Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 16 March 2009 (MN LATEX style file v2.2)

Non-parametric foreground subtraction for 21cm epoch of reionization experiments

Geraint Harker,[1⋆] Saleem Zaroubi,[1] Gianni Bernardi,[1] Michiel A. Brentjens,[2] A. G. de Bruyn,[1,2] Benedetta Ciardi,[3] Vibor Jelić,[1] Leon V. E. Koopmans,[1] Panagiotis Labropoulos,[1] Garrelt Mellema,[4] André Offringa,[1] V. N. Pandey,[1] Joop Schaye,[5] Rajat M. Thomas[1] and Sarod Yatawatta[1,2]

[1]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700AV Groningen, the Netherlands*
[2]*ASTRON, Postbus 2, 7990AA Dwingeloo, the Netherlands*
[3]*Max-Planck Institute for Astrophysics, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany*
[4]*Department of Astronomy and Oskar Klein Centre for Cosmoparticle Physics, AlbaNova, Stockholm University, SE-106 91 Stockholm, Sweden*
[5]*Leiden Observatory, Leiden University, PO Box 9513, 2300RA Leiden, the Netherlands*

16 March 2009

ABSTRACT

One of the problems facing experiments designed to detect redshifted 21cm emission from the epoch of reionization (EoR) is the presence of foregrounds which exceed the cosmological signal in intensity by orders of magnitude. When fitting them so that they can be removed, we must be careful to minimize 'overfitting', in which we fit away some of the cosmological signal, and 'underfitting', in which real features of the foregrounds cannot be captured by the fit, polluting the signal reconstruction. We argue that in principle it would be better to fit the foregrounds non-parametrically – allowing the data to determine their shape – rather than selecting some functional form in advance and then fitting its parameters. Non-parametric fits often suffer from other problems, however. We discuss these before suggesting a non-parametric method, Wp smoothing, which seems to avoid some of them.

After outlining the principles of Wp smoothing we describe an algorithm used to implement it. Some useful results for implementing an alternative algorithm are given in an appendix. We apply Wp smoothing to a synthetic data cube for the Low Frequency Array (LOFAR) EoR experiment. This cube includes realistic models for the signal, foregrounds, instrumental response and noise. The performance of Wp smoothing, measured by the extent to which it is able to recover the variance of the cosmological signal and to which it avoids the fitting residuals being polluted by leakage of power from the foregrounds, is compared to that of a parametric fit, and to another non-parametric method (smoothing splines). We find that Wp smoothing is superior to smoothing splines for our application, and is competitive with parametric methods even though in the latter case we may choose the functional form of the fit with advance knowledge of the simulated foregrounds. Finally, we discuss how the quality of the fit is affected by the frequency resolution and range, by the characteristics of the cosmological signal and by edge effects.

Key words: cosmology: theory – diffuse radiation – methods: statistical – radio lines: general

1 INTRODUCTION

Several current and upcoming facilities (e.g. GMRT,[1] MWA,[2] LOFAR,[3] 21CMA,[4] PAPER,[5] SKA[6]) aim to detect redshifted 21cm line emission from the epoch of reionization (EoR). One problem all such experiments face is to disentangle the desired cosmological signal (CS) from foregrounds which are orders of magnitude larger (Shaver et al. 1999). It is hoped and expected that these foregrounds will be smooth as a function of frequency, while the signal we wish to detect will fluctuate on small scales (Shaver et al. 1999; Di Matteo et al. 2002; Oh & Mack 2003; Zaldarriaga, Furlanetto & Hern-

⋆ E-mail: harker@astro.rug.nl

[1] Giant Metrewave Telescope, http://www.gmrt.ncra.tifr.res.in/
[2] Murchison Widefield Array, http://www.haystack.mit.edu/ast/arrays/mwa/
[3] Low Frequency Array, http://www.lofar.org/
[4] 21 Centimeter Array, http://web.phys.cmu.edu/~past/
[5] Precision Array to Probe the EoR, http://astro.berkeley.edu/~dbacker/eor/
[6] Square Kilometre Array, http://www.skatelescope.org/

quist 2004). If we subtract the smooth component, then the residual will contain contributions from fitting errors (hopefully small), the signal (hopefully largely intact) and noise. Because 21cm emission is line emission, redshift information translates into spatial information along the line of sight (modulo redshift space distortions), thus in principle allowing us to carry out 21cm tomography. In practice, however, for the current generation of instruments such as LOFAR or MWA, the noise per resolution element is expected to exceed the signal by a factor of several, and a spatial resolution element is expected to be of order the size of interesting features of the signal. Current experiments therefore aim to measure statistics such as the global signature of reionization or the power spectrum of 21cm emission. We wish to find foreground subtraction algorithms which do not introduce a large bias into these statistics or make the properties of the noise more awkward.

In this paper we propose a non-parametric technique, 'Wp smoothing' (Mächler 1993, 1995), as a way to fit the foregrounds. This method involves calculating a least-squares fit to the brightness temperature as a function of frequency along each line of sight, subject to a penalty on changes in curvature.

In previous work in which we tested our ability to extract properties of the EoR signal from a simulated data cube with realistic foregrounds (Jelić et al. 2008; Harker et al. 2009) we took a different approach, assuming a smooth functional form for the foregrounds (for example, a third-order polynomial), and then fitting the parameters of this function for each line of sight. This permitted reasonable recovery of the CS, but left us with some concerns. Firstly, the function has to be carefully chosen, both to be able to capture the shape of the foregrounds and to have the right amount of freedom: for example, in our simulations a second-order polynomial has insufficient freedom and produces biased fits, while a fourth-order polynomial has too much freedom and 'fits out' some of the signal. Secondly, we knew the original foregrounds by construction and could use this fact to test our recovery, whereas in the real observations the data themselves will provide the best estimate of the foregrounds. This latter point suggests using a non-parametric fit in which the shape of the fit is 'chosen' by the data.

We must also be careful in our selection of a non-parametric method, however. One could consider using, for example, 'smoothing splines': piecewise polynomial functions which minimize the sum of the squared residuals and a term which measures the integrated squared curvature. A smoothing parameter adjusts the relative weight of the least-squares term and the curvature term. If the least-squares term is given a large weight then the smoothing spline becomes an interpolating function, passing through all the data points, which is clearly undesirable. As the curvature term is given larger weight, the smoothing spline becomes closer to being a straight line, which leads to a systematic bias in the estimate of the foregrounds if they have any curvature. In practice, this would not be a problem if there were some intermediate value of the smoothing parameter which led to acceptable fits, or if there were a well defined procedure for choosing a smoothing parameter for a given problem. We have found, though, that there is no value for which we do not see overfitting, large bias or both.

A comparison of results using several methods, including Wp smoothing, polynomial fitting and smoothing splines, may be found in Section 4. Here we show that Wp smoothing overcomes the problems posed by parametric fits and by other non-parametric methods. As in the case for smoothing splines, for Wp smoothing we must specify the value of a smoothing parameter, λ. We suggest a way of choosing λ and examine its effects on our results, then show that some statistical properties of the signal can be extracted well after removal of the foregrounds using Wp smoothing. Before that, in Section 2, we start by briefly describing the simulations on which our results are based. Then, in Section 3, we lay some groundwork by sketching the mathematical basis of our method, and showing how we solve the differential equation which the Wp smoother fulfils. An appendix gives some intermediate results that may be useful for others who may wish to solve the equation by a different route. Some conclusions are offered in Section 5.

