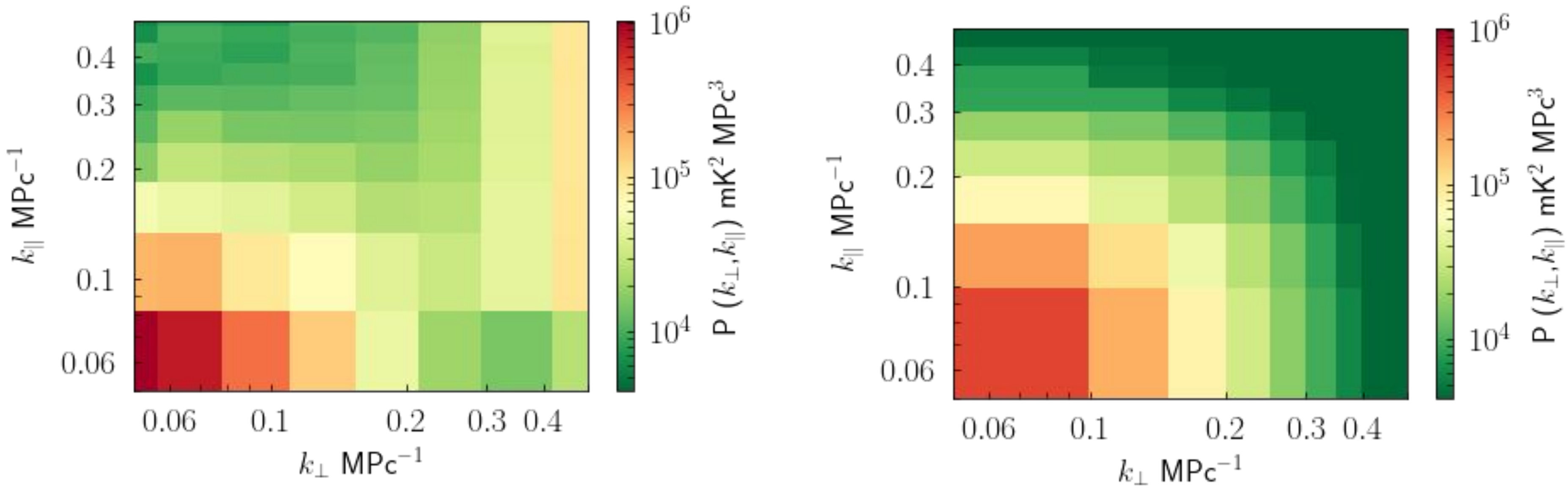
Visibility plane of the test data (1st uvf file)



- Here we extract out 1.5 km region (or 830 in wavelength unit) from the original visibility plane

