# Machine Learning and Experimental Design for Hydrogen Cosmology

David Rapetti

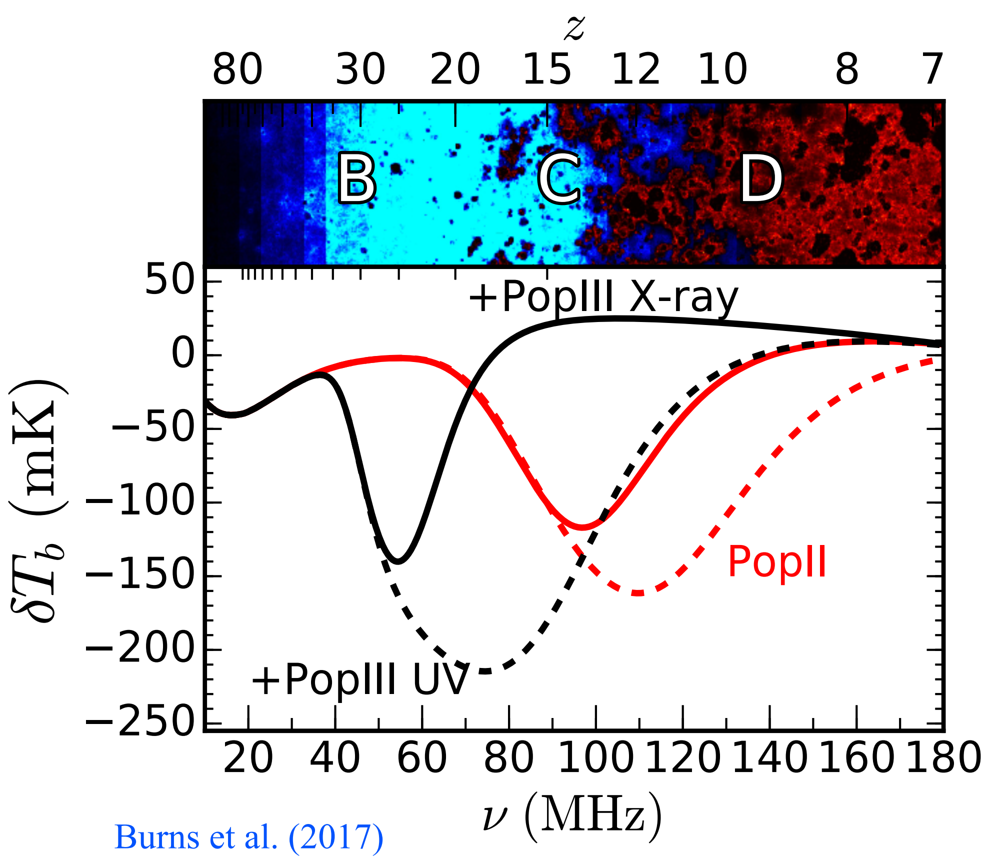
University of Colorado Boulder / NASA Ames Research Center

The work presented here is in collaboration with:

Keith Tauscher (CU Boulder), Jack O. Burns (CU Boulder), Jordan Mirocha (UCLA), Eric Switzer (NASA Goddard), Raul Monsalve (CU Boulder), Steven Furlanetto (UCLA), Judd Bowman (Arizona State)

NESS
Network for Exploration and Space Science

# HYDROGEN COSMOLOGY

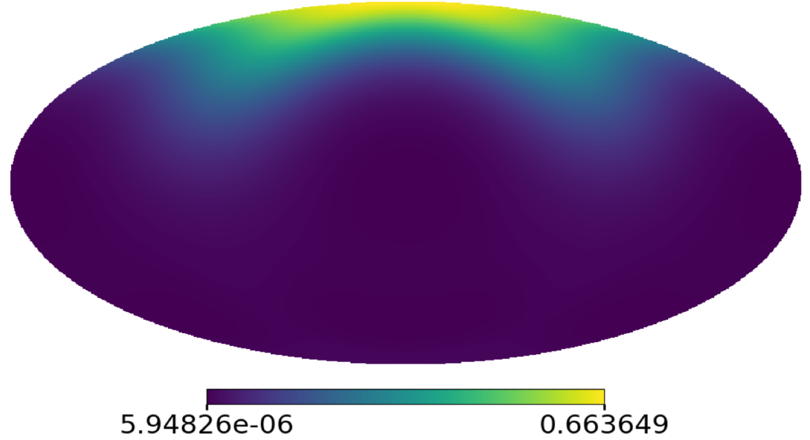



Burns et al. (2017)

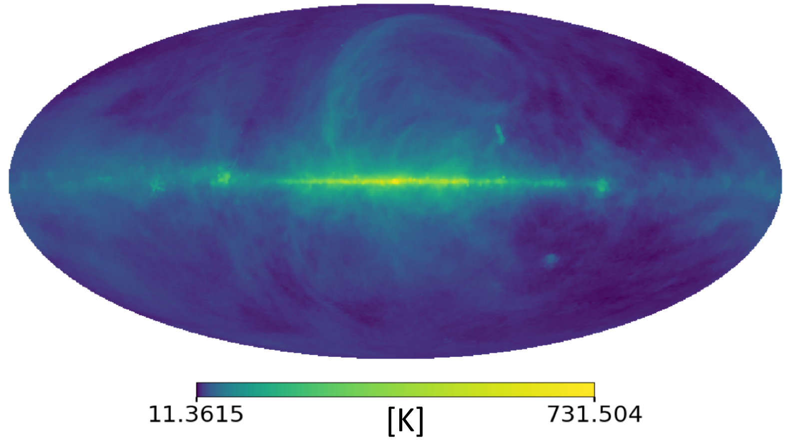
- Upper panel: Evolution of a Universe's slice from early (left) to late times (right).
- Lower panel: Standard models of the global 21-cm spectrum relative to the CMB temperature; red models with metal-rich stars (Pop II), black curves assume that metal-free stars (Pop III) also occur, but only in low-mass galaxies where atomic cooling is inefficient. The dashed and solid curves differ in specific emission and stellar properties (see Burns et al. 2017 for details).
- The epochs B, C and D correspond to the ignition of the first stars, the initial accretion of black holes, and the onset of reionization, respectively.
- Figure from adapted in turn from Pritchard & Loeb (2010) using newer models from Mirocha et al. (2017).