2 SIMULATIONS OF EOR DATA

We test our fitting techniques on the same synthetic data cubes as we have used in previous work (Harker et al. 2009). They have three components: the CS, the foregrounds and the noise. The data cube consists of spatial slices of 256^2 pixels, representing an observing window with an angular size on the sky of $5^\circ \times 5^\circ$. This corresponds to a square of side $624\ h^{-1}\ \mathrm{Mpc}$ (comoving) at $z = 10$ in the cosmology assumed in the simulation. There are 170 such slices, spaced at intervals of $0.5\ \mathrm{MHz}$ in observing frequency between 115 and $200\ \mathrm{MHz}$. These frequencies correspond to redshifts of the 21cm line of between 11.35 and 6.12. At $150\ \mathrm{MHz}$, $\Delta\nu = 0.5\ \mathrm{MHz}$ corresponds to $\Delta z \approx 0.03$, or a comoving radial distance of around $7\ h^{-1}\ \mathrm{Mpc}$.

We estimate the CS primarily using the simulation f250C of Iliev et al. (2008). The distribution of dark matter in a $100\ h^{-1}\ \mathrm{Mpc}$ box was followed using 1624^3 particles on a 3248^3 mesh, and the ionization fraction was then calculated in post-processing on a 203^3 mesh. The parameters of the assumed cosmology were $(\Omega_\mathrm{m}, \Omega_\Lambda, \Omega_\mathrm{b}, h, \sigma_8, n)$=(0.24, 0.76, 0.042, 0.73, 0.74, 0.95). A datacube of $203 \times 203 \times 3248$ points was generated from the periodic simulation boxes according to the method described by Mellema et al. (2006), where the long dimension is the frequency dimension. We then tiled copies of this cube in the plane of the sky in order to fill our observing window, before interpolating onto our $256 \times 256 \times 170$ grid. The tiling means that there are periodic repetitions in the CS in the plane of the sky, which may introduce problems if we were to study spatial statistics, for example the power spectrum. We do not study such statistics here, however. We are interested mainly in how well the signal is recovered given the foregrounds and noise, the maps of which are generated for the full observing window and therefore have no periodic repetition. Two pixels which receive their CS contribution from the same pixel of the original CS map may none the less have very different contributions from the foregrounds and noise.

For comparison with our results using f250C, in Section 4.4 we study two simulations described by Thomas et al. (2009). These use a one-dimensional radiative transfer code (Thomas & Zaroubi 2008) in conjunction with a dark matter simulation of 512^3 particles in a $100\ h^{-1}\ \mathrm{Mpc}$ box. They differ only in the source properties: in one simulation it is assumed that the Universe is reionized by QSOs, and in the other by stars. We label these simulations 'T-QSO' and 'T-star' respectively. Data cubes are derived from the periodic simulation boxes in a similar fashion as was done for the f250C simulation.

We use the foreground simulations of Jelić et al. (2008). These incorporate contributions from Galactic diffuse synchrotron and free-free emission, and supernova remnants. They also include unresolved extragalactic foregrounds from radio galaxies and radio clusters.

We include the effects of the instrumental response of LOFAR on the signal and foregrounds by performing a two-dimensional

Fourier transform on each image, multiplying by a sampling function which describes how densely the interferometer baselines sample Fourier space, and then performing an inverse transform. At present we use the same sampling function at all frequencies; in reality, the '*uv* coverage' (the region of the Fourier plane where the sampling function is not zero) changes with frequency, so this amounts to ignoring information from parts of the Fourier plane which are not sampled at all frequencies.

Noise images are produced by generating uncorrelated Gaussian noise at grid points in the Fourier plane where the sampling function is not zero, transforming to the image plane, and then normalizing this noise image so that it has the correct *rms*. The noise *rms* is calculated as in Jelić et al. (2008), and includes a frequency-dependent part from the sky and a frequency-independent part from the receivers, such that at 150 MHz it has a value of 52 mK The instrumental corruptions introduced by the observing process will clearly be rather more complex than we have assumed here; Labropoulos et al. (2009) discuss this in more detail, with a view towards developing a complete end-to-end model of the effects on the signal of foregrounds, the atmosphere and the instrument.

3 THE WP METHOD

In this section we provide some justification for trying Wp smoothing as a foreground fitting technique and briefly review the relevant mathematical results given by Mächler (1993, 1995). We then describe our algorithm for implementing Wp smoothing.

3.1 Background

If pressed to explain what one meant by trying to find a 'smooth' curve that fit some data set, one might be tempted to say, for example, that the curve had no 'wiggles'. A function with constant curvature might well be considered extremely smooth in this sense. In the case of the smoothing splines mentioned in the introduction, however, the roughness of a curve is given by its integrated squared curvature. By this measure, a function with constant moderate curvature could well be computed as being less smooth than an almost straight line with small wiggles superimposed. This is the motivation for considering, instead, the integrated *change* of curvature.

To be more precise, suppose we have a set of observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ which we wish to fit with a smooth function $f(x)$. Each y_i may have an associated error, σ_i. In our context, the x_i are a series of observing frequencies and the y_i the corresponding differential brightness temperature, all at a given point on the sky. σ_i is the *rms* noise in the map at frequency x_i. Given a function $f(x)$, its curvature, defined as the reciprocal of the radius of curvature, is given by $\kappa(x) = f''(x)(1+f'(x)^2)^{-3/2}$, and its standardized change of curvature (or change of log curvature) is given by

$$\frac{\kappa'}{\kappa} = \frac{f'''}{f''} - 3\frac{f'f''}{1+f'^2} \approx \frac{f'''}{f''} \,. \tag{1}$$

The approximation shown holds exactly at local extrema ($f' = 0$) and at inflection points ($f'' = 0$) and is adopted for convenience.

The first thing to note is that the standardized change of curvature becomes singular at inflection points. Thus the number of inflection points that a function possesses is the most important determinant of its roughness, and we need some sort of procedure to specify the number and position of the inflection points of our smoothing function. Once this is done, we need a way to measure the roughness 'apart from inflection points' to finally specify the function. The importance of inflection points is reflected in the name of the method, 'Wp' being short for the German word 'Wendepunkt', meaning 'inflection point'.

Suppose, then, that the inflection points $w_j, j = 1, 2, \ldots, n_w$, are given. Then we may write

$$f''(x) = p_{\mathbf{w}}(x)\mathrm{e}^{h_f(x)} \tag{2}$$

where

$$p_{\mathbf{w}}(x) \equiv s_f(x - w_1)(x - w_2)\ldots(x - w_{n_w}) \,, \tag{3}$$

$s_f = \pm 1$ and h_f is a function as many times differentiable as f''. Now,

$$\frac{f'''}{f''} = \frac{\mathrm{d}}{\mathrm{d}x}\log f'' = (\log p_{\mathbf{w}})' + h_f' \,, \tag{4}$$

or, rearranging,

$$h_f' = \frac{f'''}{f''} - \sum_{j=1}^{n_w}\frac{1}{x - w_j} \,. \tag{5}$$

This separates our measure of roughness into a part which depends on the number and position of the inflection points and a part which depends on the other properties of f.