Credit: Santanu Das

TGE on test data : HI + Noise



Estimated

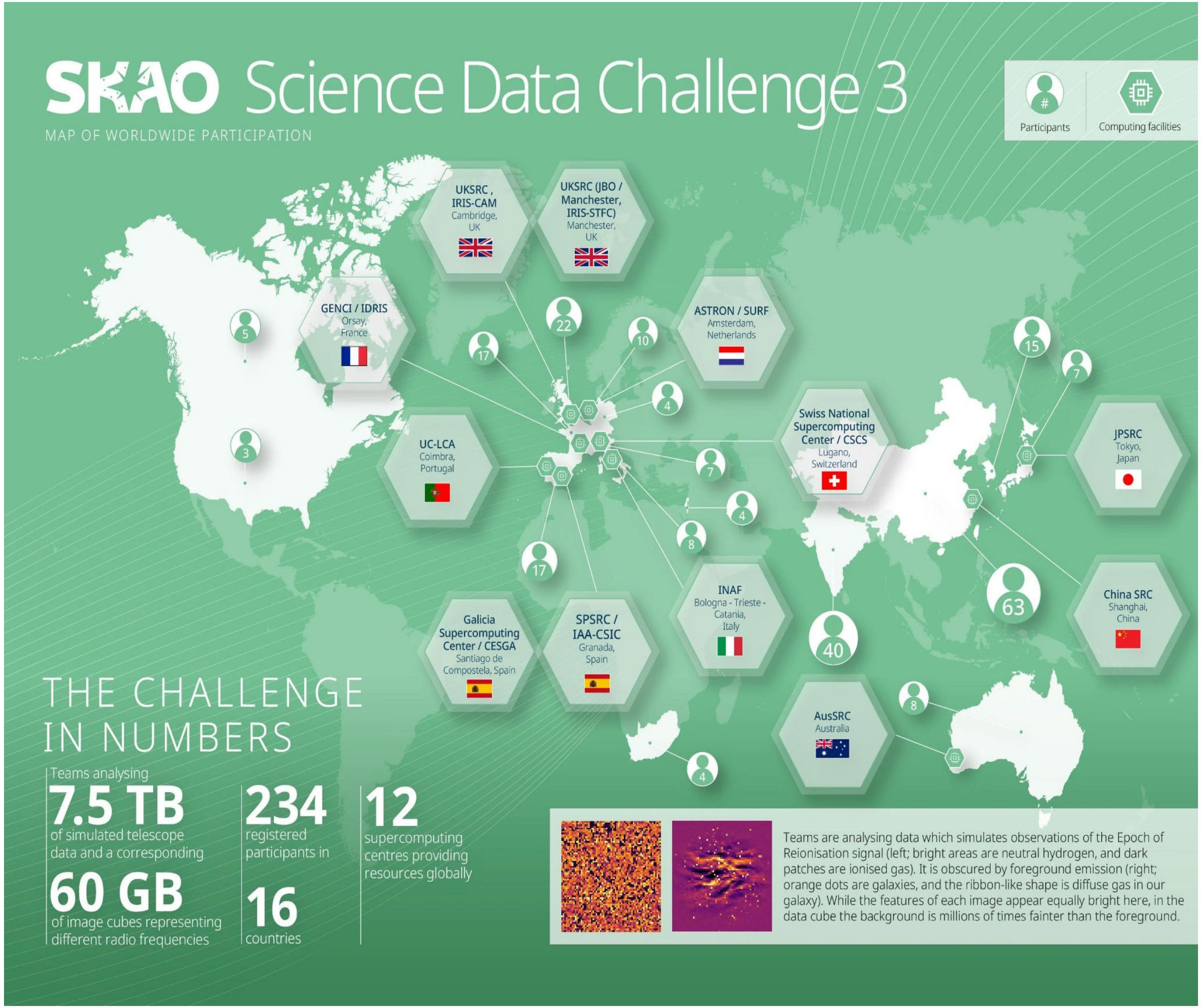
f=0.8

True

Overall normalization and PS estimation is working

Total run time (1 data+10 Mg) 35 hrs

Credit: Srijita Pal



Rank	Team	Score
1	HIMALAYA	74758
2	DOTSS-21cm_ML-GPR	71573
3	DOTSS-21cm_Advanced_ML-GPR	71135
4	ERWA	63670
5	DOTSS-21cm_Avoidance	51889
6	Shuimu-Tianlai	43422
7	Wizards_of_Oz_3D	33295
8	Akashganga	31864
9	REACTOR	21888
10	SKACH	12103
11	KUSANAGI	
12	Cantabrigians	
13	Hausos	
14	KUSANAGIb	
15	Nottingham-Imperial	
16	Pisano_Galaxy_Moppers	
17	HAMSTER	
18	Foregrounds-FRIENDS	
19	KORSDC	

Same team

SDC3 Inference

Extraction of reionization parameters (SWG contacts: Mesinger & Mellema)

Target Participants: SWGs like CD/EoR.