# FOREGROUND TRAINING SET

DARE beam at 80 MHz




All-sky 408 MHz map from Haslam et al. (1982)



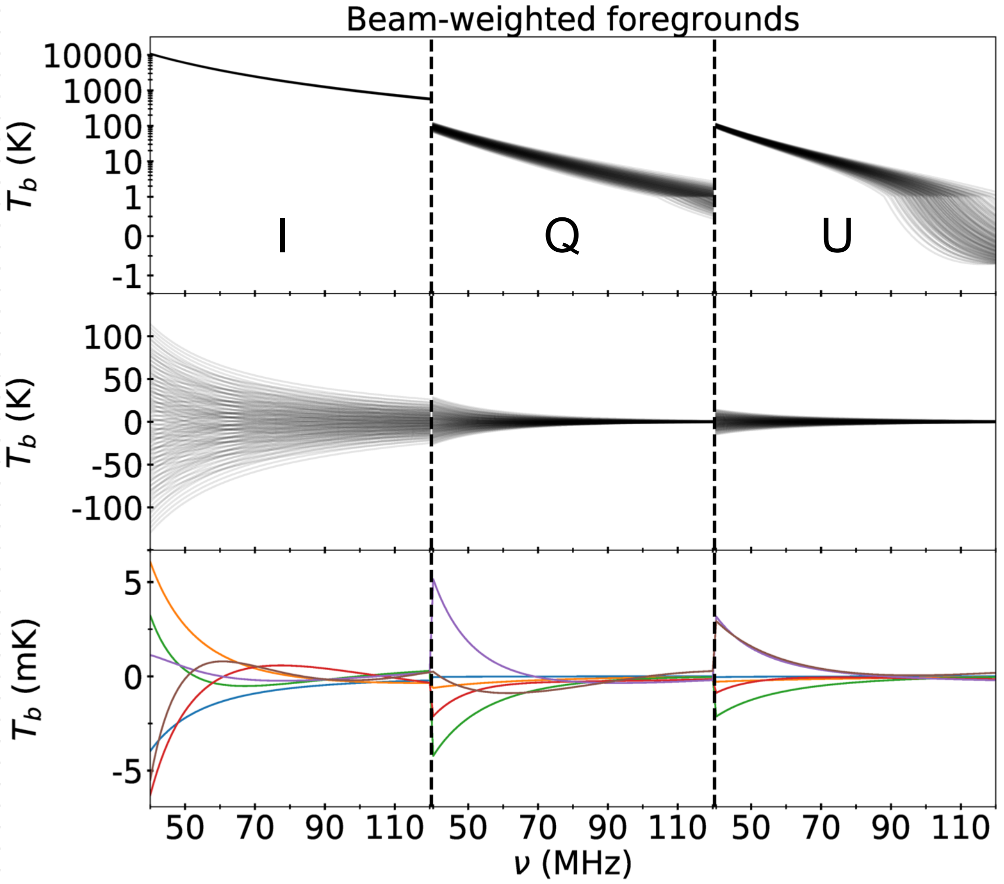

- Antenna temperature simulated convolving beam, $B(\nu,\Omega)$, and sky, $T_{sky}(\nu,\Omega)$, through

$$T_A(\nu) = \frac{\int B(\nu,\Omega)\, T_{sky}(\nu,\Omega)\, d\Omega}{\int B(\nu,\Omega)\, d\Omega}$$

- CST code used to model beam
- Sky maps from Guzmán et al. (2010) and Haslam et al. (1982)

# EXPERIMENTAL DESIGN: INCLUDING STOKES PARAMETERS INTO THE LIKELIHOOD FUNCTION

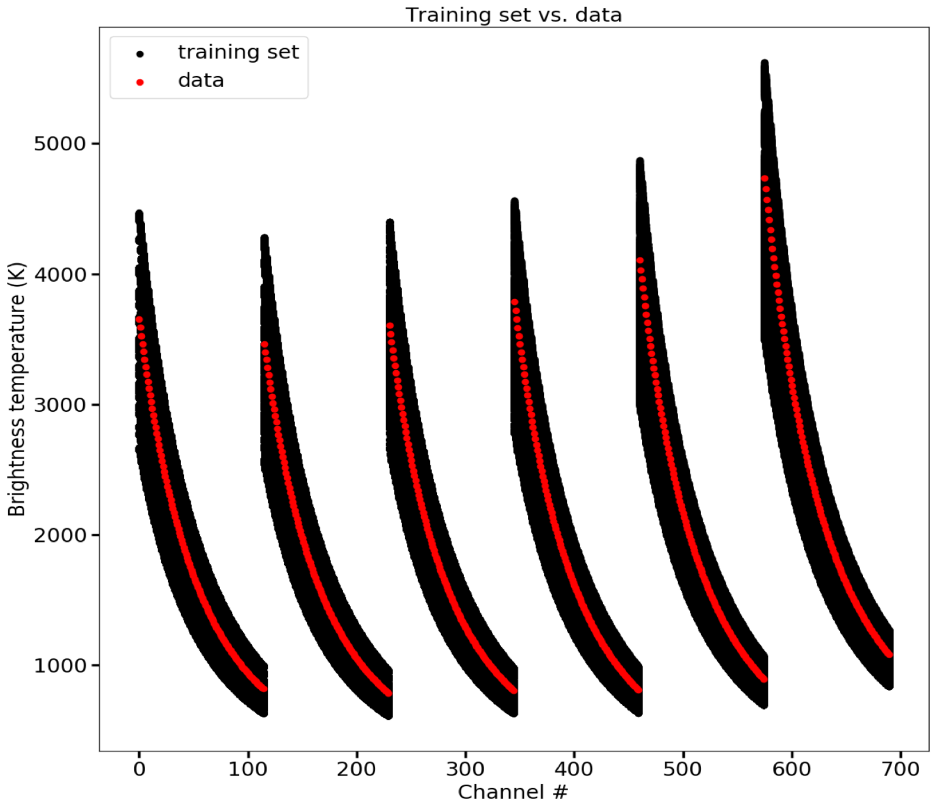
Tauscher et al. (2018)




- Beam-weighted foreground training set for a single rotation angle about one of the 4 antenna pointing directions (top).
- The same training set with its mean subtracted (middle).
- The first 6 SVD basis functions obtained from the training set (bottom).
- The different rotation angles about each antenna pointing direction are part of the same training set so that SVD can pick up on angle-dependent structure and imprint it onto the basis functions.

# EXPERIMENTAL DESIGN: INCLUDING DRIFT SCAN INTO THE LIKELIHOOD FUNCTION

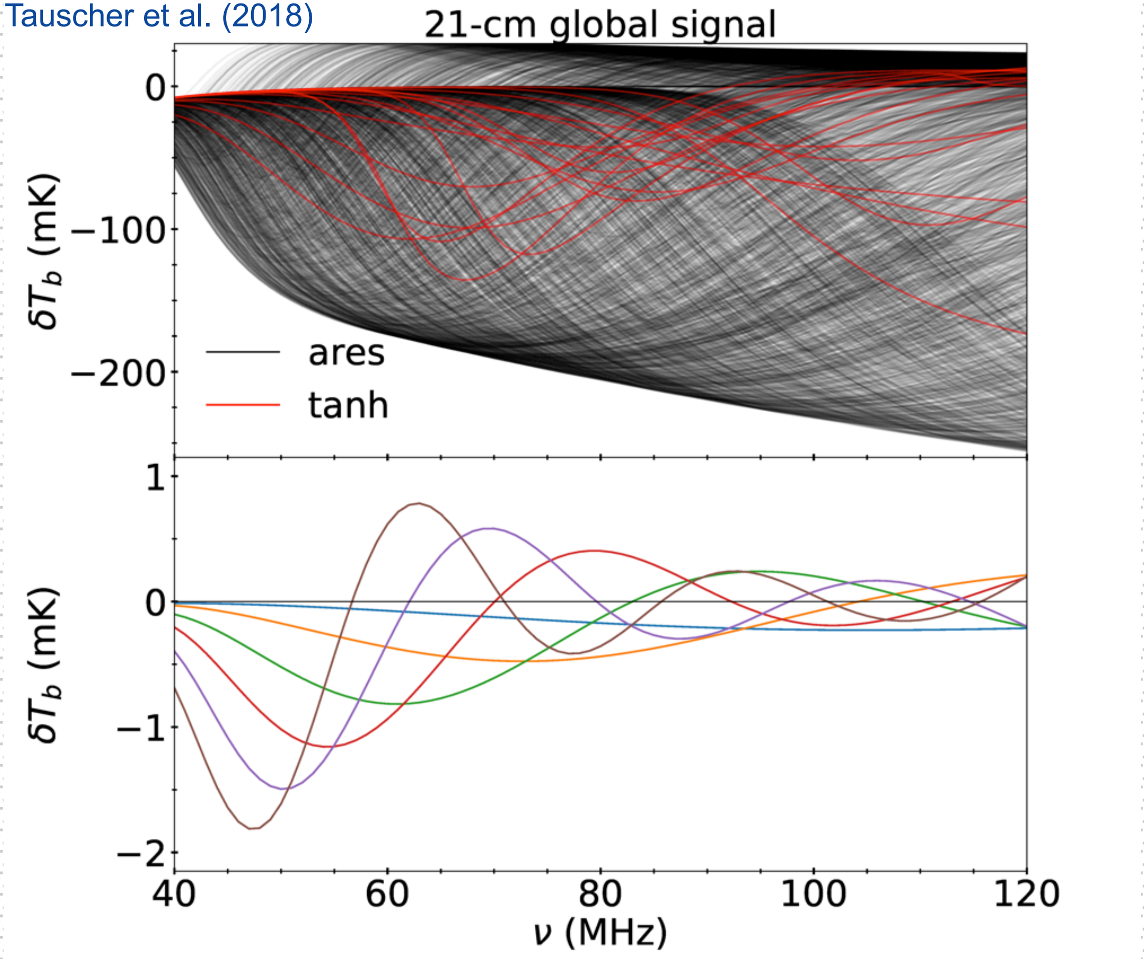
Preliminary (see also Tauscher's poster)