We may then express the smoothing problem, given the position of the inflection points, as follows. We wish to find the function f which minimizes

$$\sum_{i=1}^{n}\rho_i(y_i - f(x_i)) + \lambda\int_{x_1}^{x_n} h_f'(t)^2\mathrm{d}t \tag{6}$$

where λ is a Lagrange multiplier, the integral term measures the change in curvature 'apart from inflection points' and the function ρ_i determines the size of the penalty incurred when $f(x_i)$ deviates from y_i. For simple least-squares minimization, for example, $\rho_i(\delta) = \frac{1}{2}\delta^2$ for all i, where δ is the difference between the data and the fitting function.

The solution of this minimization problem must then satisfy an ordinary differential equation (ODE) and boundary conditions derived by Mächler (1993, 1995), who also considered more general cases, for example using higher derivatives of h_f in the integral term. The ODE found by Mächler for the Wp smoothing case is as follows:

$$h_f'' = p_{\mathbf{w}}\mathrm{e}^{h_f}L_f \tag{7}$$

where, using the notation $a_+ = \max(0, a)$,

$$L_f(x) = -\frac{1}{2\lambda}\sum_{i=1}^{n}(x - x_i)_+\psi_i(y_i - f(x_i)) \tag{8}$$

and $\psi_i(\delta) = \frac{\mathrm{d}}{\mathrm{d}\delta}\rho_i(\delta)$. The solution must satisfy some simple boundary conditions,

$$h_f'(x_1) = h_f'(x_n) = 0 \,, \tag{9}$$

as well as some rather more problematic boundary conditions,

$$\sum_i \psi_i(y_i - f(x_i)) = \sum_i x_i\psi_i(y_i - f(x_i)) = 0 \,. \tag{10}$$

We may write ψ_i explicitly as $\psi_i(\delta) = \delta$ for least squares, or, taking the errors into account, as $\psi_i(\delta) = \delta/\sigma_i$. Equivalently, each data point is associated with a weight $c_i = 1/\sigma_i$; this is our default

weighting scheme, but we consider other choices in Section 4.3. Alternatively, a more robust method may use

$$\psi_i(\delta) = \begin{cases} C & \text{if } \delta/\sigma_i > C\,, \\ \delta/\sigma_i & \text{if } |\delta/\sigma_i| \leqslant C\,, \\ -C & \text{if } \delta/\sigma_i < -C \end{cases} \tag{11}$$

for some $C > 0$.

Not only are the boundary conditions problematic, but the ODE itself, Equation 7, includes on the right-hand side a contribution from $f(x_i)$ for each x_i, meaning that the equation is not in the 'standard form' assumed by off-the-shelf solvers for boundary value problems (BVPs).

Recall that the minimization is performed with s_f and $\{w_i\}$ fixed. To apply the procedure to an arbitrary data set, then, requires a further minimization over the number and position of the inflection points. We therefore require some method to give a starting approximation for f, f', h_f, h'_f, n_w, $\{w_i\}$ and s_f. For our particular application we need not consider arbitrary data sets: the properties of the foregrounds seem to allow us to achieve acceptable fits with $n_w = 0$. This might be expected if the foregrounds were to consist of a superposition of power laws with varying spectral index, for example arising from different sources along the line of sight. We therefore impose this condition throughout this paper and do not discuss how to perform a minimization over n_w and $\{w_i\}$.

In principle we should also like some method to choose the Lagrange multiplier, λ. Wp smoothing remains well defined for $\lambda \to 0$ and $\lambda \to \infty$. Indeed, an attractive feature of the method is that for $\lambda \to 0$, f does not become an interpolating function as happened for smoothing splines: rather, it becomes the best-fitting function having the given inflection points. Meanwhile, for $\lambda \to \infty$, f becomes the best-fitting polynomial of degree $n_w + 2$ with the given inflection points, rather than becoming a straight line which automatically underestimates the curvature.

Mächler (1993) suggests using the autocorrelation function of the residuals to estimate λ, reducing λ from a large value in stages until the residuals become uncorrelated. This could be problematic for our application, since there may be real correlations in the noise between frequency bands due to the CS we aim to find.

We might note instead that because λ controls the degree of regularization we apply during the fitting, with smaller λ affording a greater degree of freedom in the functional form, a choice of λ expresses our prior knowledge of how smooth we expect the foregrounds to be. If that knowledge is uncertain, a fully consistent approach would be to estimate what level of freedom is justified by the data themselves. Such a framework could also encompass the choice of n_w, which may be viewed as a more important regularization parameter. This sort of problem, and the topic of mixed signal separation in general, is of course the subject of an extensive literature in information theory and Bayesian inference. Unfortunately, Wp smoothing seems to present a rather awkward case for such methods. Since it is already quite computationally expensive to calculate the Wp smoothing solution for even a single value of λ, we have not chosen to go via this route.

We have instead taken a more heuristic approach, smoothing using different values for λ and using our knowledge of the simulated foregrounds to test the quality of the fit according to various criteria. We detail these criteria, and use our results on simulated datacubes to choose the value of λ used for the subsequent parts of the paper, in Section 4.1.

In the following subsection we give some details of the algorithm we use to solve Equation 7. A reader uninterested in these details should skip directly to Section 4, in which we describe our results.

3.2 Algorithm

An algorithm to solve Equation 7 subject to the boundary conditions given by Equations 9 and 10 is reportedly given by Mächler (1989). Since there is no publicly available implementation of this algorithm, and since we will not deal with the most general case, we have experimented with different approaches. The first is to rewrite the differential equation as in Appendix A, such that it can be solved by a standard BVP solver. The second, which we have found to be faster and more stable (though giving identical results) is to discretize the differential equation into a finite difference equation defined on a grid, and then solve the resulting algebraic system using standard methods.

We choose a mesh such that the abscissae of the data points are also mesh points. That is, we have a mesh $X_1, X_2, \ldots, X_N$, where $N \geqslant n$, and where $X_{m_i} = x_i$ for $i = 1, \ldots n$, with $m_1 = 1$ and $m_n = N$. A mesh with two additional points between each pair of data points (that is, with $N = 3n - 2$) seems to be adequate, in that adding more mesh points does not change the solution at the position of the data points to high accuracy.

Let $f(X_i) = f_i$ and $h(X_i) = h_i$ (which implies that $f(x_i) = f_{m_i}$). Further, let $\Delta_j = (X_{j+1} - X_j)(X_j - X_{j-1})$. Then we may discretize Equation 2 as

$$f_{j+1} - 2f_j + f_{j-1} - \Delta_j p_{\mathbf{w}}(X_j)\mathrm{e}^{h_j} = 0 \quad . \tag{12}$$

Similarly, we may rewrite Equation 7 as

$$\begin{aligned} 0 = {} & h_{j+1} - 2h_j + h_{j-1} \\ & - \Delta_j p_{\mathbf{w}}(X_j)\mathrm{e}^{h_j} \left[-\frac{1}{2\lambda} \sum_{i=1}^{n} (X_j - x_i)_+ \psi_i(y_i - f_{m_i}) \right] \end{aligned} \tag{13}$$

where in each case the index j runs from 2 to $N - 1$. The boundary conditions of Equation 9 become

$$h_2 - h_1 = h_N - h_{N-1} = 0\,, \tag{14}$$

while those of Equation 10 become

$$\sum_i \psi_i(y_i - f_{m_i}) = \sum_i x_i \psi_i(y_i - f_{m_i}) = 0\,. \tag{15}$$

We solve the system of Equations 12–15 using the MATLAB routine 'fsolve'. Our method is therefore essentially a relaxation scheme, but one in which the unusual form of Equations 13 and 15 does not allow us to take the shortcuts used by standard relaxation schemes, which exploit the special form of algebraic systems arising from finite difference schemes.