Input Data: EoR PS + noise and residual foreground contamination

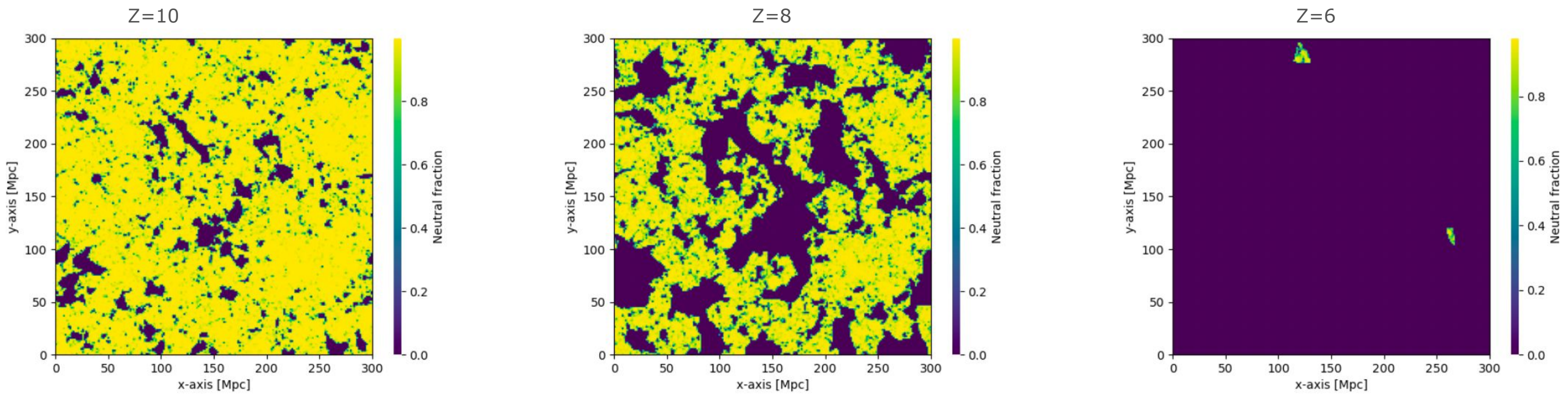
Challenge will be based on:

a) ability to extract the IGM and source properties

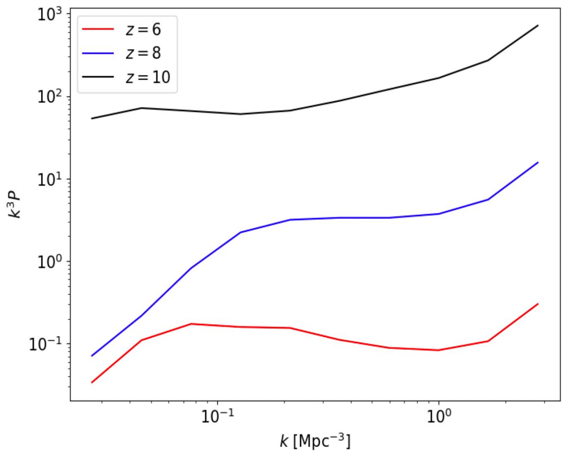
Verification of the results from participants