- Beam-weighted foreground training set for each LST bin.
- For a zenith pointing antenna from Earth, the drift scan data from different times are part of the same training set so that SVD can pick up on LST-dependent structure and imprint it onto the basis functions.

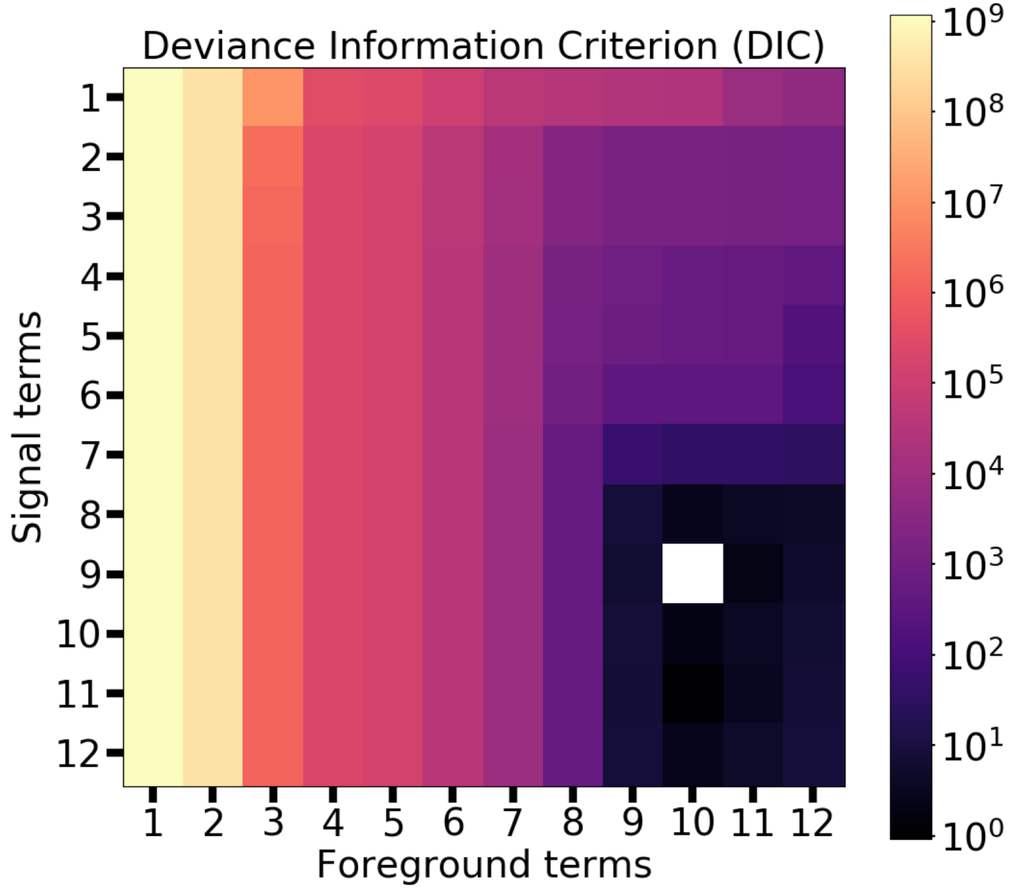
# GLOBAL 21-CM SIGNAL TRAINING SET

Tauscher et al. (2018)




- The signal training set used for our analysis was generated by running the *ares* code 7 × 105 times within reasonable parameter bounds in order to fill the frequency band.
- The top panel shows a thinned sample of that set (black curves). The SVD modes are ordered from most to least important.
- The modes are normalized so that they yield 1 when divided by the noise level, squared, and summed over frequency, antenna pointing, and rotation angles about the antenna pointing.