The initial guess for the solution is also important, and a poor guess can greatly increase the execution time. A method for finding an initial guess for a generic dataset would need to provide an estimate of the number and position of the inflection points. We have found, however, that we can fit the foregrounds using estimates with no inflection points – or, to put it another way, no wiggles – i.e. $n_w = 0$. Imposing this condition simplifies the problem. It is necessary to provide an initial guess for f which also has no inflection points within the range being fitted, and we have found that using a power law works reasonably well.

Using this scheme, fitting the foregrounds for one line of sight for our fiducial value of λ (see Section 4.1) takes less than one second on a typical workstation; going to smaller λ does increase the

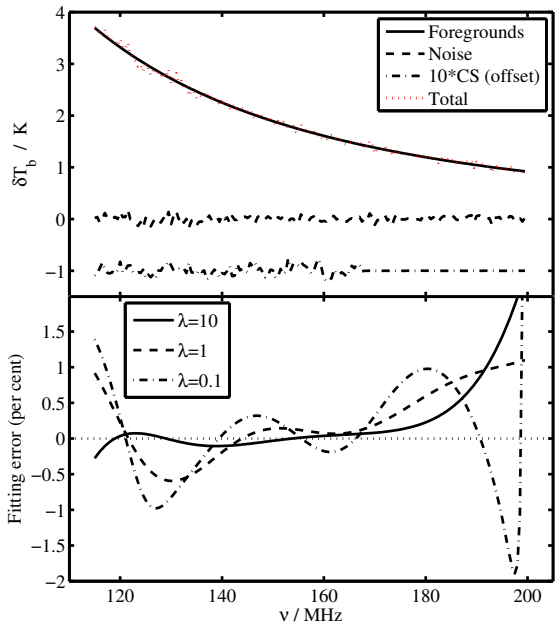

Figure 1. We show, in the top panel, the differential brightness temperature as a function of frequency for a particular line of sight. We also show the individual contributions to this total from the foregrounds, noise and CS. We have multiplied the size of the CS by a factor of 10, and offset the corresponding line by -1 K for clarity. The bottom panel shows the difference between the Wp smoothing estimate of the foregrounds and their true value, expressed as a percentage, for three values of the smoothing parameter λ.

execution time, however. Our simulated data cube has 256^2 lines of sight, and since the fitting for each one is independent the calculation can be trivially split between several processors, meaning processing the cube typically takes of order one hour on our setup.

4 RESULTS

To illustrate the problem we are attacking, in the top panel of Fig. 1 we show the contribution to the differential brightness temperature $\delta T_{\rm b}$ along an example line of sight from the foregrounds, noise and CS. For this particular line of sight, the total intensity is positive at all frequencies. Because at each frequency the mean over all lines of sight in one of our images must be zero, however (since an interferometer cannot measure the mean, which for the foregrounds could be as much as tens or hundreds of kelvin at these frequencies), if we had chosen a different line of sight then we could have seen $\delta T_{\rm b} < 0$ for all ν, or have seen some positive and some negative values due to noise. The latter situation is atypical, however, since the fluctuations in the foregrounds are of order a few kelvin (making the line of sight shown in Fig. 1 fairly typical) whereas the noise fluctuations are of order tens of millikelvin. The size of the CS has been increased by a factor of 10 for the plot, so that the fluctuations are visible; we have also offset the line by -1 K for clarity. The CS is very nearly zero for $\nu \gtrsim 170$ MHz, owing to reionization.

The bottom panel of Fig. 1 shows how well we estimate the foregrounds by applying Wp smoothing to the total signal along this line of sight, for three different values of λ. Though no conclusions can be derived from a single line of sight, we can see that

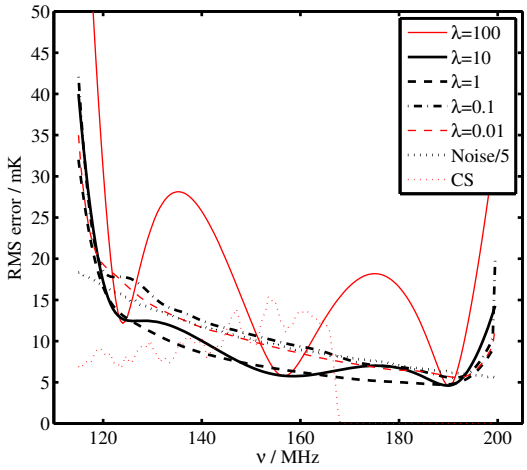

Figure 2. The *rms* difference between the known, simulated foregrounds, and the foregrounds estimated from Wp smoothing for our 256^2 lines of sight, as a function of frequency. The solid, dashed and dot-dashed lines show estimates using different values of the smoothing parameter, λ, as given in the legend. With dotted lines we show the *rms* of the noise, scaled down by a factor of 5 to facilitate comparison, and the *rms* of the CS.

accuracies of around one per cent or better are reached, and this turns out to be quite typical.

In the remainder of this section we compare foreground subtraction using Wp smoothing with that using parametric fitting and smoothing splines, and study how its performance is affected by changes in the frequency resolution and range, in the weights c_i and in the model for the CS. We start, though, by choosing a value for the smoothing parameter, λ, and describing the criteria we use to determine the quality of the fitting.

4.1 Choice of smoothing parameter

Perhaps the most natural way to estimate the quality of the fit is to look at the *rms* difference between the simulated foregrounds, which are known exactly, and the estimates for the foregrounds extracted from the complete data cube. We show this *rms* difference as a function of observing frequency, for five different values of λ, in Fig. 2. We also show the frequency dependence of the noise, which we have scaled down by a factor of 5 for ease of comparison, and the *rms* of the CS. The fact that we must scale the noise for this comparison shows immediately that the fitting errors are much smaller in magnitude than the noise. Indeed, this is a relatively easy target to achieve with parametric or non-parametric fits, and is achieved for all the values of λ shown. Good 'by eye' fits are also easy to obtain for individual lines of sight. Other than for $\lambda = 100$, over most of the frequency range the magnitude of the fitting errors appears to scale roughly with the noise, as one might expect. At the edges, however, the errors become much larger, growing to approximately twice the size of the errors in nearby interior bins. This does not seem unreasonable: for interior points the fit is constrained from both sides, while for edge points it is constrained only from one side. We study edge effects in more detail in Section 4.3.

We have chosen to show the result for $\lambda = 1$, but we find that for values of λ near 1 we obtain very similar results. For example, lines for $\lambda = 0.5$ or 2 would be almost indistinguishable. The

$\lambda = 1$ line therefore represents very nearly the minimum *rms* error we can achieve using this method. For $\lambda = 0.1$ (light smoothing) the fit becomes noticeably worse: on any one line of sight, random features of the noise pull the fitting function around too easily. This just increases the *rms* error by leaking noise into the fitting errors. For $\lambda = 10$ (heavy smoothing) there is also a small increase in the average *rms* error compared to $\lambda = 1$. Oscillations in the error are also clearly visible, however: the excessive smoothing prevents the fitting function from accurately taking the shape of the underlying foregrounds, introducing an additional, systematic error in parts of the frequency range. We will examine this in more detail below when we study the cross-correlation of foreground maps at a given frequency with the fitting residuals. For now, we note merely that this sort of error is potentially more pernicious than a mere increase in the noise, since it allows spatial fluctuations in the foregrounds to leak into the signal.