Comparison with the input EoR history **(ionization fraction)**

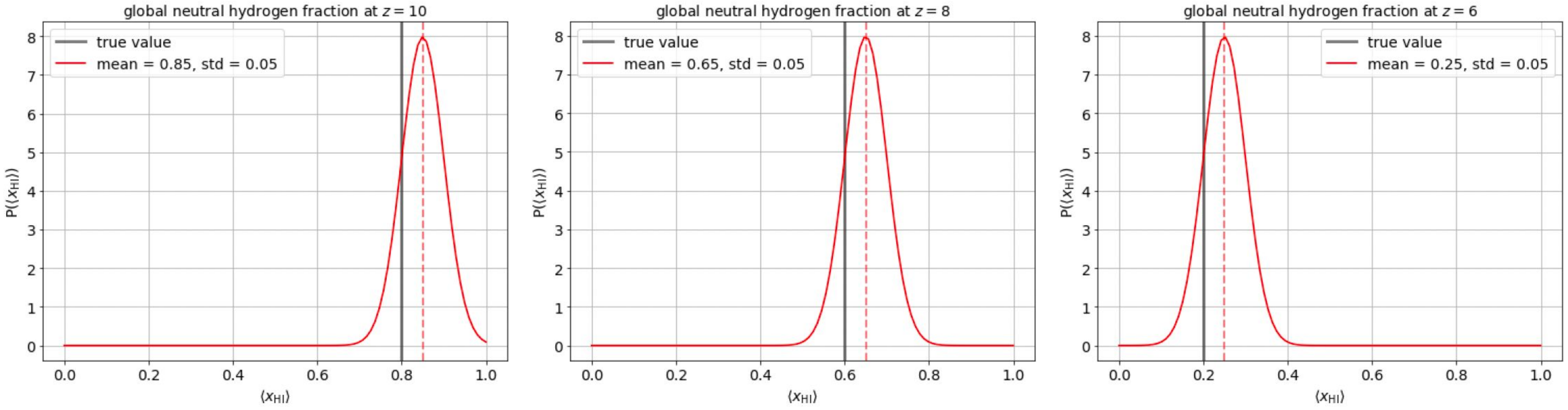
Credit: A. Bolandi



EoR power spectra corresponding to the 3 cubes



Credit: A. Bolandi



Propoposed score: $\prod_{z1}^{z3} P(x_{\mathrm{HI}})$

Figure credit: Eunseong Lee

Resources (preliminary) - inference

Inference - Dataset size minimal; disk space per team 100 GB

- If performing "forward modelling" inference (or emulator + training):
  - Around 256 cores having 2GB (preferably 4) GB RAM each (with some flexibility)
  - Quota few 100K core h per team
- If using analytical models / emulators:
  - 8-32 cores
  - Quota few K cores h

Thank you

Credit: A. Bolandi

Summary

We participated in the SDC3a - foreground removal.

Image based and visibility based power spectrum estimators are applied in the test and actual data.

We are in the 8th position in global scoreboard.

We can plan for SDC3b, coming next year.

Thank you

Around 50% of the stations will be located within a 1 km diameter core, with the remaining stations organised in clusters of 6 stations on three modified spiral arms. The maximum baseline length will be around 70 km.

field of view, ranging from about 40 square degrees at 50 MHz to about 18 square degree at 1.4 GHz.

Gaussian Process Regression (GPR) to model covariance of the each component of the data

Data Cov :- C = C_fg + C_mix + C_21 + C_N

GPR :- $$K_{\text{total}} = K_{\text{fg}} + K_{\text{mix}} + K_{21} + K_{\text{N}}$$

Kernel function/covariance functions are not data covariance function. Data covariance is not known exactly. We wish to find the kernel functions K that best fit the covariance of our data C. For example, if C_fg is the foreground data covariance, then we want to find a kernel function K_fg, that best describes that and the best-fitting hyperparameters.

$$\mathrm{E}[\mathbf{f}_{\text{fg}}] = \mathrm{K}_{\text{fg}}\left[\mathrm{K}_{\text{fg}} + \mathrm{K}_{21} + \mathrm{K}_{\text{n}}\right]^{-1}\mathbf{d},$$
$$\mathrm{cov}[\mathbf{f}_{\text{fg}}] = \mathrm{K}_{\text{fg}} - \mathrm{K}_{\text{fg}}\left[\mathrm{K}_{\text{fg}} + \mathrm{K}_{21} + \mathrm{K}_{\text{n}}\right]^{-1}\mathrm{K}_{\text{fg}}$$

Here we used this kernel for optimal FG filtering

$$k_{\text{total}} = k_{\text{fg,smooth}} + k_{\text{mode-mix}} + k_{\text{ex}} + k_{21cm} + k_{N,gauss}$$

Residual :- $$r = d - E[f_{\text{fg}}]$$