# MODEL SELECTION: OPTIMIZING THE NUMBER OF SIGNAL AND SYSTEMATIC MODES



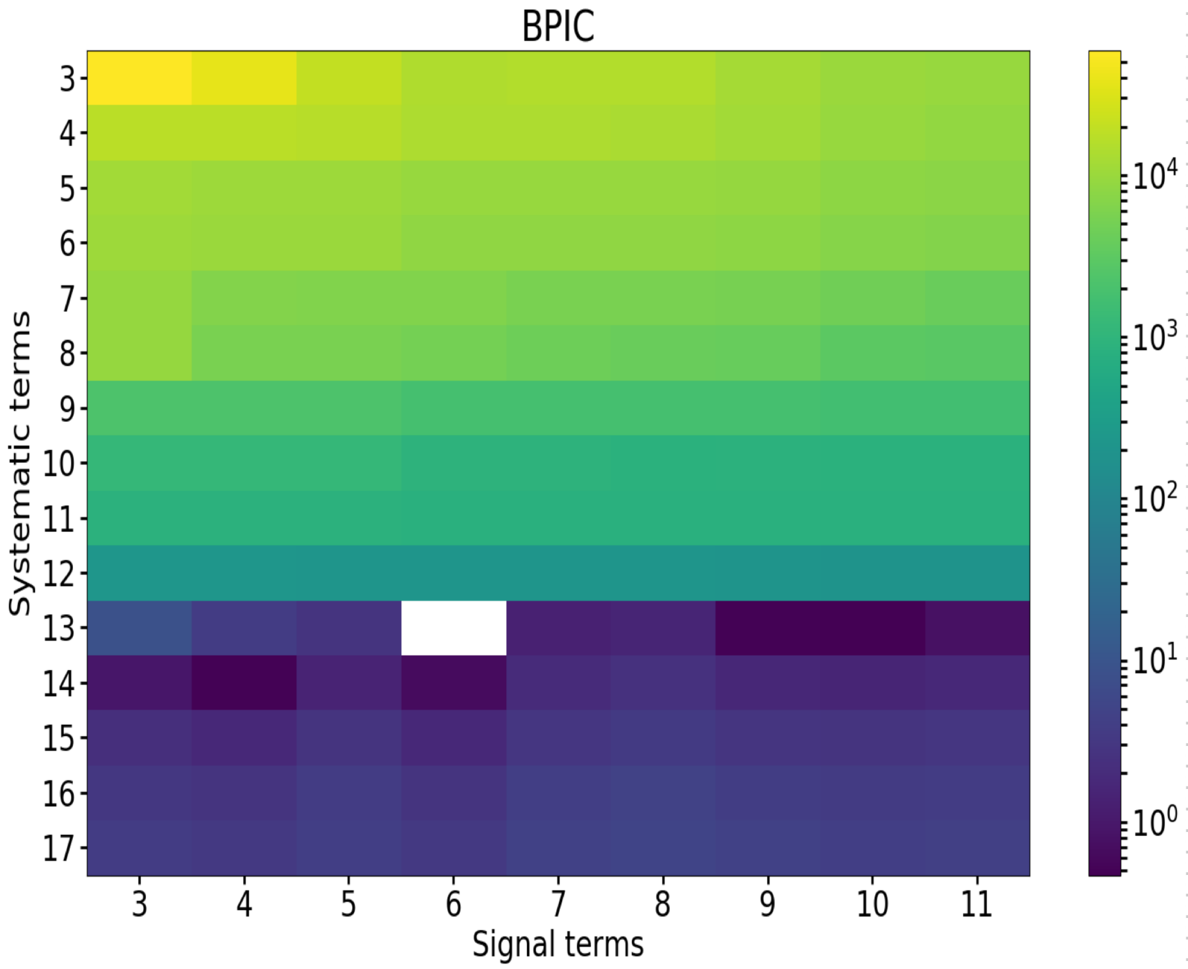

Tauscher et al. (2018)

- Grid of values of the Deviance Information Criterion (DIC).

$$\mathrm{DIC} = -2\ln\mathcal{L}_{\max} + 2p$$

- The colors indicate the difference between the DIC and its minimal value, marked by the white square.
- This same process can be done with any information criteria (BIC, AIC, BPIC, etc.).
- Although only a 12 × 12 grid is shown here, all of the information criteria were calculated over a 60 × 30 grid.

# MODEL SELECTION:
# ANOTHER EXAMPLE USING BPIC




- Grid of values of the Bayesian Predictive Information Criterion (BPIC; Ando 2007).

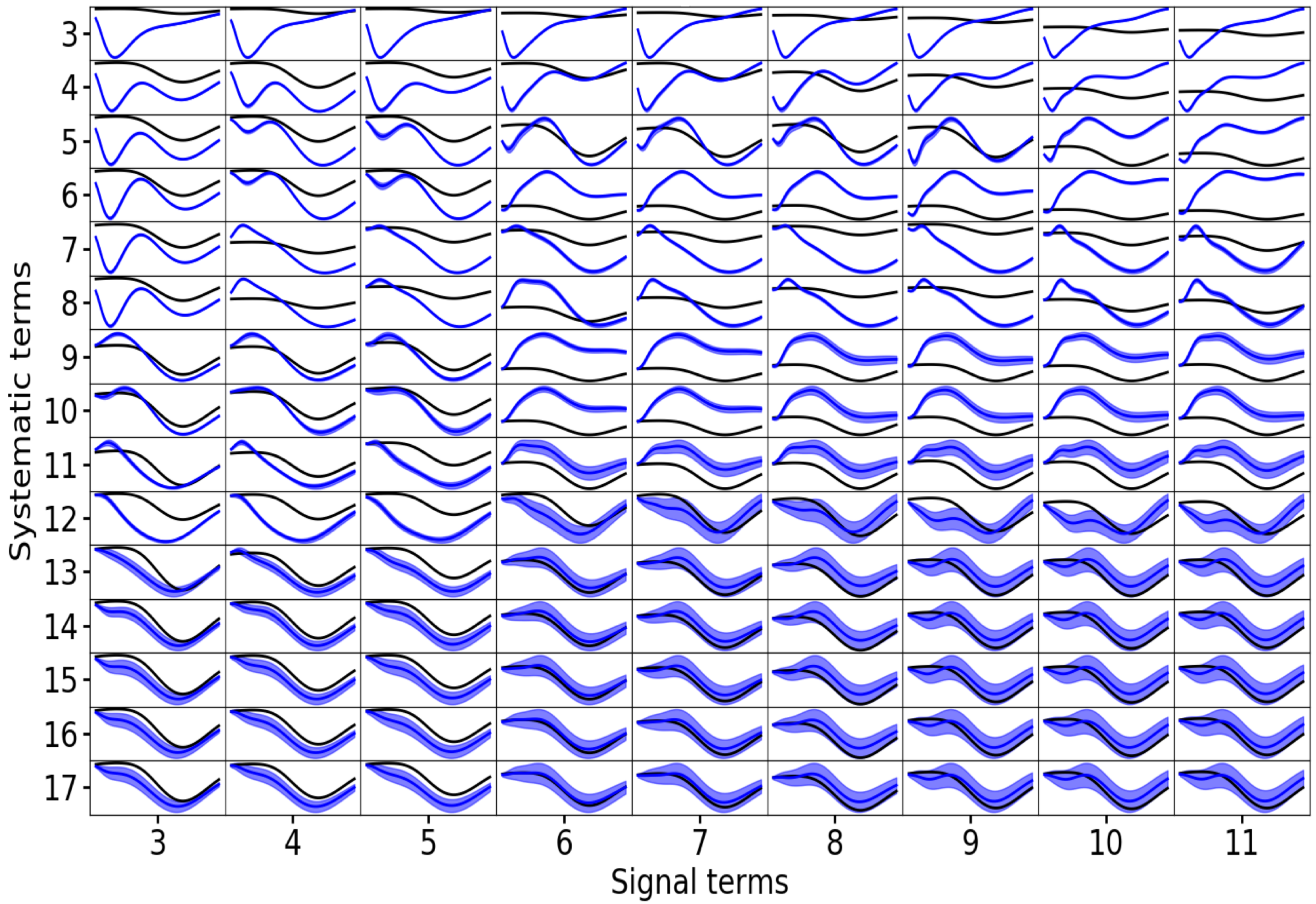
$$\mathrm{BPIC} = \boldsymbol{\delta}^T \boldsymbol{C}^{-1} \boldsymbol{\delta} + N_p + 2\ \mathrm{Tr}(\boldsymbol{C}^{-1} \boldsymbol{\Delta} \boldsymbol{C}^{-1} \boldsymbol{D})$$

$$\boldsymbol{\Delta} = \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T,\ \ \boldsymbol{\delta} = \boldsymbol{G}\boldsymbol{\xi} - \boldsymbol{y},\ \ \text{and}\ \ \boldsymbol{D} = [\mathrm{diag}(\boldsymbol{\delta})]^2$$

where y is the full data vector. See further definitions in Tauscher et al (2018).

- The colors indicate the difference between the BPIC and its minimal value, marked by the white square.