The results for $\lambda = 100$ and $\lambda = 0.01$, plotted using thin lines, are intended to indicate how the fitting behaves in the limit $\lambda \to \infty$ and $\lambda \to 0$ respectively. For $\lambda = 100$ the oscillations which are also visible in the $\lambda = 10$ fit become very large, resulting in a poor fit. This is not unexpected, since the $\lambda \to \infty$ limit for Wp smoothing for $n_w = 0$ is the best-fitting quadratic function, which we know from previous work to be a poor model for our data compared to, for example, a cubic function. The $\lambda \to 0$ limit is more interesting, since it corresponds to the best-fitting function with no inflection points in the interval. For $\lambda = 0.01$ and $\lambda = 0.1$ the fits are very similar, and do not give an *rms* error much worse than the best value for λ. As we mentioned in Section 3.1, this well behaved limit is one of the attractive features of Wp smoothing.

The contents of Fig. 2 are computed by taking an *rms* over all 256^2 lines of sight in our data cube. The results for $\lambda \leqslant 10$ do not change appreciably if we use only, say, 32^2 lines of sight, and do not depend on the position of the selected sub-region. Only the $\lambda = 100$ result changes: if we choose a sub-region where the foregrounds are relatively intense, the size of the oscillations is reduced considerably, in some cases producing an *rms* very similar to the $\lambda = 10$ result. The oscillations come from regions where the foregrounds are less intense, and where a quadratic function is clearly unable to match the shape of the foregrounds as a function of frequency.

The first objective of the LOFAR EoR key project is simply to make a detection of emission from the EoR, and to find the redshift evolution of the global emission which would be a signature of reionization. If we look at the variance of the residuals after the foregrounds have been subtracted from the data, then subtract the (known) variance of the noise, any remaining variance is expected to arise from fluctuations in the CS. This change in the variance of the fluctuations as a function of redshift constitutes a detection of the global signature of reionization. Fig. 3 shows how well this variance is recovered for different values of the smoothing parameter, λ. The black, dotted line shows the variance of the input CS, while the other three lines show the estimates recovered from the full data cube. We do not plot a line for $\lambda = 0.1$ because it overplots the $\lambda = 0.5$ line almost exactly. For the majority of the redshift range, $z \approx 6$–10, the Wp smoothing with $\lambda = 10$ does reasonably well in recovering the variance of the CS (much larger λ, as expected from Fig. 2, does poorly, the fitting errors adding to the *rms* of the residuals and resulting in a large overestimate of the variance). It does better than $\lambda = 0.5$ or 2 in this range, a property which is not clearly reflected in Fig. 2. In this sense, Fig. 3 does a better job of showing the effect of over-fitting, which reduces the variance of the fitting residuals and causes us to underestimate the CS.

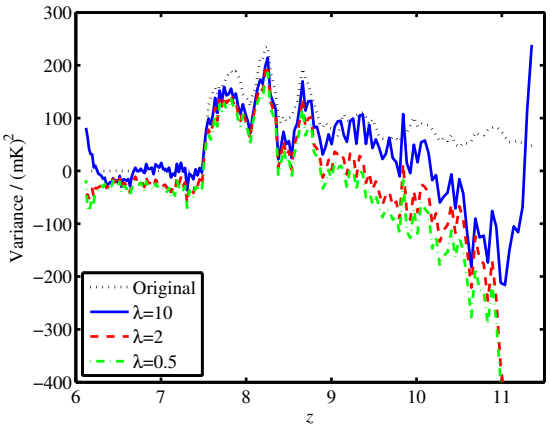

Figure 3. The recovered global signal from the EoR as a function of redshift, for three different values of λ. The variance of the fluctuations of the input CS is shown as the black, dotted line. The other three lines show estimates of this quantity extracted from the simulated data cube. Negative estimates for this variance arise because of over-fitting: the variance of the residuals after foreground subtraction is smaller in this case than the noise variance.

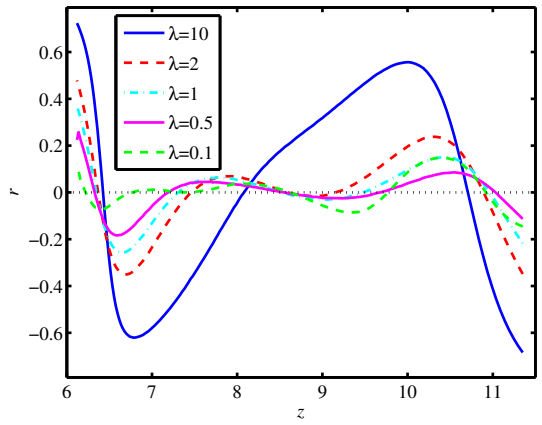

Figure 4. The Pearson correlation coefficient between maps of the foregrounds and the corresponding residual maps (after foreground subtraction from the full data cube) at the same observing frequency. $r = \pm 1$ correspond to perfect correlation and perfect anticorrelation, respectively. The line styles are shown in the same order in the legend as they appear at the far left-hand side of the plot.

By contrast, Fig. 4 shows the effect of under-fitting. Here we show the Pearson correlation coefficient, r, between images of the foregrounds at a given observing frequency, and images of the fitting errors (difference between the fit and the known foregrounds) at the same frequency. If the pixels of the foreground image and of the image of fitting errors have the values a_i and b_i respectively, where $i = 1, \ldots, 256^2$, then r is given by

$$r^2 = \frac{\left[\sum_i (a_i - \bar{a})(b_i - \bar{b})\right]^2}{\sum_i (a_i - \bar{a})^2 \sum_i (b_i - \bar{b})^2} \tag{16}$$

where $\bar{a}$ and $\bar{b}$ are the mean of a_i and b_i respectively. $r = 1$ corresponds to perfect correlation and $r = -1$ to perfect anticorrelation. We see immediately in Fig. 4 that for heavy smoothing, $\lambda = 10$,

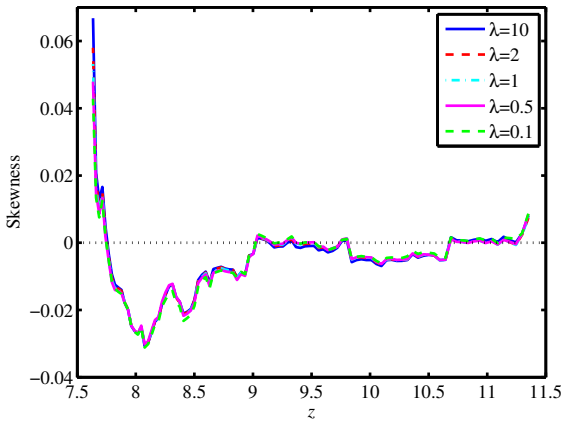

Figure 5. The skewness of the one-point distribution of pixel intensity as a function of redshift, after subtracting the foregrounds using Wp smoothing with different values of λ, and then applying a Wiener deconvolution to each image as in Harker et al. (2009). The lines in the figure are boxcar smoothed to improve the clarity of the plot.

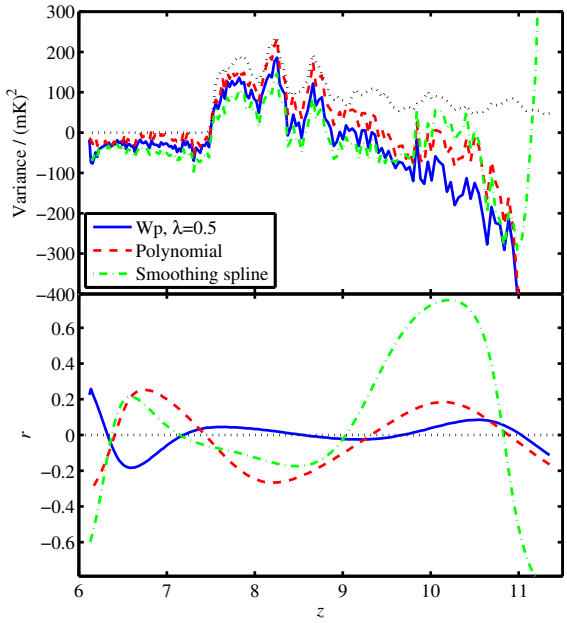

Figure 6. We compare the performance of Wp smoothing with $\lambda = 0.5$ (solid blue lines) with a third-order polynomial fit (dashed red lines) and smoothing splines with $p = 3 \times 10^{-5}$ (dot-dashed green lines; see Equation 17 for a definition of p). The top panel is similar to Fig. 3 and shows how well each method recovers the variance of the fluctuations in the CS (black dotted line) as a function of redshift. The bottom panel is similar to Fig. 4 and shows the Pearson correlation coefficient between the fitting errors and the foregrounds.

there are quite strong correlations ($r = \pm 0.6$) between the foregrounds and the fitting errors in some parts of the frequency range. The level of correlation reduces as λ is reduced, though there appears to be little to choose between $\lambda = 0.5$ and 0.1.

This shows that, as one might expect, heavier smoothing is more likely to allow spatial power to leak from the foregrounds into the residual maps used for later analysis. What constitutes an acceptable level of leakage will depend on the properties of the real foregrounds: if they do not contain more power (especially small-scale power) than our simulated foregrounds, and if *rms* errors of the order of those shown in Fig. 2 can be achieved, even correlations as large as those shown for $\lambda = 10$ in Fig. 4 may not seriously harm the recovery of power spectra or other statistics. None the less, heavy smoothing, which retains more of the desired signal (see Fig. 3) at the expense of systematic correlations with the foregrounds, can be regarded as a more aggressive foreground cleaning strategy. Light smoothing runs more of a risk of cleaning away the signal, but may be less susceptible to systematics, and so may therefore be regarded as a more conservative detection strategy.

The way the recovered variance falls away at high redshift (where the noise is larger) in Fig. 3 seems to suggest that more regularization is required there, i.e. that we may want to consider varying λ as a function of frequency. In practice, doing so does not appear to deliver any significant overall improvement in performance. We do note, though, that Equation 6 implies that a change in λ is degenerate with an overall scaling of the weights, and that we directly address changes in the weighting scheme in Section 4.3.

The recovery at high redshift can be improved, however, if we look at the variance of spatially smoothed maps. This is because the noise and fitting errors are most dominant on small scales, and smoothing removes small scale power. Since this paper is concerned with the quality of the fitting, and since the variance of unsmoothed maps seems to provide a stringent test of this, we do not further explore scale dependence here. The recovery and analysis of the power spectrum will instead be studied in a forthcoming paper.

In Fig. 5 we show how changing λ affects our recovery of the changes in the skewness of the one-point distribution of the signal. As in Harker et al. (2009) we apply a Wiener deconvolution to foreground-subtracted images, and plot the skewness of the distribution of pixel intensity in these images as a function of redshift. The recovered changes in skewness are very robust to altering λ, with nearly identical histories being produced. Reionization causes a fall in the skewness to negative values (recall that if the neutral hydrogen density merely traced the cosmological density field then we would expect positive skewness), followed by a rise at low redshift. We should note that in the foreground simulation used here the images do not possess large skewness, so that correlations between the foregrounds and the fitting errors do not cause a serious problem. If the real foregrounds turn out to be more skewed, Fig. 5 suggests we would be wise to choose a value of λ which minimizes the correlations, perhaps at the expense of reducing the variance of the recovered signal. Since with our current foreground models the extracted skewness does not appear to be sensitive to the value of λ, for the remainder of the paper we do not use the recovered skewness to test our fitting.

There is quite a wide range of reasonable values for λ which achieve a compromise between over- and under-fitting. For the purposes of comparison to other techniques in the remainder of this section we adopt $\lambda = 0.5$, since Fig. 4 shows there seems to be little or no benefit from moving to smaller values (for which the fit is slower to compute).

4.2 Comparison to other fitting methods

We compare the performance of Wp smoothing with $\lambda = 0.5$ with two other techniques in Figs. 6 and 7. The top panel of Fig. 6 shows how well the three methods recover the variance of the fluctuations

in the CS as a function of redshift, as in Fig. 3, while the bottom panel shows the Pearson correlation coefficient between the fitting errors and the foregrounds, as in Fig. 4. The four panels of Fig. 7 show the fitting errors for four different lines of sight. The line styles are the same for both figures: the solid blue lines show the Wp results, the red dashed lines show the results when we estimate the foregrounds by fitting a third-order polynomial in $\log \nu$ to each line of sight, and the green dot-dashed lines show the results using smoothing splines to fit the foregrounds. Smoothing splines are a non-parametric method which we considered as an alternative to Wp smoothing. The smoothing spline fit is a piecewise polynomial function f minimizing

$$p \sum_{i}^{n} c_i [y_i - f(x_i)]^2 + (1-p) \int_{x_1}^{x_n} [f''(x)]^2 \mathrm{d}x \qquad (17)$$

where p is a smoothing parameter. $p = 0$ gives a straight line fit, while for $p = 1$ f becomes an interpolating cubic spline. For Figs. 6 and 7 we used $p = 3 \times 10^{-5}$.

Fig. 6 suggests that the smoothing spline fit does poorly compared to the Wp smoothing: not only does it suppress the variance of the residuals more than Wp smoothing for our chosen values of λ and p over most of the frequency range (a symptom of over-fitting), but it simultaneously produces fitting errors which correlate more strongly with the foregrounds (a symptom of under-fitting). For a small frequency interval near $z = 10$, the smoothing spline fit appears to suppress the variance less than the other methods. This, however, is precisely the interval where the correlations of the errors with the foregrounds are strongest, which illustrates our point about the dangers of foreground leakage. Similarly to the Wp case, we can improve the performance of the smoothing spline fits according to either the over-fitting or under-fitting criterion by tuning p, but this comes at the expense of worse performance according to the other criterion. Wp smoothing therefore appears to be a superior method for this problem.