# MODEL SELECTION:
# ANOTHER EXAMPLE USING BPIC

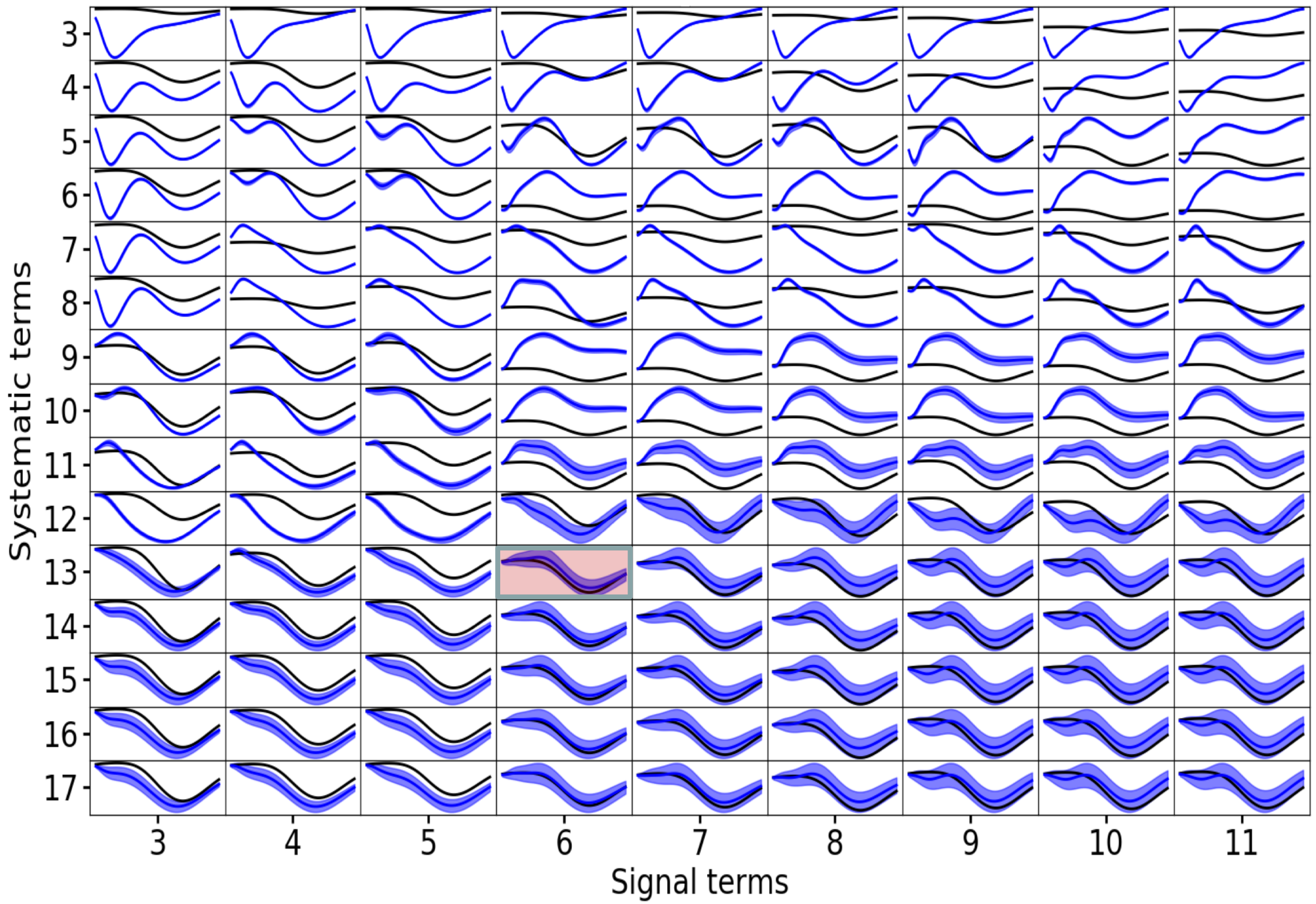



**Signal Extraction optimization:**

The **black line** for all panels is the input 21-cm signal.

The blue bands are the pipeline reconstructions of the signal for a given number of SVD signal and systematic parameters/modes.

# MODEL SELECTION:
# ANOTHER EXAMPLE USING BPIC



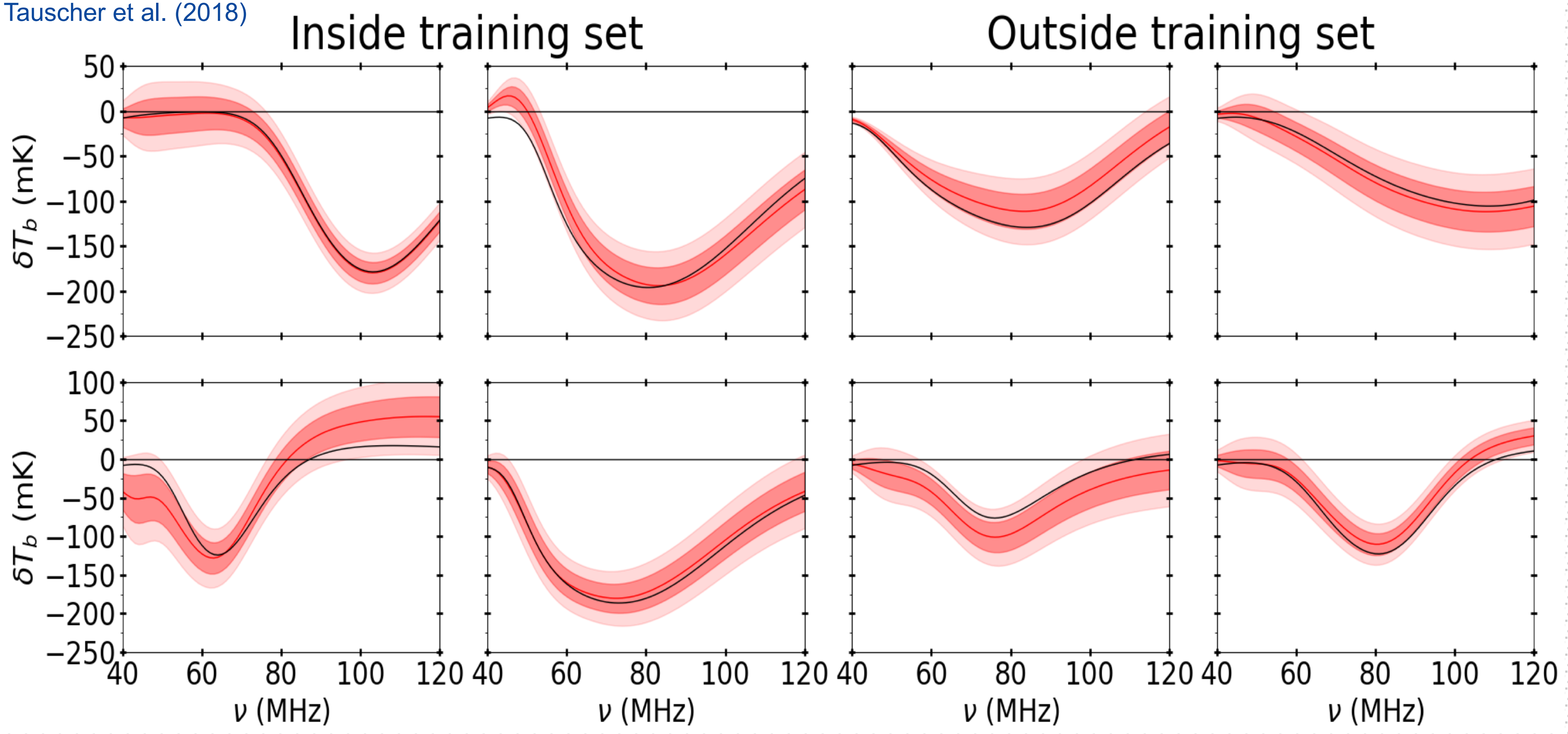

**Signal Extraction optimization:**

The **black line** for all panels is the input 21-cm signal.