Comparison to the parametric (third order polynomial) fit gives a more mixed result. For $\lambda = 0.5$ the Wp smoothing loses more of the signal, but induces smaller correlations between the fitting errors and the foregrounds. Wp smoothing does, though, give us the freedom to change λ continuously to trade off performance in these two tests. A similar trade off is possible by changing the order of the polynomial used for the parametric fit, but changing the order in this way corresponds to a rather drastic jump in the properties of the fit, and seems not to be very useful in practice. We must also emphasize that by using Wp smoothing we are only making rather general assumptions about the smoothness of the foregrounds (and, for our current choice of implementation, the number of inflection points of the foregrounds). Clearly, if we were to know the functional form of the foregrounds in advance then we would be justified in parametrically fitting the foregrounds with the correct function. If, though, we can achieve comparable results for realistic simulated foregrounds using parametric or non-parametric methods, it would be preferable to use the non-parametric technique on the observational data in case the real foregrounds do not match our expectations. The fact that Wp smoothing can achieve a fit of parametric quality without assuming a functional form for the foregrounds justifies its use for EoR experiments, and suggests further investigation of non-parametric techniques to address this problem.

The four example lines of sight shown in Fig. 7 are intended to illustrate some of the differences between the methods. The foregrounds differ in amplitude between these lines of sight: from top to bottom, their value at 150 MHz is 1.89, 1.65, 4.93 and -1.14 K.

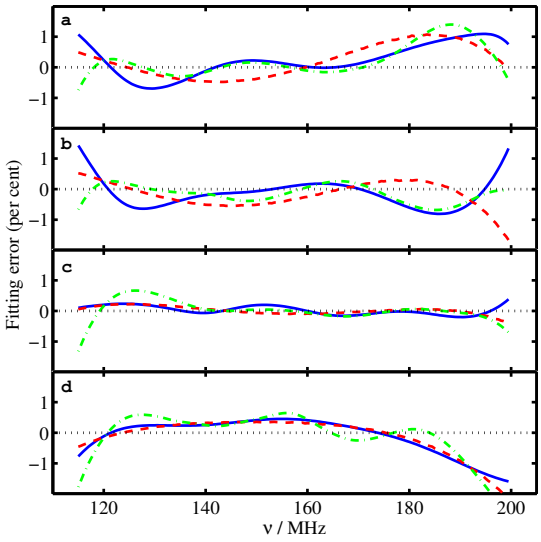

Figure 7. We show the fitting errors along four example lines of sight for Wp smoothing with $\lambda = 0.5$, a third-order polynomial fit and smoothing splines with $p = 3 \times 10^{-5}$. Line styles are as for Fig. 6. From top to bottom, the level of the foregrounds at 150 MHz for each of the lines of sight is 1.89, 1.65, 4.93 and -1.14 K. The top panel shows the same sight line as Fig. 1.

Comparing panels a and b, one may notice that the shape of the error curve for the polynomial fitting is very similar in these two cases, while the Wp smoothing curve differs between the two panels at the high frequency end. This is a manifestation of the systematic errors made by the parametric fit which seem to be alleviated somewhat by non-parametric methods. The line of sight in panel c comes from a point on the sky where the foregrounds are relatively intense. The noise does not scale with the foregrounds, and so the fitting is able to determine the foregrounds more accurately in a relative sense. This suggests that the large amplitude of the foregrounds relative to the CS may be less of a concern than the scale-dependence of their fluctuations on the sky, since small-scale fluctuations which leak into the residual maps because of biased fitting may be confused with the CS. Finally, panel d of Fig. 7 shows how the fits produced by the smoothing spline method are more prone to oscillations than those produced by Wp smoothing or by polynomial fits. The statistical signature of these oscillations is the over-fitting shown by the top panel of Fig. 6. One must be careful not to over-interpret results for individual lines of sight, however, and so in the remainder of the paper we restrict ourselves to statistical comparisons.

4.3 Changes in frequency resolution and weighting

It is very noticeable in Fig. 2 that the errors on the fit become larger at the ends of the frequency range. Similarly, in Fig. 4, while there is a very small cross-correlation between the foregrounds and fitting errors for $z \approx 7$–10 for our $\lambda = 0.5$ fit, the performance degrades slightly at the lowest redshifts (highest frequencies). It would be desirable to have a fit of more uniform quality, since otherwise we truncate the useful frequency range, and since we might worry that an apparent signal is merely a side-effect of more serious foreground contamination at some redshifts than others. It

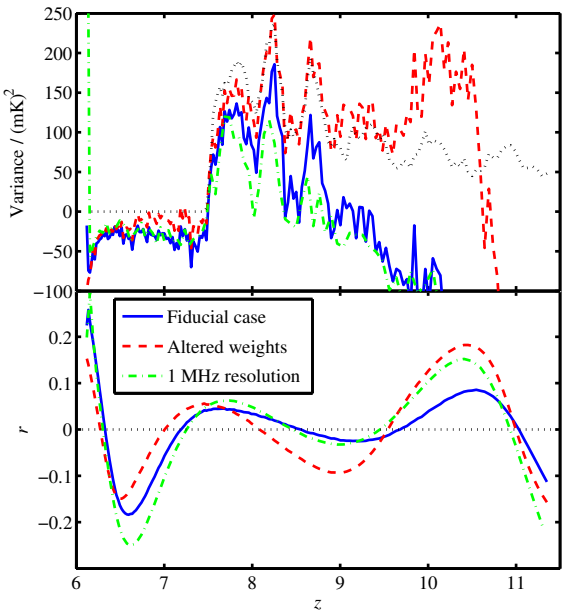

Figure 8. We show an example of the effects of a different weighting scheme and a lower frequency resolution on the recovery of the variance of the CS (top panel), and on the correlation between the fitting errors and the foregrounds (bottom panel). The solid blue line shows the results for $\lambda = 0.5$ and our fiducial weighting scheme and frequency resolution. The dashed red line shows the result if we adjust our weighting scheme to give points near the end more weight, while the dot-dashed green line shows the effect of halving the frequency resolution. Note that the axes cover a smaller range than in Fig. 6.

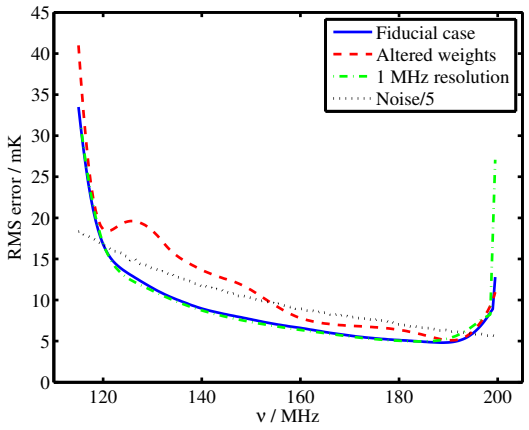

Figure 9. We show the *rms* error of the foreground fitting for our fiducial fit, a fit with modified weights, and a fit for a data cube with half the frequency resolution of our standard data cube. The line styles are as for Fig. 8. The solid and dot-dashed lines almost coincide for most of the frequency range.

seems plausible that adjusting the weights c_i used in the fitting may improve the fit at the ends of the interval at the expense of the interior. Modest changes in the weights (for example, using uniform weights rather than inverse noise weights) have little effect. Large enough changes do have an impact, though, as we show in Fig. 8. Here we compare our fiducial weighting scheme (solid blue line) with an alternative weighting scheme (dashed red line) in which extra weight is given to points near the ends of the interval. To be precise, we multiply the ith 'natural' weight $1/\sigma_i$ by $1/(1-d_i^2)$ where $d_i = 1.7(i-1)/(n-1) - 0.9$. We then normalize the new weights to have the same mean as the fiducial weights, in order that the value of λ can remain unchanged. The top panel of the figure shows the recovered variance, while the bottom panel shows the correlation coefficient between fitting errors and foregrounds, as in Fig. 6.