The blue bands are the pipeline reconstructions of the signal for a given number of SVD signal and systematic parameters/modes.

# SIGNAL EXTRACTION WITH THE CODE PYLINEX

Tauscher et al. (2018)

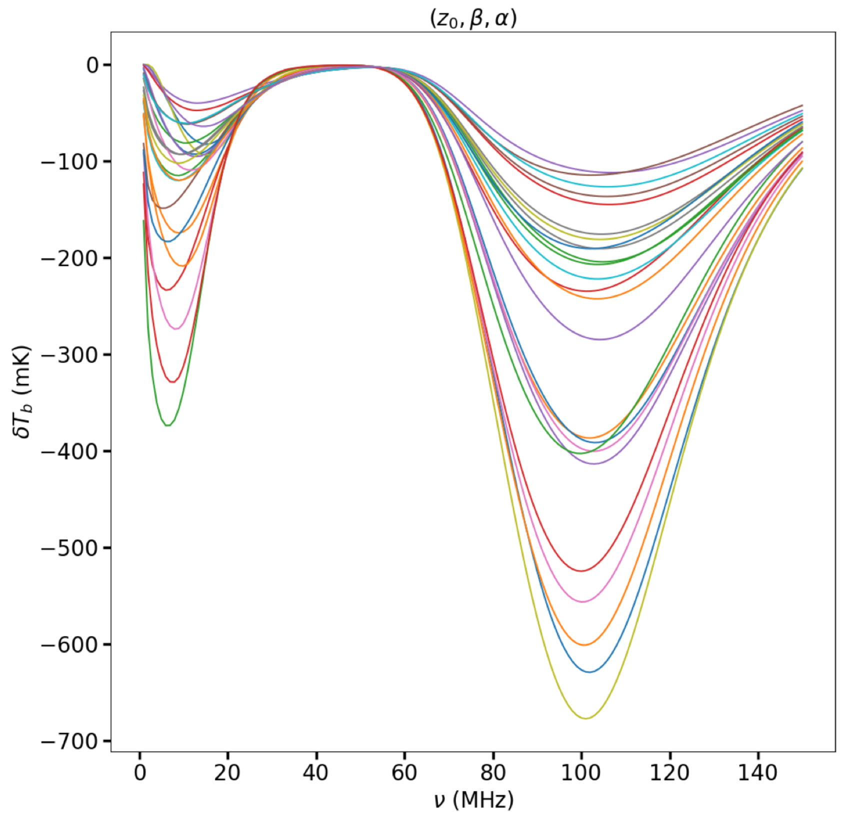


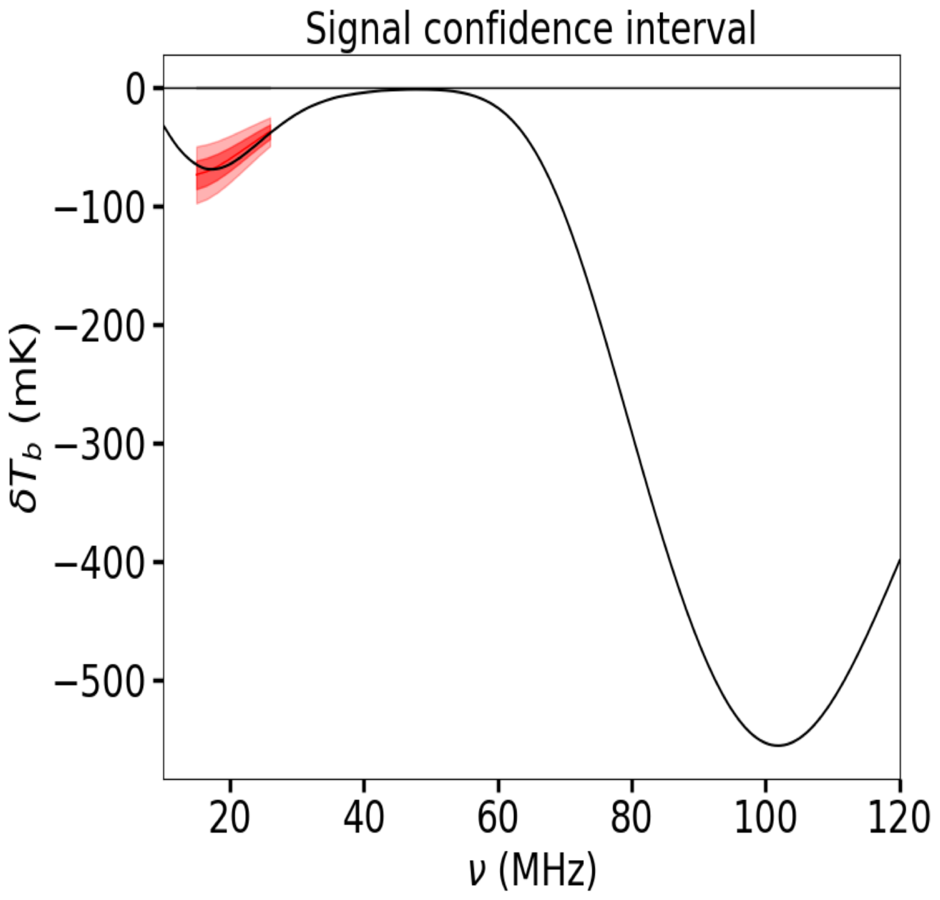

**Signal Estimates** from linear models defined by SVD eigenmodes. The **black** curves show the input signals, the red curves the signal estimates, the dark/light red bands the posterior 68/95% confidence intervals.

The input signals for the 4 left plots came from the *ares* signal training set, and the 4 on the right from the tanh model (see e.g. Harker et al. 2016).

See the pylinex in this link: https://bitbucket.org/ktausch/pylinex

# SIGNAL EXTRACTION WITH THE CODE PYLINEX



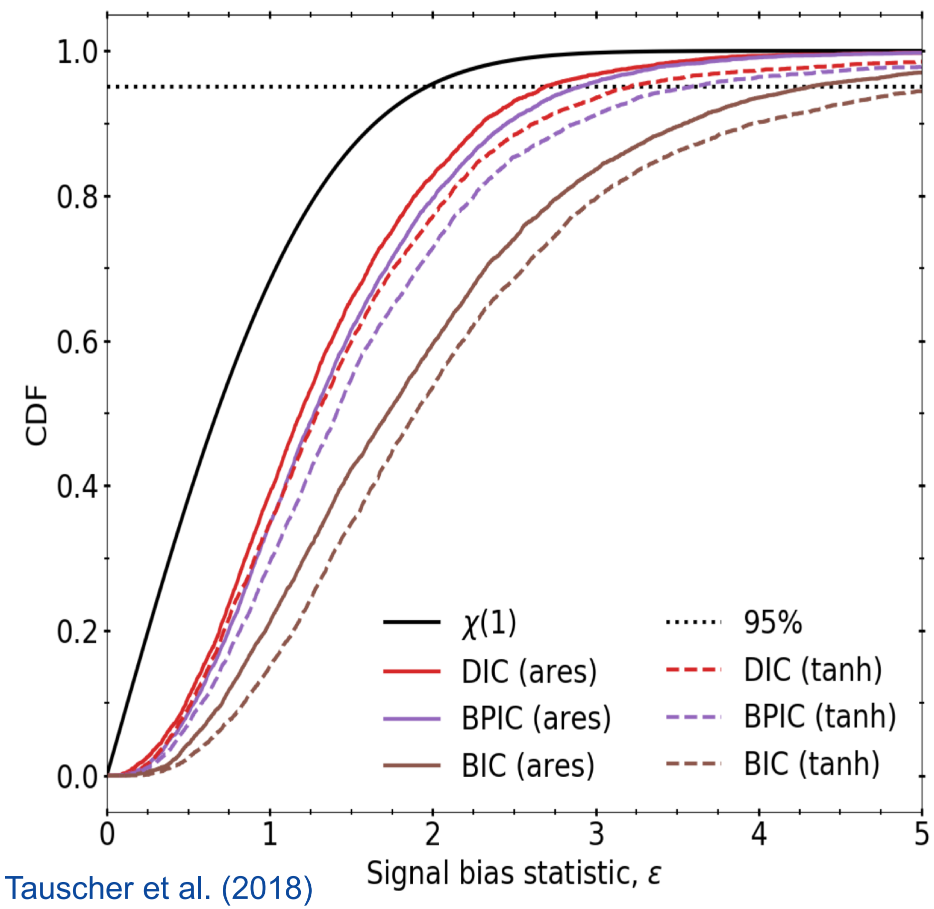

**Example of training set with non-standard cooling rates** with ares allowing larger amplitudes consistent with that of EDGES.

Including both the dark ages and the cosmic dawn troughs.

For a given input signal (**black** curve), the dark/light red bands correspond to the signal estimate for DAPPER in the range 15-26 MHz.

See the pylinex in this link: https://bitbucket.org/ktausch/pylinex

# SIGNAL BIAS STATISTIC




Tauscher et al. (2018)

- The signal bias statistic is a measure of the root mean square error weighted bias of the signal fit:

$$\varepsilon_{21\text{-cm}} = \sqrt{\frac{\boldsymbol{\delta}_{21\text{-cm}}^T \boldsymbol{C}^{-1} \boldsymbol{\delta}_{21\text{-cm}}}{N_\nu}}$$

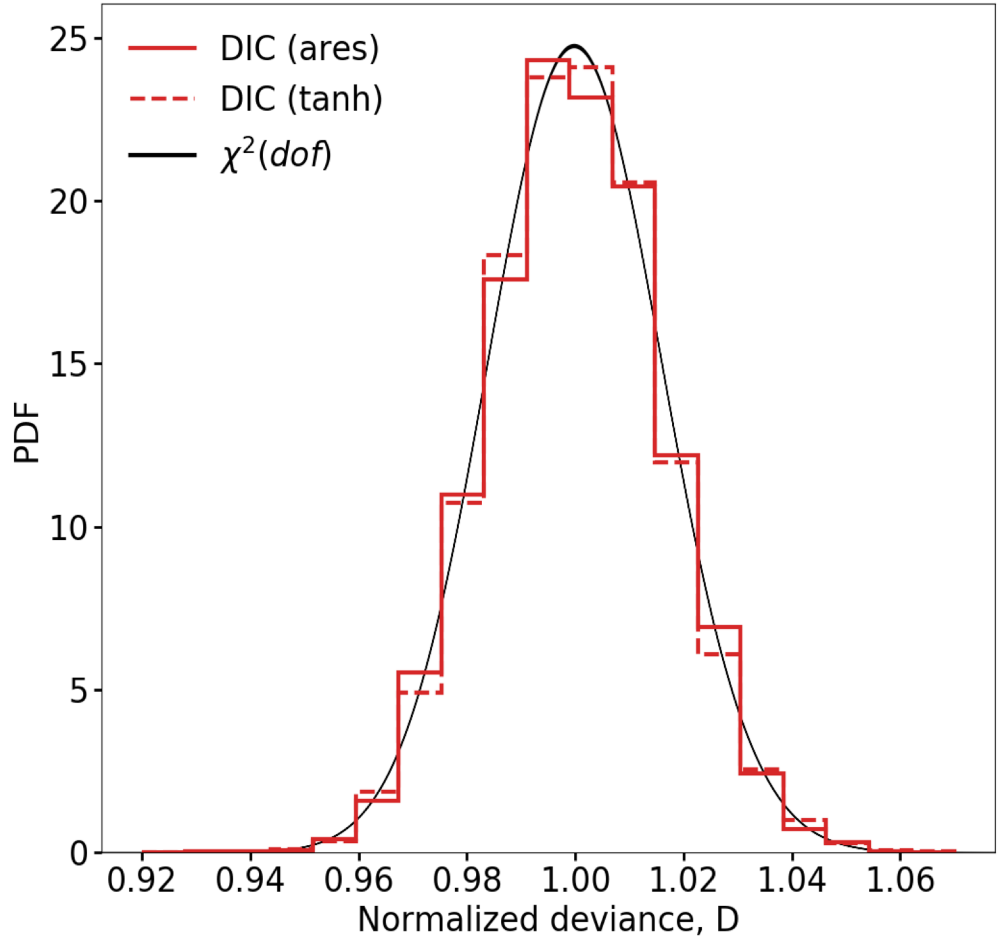
- Estimate of the Cumulative Distribution Function (CDF) of the signal bias statistic from 5000 input simulated datasets.

- A bias statistic of ε roughly corresponds to a bias at the εσ level. The **solid black** reference line is for the distribution which associates 1σ with 68% confidence and 2σ with 95%.

- To guide the eye, the **dotted black** line indicates the 95% level.

# NORMALIZED DEVIANCE

Tauscher et al. (2018)



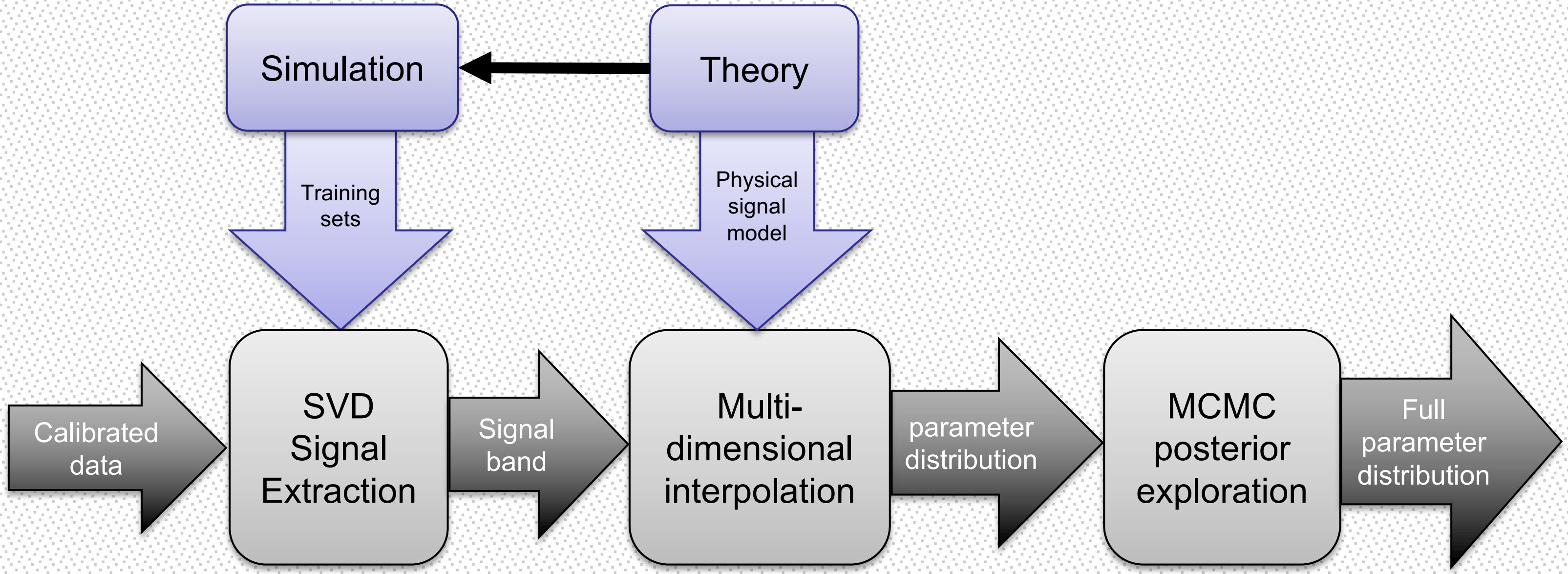

- The deviance normalized by the degrees of freedom contains information about how well the training sets fit the data:

$$D = \frac{\boldsymbol{\delta}^T \boldsymbol{C}^{-1} \boldsymbol{\delta}}{N_{\text{dof}}}$$

- Histogram of the Probability Distribution Function (PDF) for 5000 values of the normalized deviance from fits with different input signals, beam-weighted foregrounds, and noise when using the DIC to choose the best model.

- D should follow a distribution approximated by the extremely thin black region, which is a combination of chisquare distributions associated with the range of degrees of freedom chosen for the extractions.

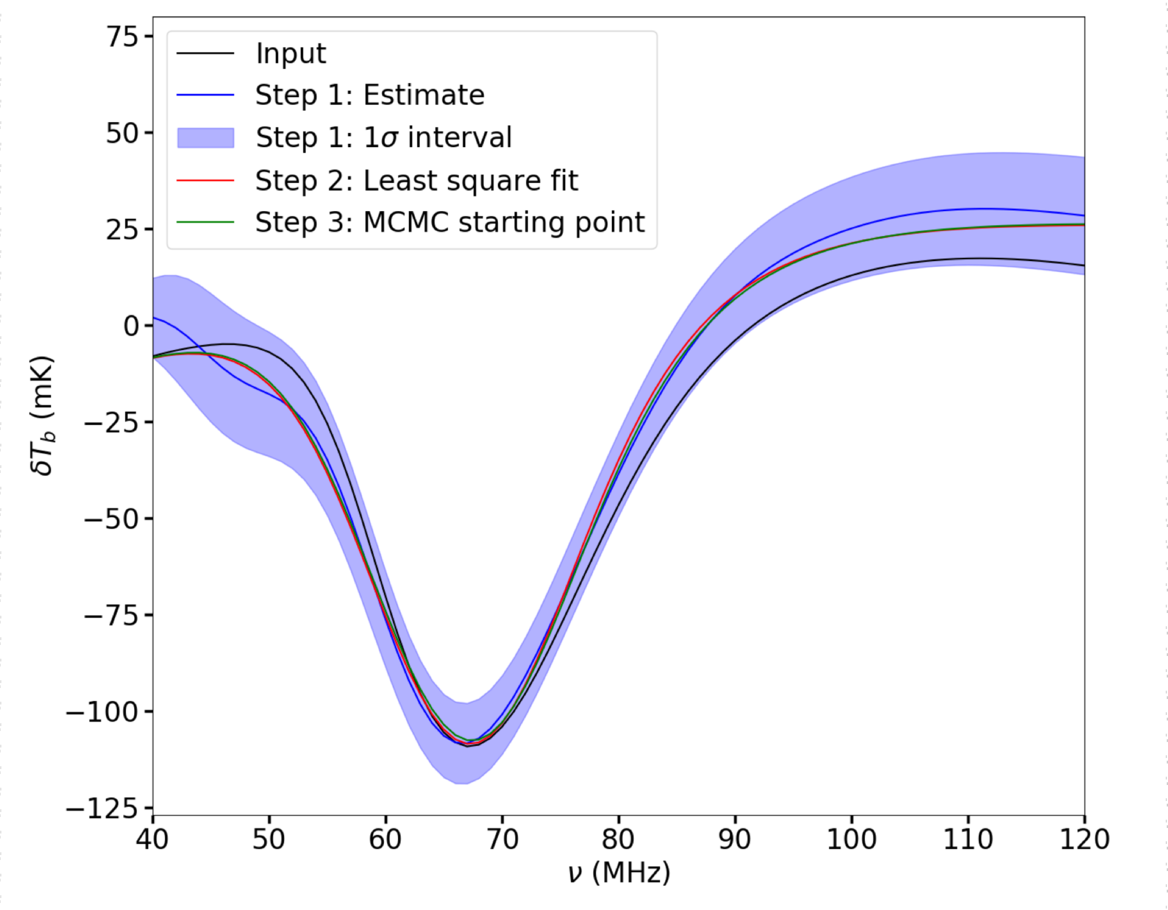
# SVD/MCMC DATA ANALYSIS PIPELINE (PRELIMINARY)




- After extracting the signal in frequency space in the first step of the pipeline we need to transform this result into a constraint in physical parameter space.
- For this, we use a multi-dimensional interpolation using a Delaunay mesh for the change in parameter space and then a Markov Chain Monte Carlo search to constrain the full probability distribution.

(Rapetti et al. 2018, in preparation)

# MULTI-DIMENSIONAL INTERPOLATION USING A DELAUNAY MESH (PRELIMINARY)




(Rapetti et al. 2018, in preparation)

- We generalize linear interpolation to arbitrary input and output dimensions.
- We use this interpolation to perform a least square fit (red line) using the training set.
- Importantly, note that having an starting point (green line) within the estimated error (blue band) provided by the first (very fast) step of the pipeline is crucial for the convergence of the MCMC in a vast parameter space where we do not have otherwise any prior information on the solution and its uncertainty (for the jump proposal).

# CONCLUSIONS

- A challenge of extracting the global 21-cm signal is the large foregrounds.
- However, unlike the foregrounds, the signal is spatially uniform, has well-characterized spectral features, and is unpolarized.
- We benefit from these differences using our novel approach for signal extraction and physical parameter constraints, using an SVD/MCMC pipeline.
- We obtain a highly significant improvement by using a pioneering experimental design of induced polarization and we can do the same with a time series drift scan. Note that these are not mutually exclusive.
- Our pipeline can be used for both lunar orbit and lunar surface low-frequency radio telescopes.
- We are also working on running our pipeline on current/ongoing ground based data from EDGES and CTP using our Pattern Recognition/Information Criteria/MCMC pipeline to measure the expected absorption features in the Global 21-cm spectrum.