It seems that this adjustment of the weighting scheme is at least a limited success. The correlation between fitting errors and foregrounds becomes slightly smaller at low redshift, at the expense of increased correlations in the interior of the redshift range. The recovered variance of the signal is, moreover, closer to the original in the most interesting part of the redshift range. Unfortunately, the origin of this improved agreement is not a better fit, but a worse one. This is demonstrated in Fig. 9, in which we show the *rms* error of the foreground fitting. The line styles are the same as for Fig. 8. The modified weighting scheme significantly increases the fitting errors. The improved recovery of the signal variance in Fig. 8 therefore seems to be a fluke caused by leaking more noise into the fitting residuals, and it would be hard to recommend this as a strategy for signal recovery. In fact, after experimenting with various weighting schemes, modifying them seems to be an unpromising avenue: modest changes have a marginal effect, while large changes tend to significantly increase the overall error.

Also shown in Fig. 8 is the effect of reducing the frequency resolution to 1 MHz rather than 0.5 MHz. Though this halves the number of bins, it also reduces the noise per bin by a factor of $\sqrt{2}$. In the top panel we see that the recovered variance is smaller, but this is not due to a poorer fit being achieved (as one can see from Fig. 9): rather, the variance of the original signal itself is reduced when binned up, since adjacent 0.5 MHz frequency slices are decorrelated to some extent. The amount of variance lost by the fitting process is similar in either case. The reduction in the number of data points does, however, degrade the quality of the fit in the sense that the correlation between fitting errors and foregrounds increases, as one can see in the lower panel of Fig. 8. Increasing the number of frequency channels stored and analysed may be expensive, unfortunately. Since we can achieve low foreground contamination in our 0.5 MHz case, a further increase in frequency resolution may only significantly reduce the fitting contamination if a smaller frequency range is being observed and so a larger number of bins is required to avoid edge effects. Otherwise, a more stringent criterion for selecting the frequency resolution would be to choose it such that the decorrelation within a resolution element is not too large.

4.4 Alternative signal models and frequency ranges

So far we have shown results using only the f250C simulation of Iliev et al. (2008). We now show the effect on the signal extraction of taking our CS from the two simulations, T-QSO and T-star (see Section 2) described by Thomas et al. (2009). The top panel of Fig. 10 shows the variance of the CS derived from each of these three simulations as a function of redshift. This variance goes to zero at low redshift as reionization destroys the neutral hydrogen responsible for 21cm emission. The speed of this decline varies between simulations. The solid blue line (f250C) is most rapid, followed by the dashed red line (T-QSO) then the dot-dashed green line (T-star). This set of simulations is therefore useful to check that the quality of our fits is not unduly influenced by details of the

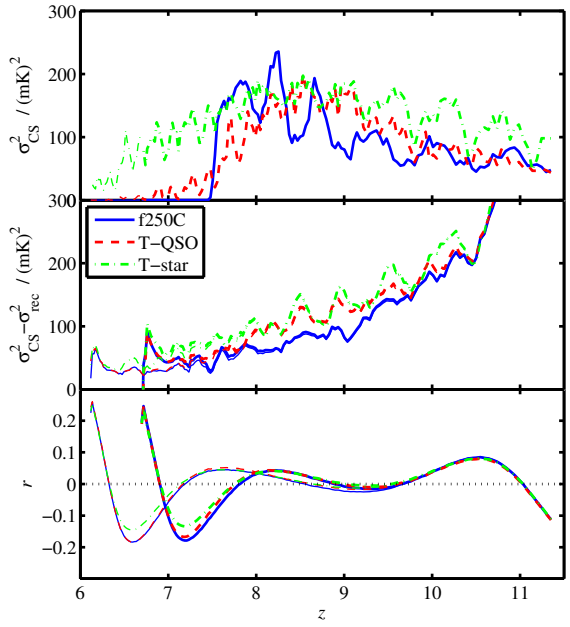

Figure 10. We study the effect on our extraction of using a different model for the CS and of truncating the frequency range used for the fit. Thin lines show results using our normal frequency range, $\nu = 115$–200 MHz, while thick lines show results using $\nu = 115$–185 MHz. In both cases the frequency resolution is 0.5 MHz. The top panel shows the variance in three different simulations of the CS as a function of redshift. The solid blue line uses the f250C simulation of Iliev et al. (2008) which we have been using throughout the paper, and is the same as the dotted line of Fig. 3. The red dashed line and the green dot-dashed line are for the simulations of Thomas et al. (2009) which assume that reionization is carried out by QSOs and by stars, respectively. The middle panel uses the same colour coding, and shows the difference between the recovered variance and the true variance of the CS. The bottom panel shows the Pearson correlation coefficient between the fitting errors and the foregrounds.

signal. We also use all three simulations to test our procedure over a shorter frequency interval, from 115 to 185 MHz ($z = 11.35$–6.70) rather than our fiducial 115–200 MHz ($z = 11.35$–6.12). We aim to check whether the different low redshift (high observing frequency) behaviour leads to this truncation having a different effect. This is an important test because it may not be possible to observe over the entire frequency range at once with LOFAR. Rather, we may have to split the frequency range into 32 MHz chunks which are observed consecutively. It may then be necessary to choose between increasing observing time at, say, 115–180 MHz and increasing the frequency range to 115–210 MHz.

The middle and bottom panels of Fig. 10 test the quality of the fit. The colour coding is the same as for the top panel. The thick lines show the results for an analysis using only 115–185 MHz while the thin solid lines show our fiducial 115–200 MHz case. In the middle panel we show the difference between the recovered variance of the CS and the original variance from the datacube without noise or foregrounds. For the thin solid line (f250C, 115–200 MHz), this is equal to the difference between the dotted line and the solid blue line in the top panel of Fig. 8; that is, it shows the amount of variance lost through overfitting. We see that the three different simulations show similar behaviour, though our procedure performs slightly better for f250C than for the other two simulations. For the majority of the frequency range, the thick and thin lines are indistinguishable, meaning that the effects of truncating the frequency range seem to be limited to the edge regions in this case. In the bottom panel we plot the Pearson correlation coefficient between the fitting errors and the foregrounds. The results using the three simulations are again very similar. The effect of the truncation is visible for a larger part of the range than was the case in the middle panel, but the correlations do not become larger: rather, the whole pattern just appears to be squashed.

We find that for larger values of λ (heavier smoothing) the effect of truncation on the correlation extends over a larger part of the frequency range. Recall that larger values of λ give a more aggressive signal recovery strategy, and one which allows us to detect an excess from the CS to higher redshift (that is, the lines of Fig. 3 do not fall away as rapidly at high redshift if λ is large). If we wish to pursue such an aggressive strategy, or indeed if it turns out to be necessary to do so to detect the signal, then extending to higher frequencies may turn out to be required: heavier smoothing makes more use of the longer lever arm provided by extending the range of the fit.

For our fiducial value of λ, though, we infer from Fig. 10 that if Wp smoothing is used to fit the foregrounds, shortening the frequency interval should not affect the quality of signal recovery in the interior too badly, either for extended or rapid reionization. The most important consideration when choosing what range of frequencies to observe is that we should be prepared to discard (or view with considerable caution) some bins at either end of the frequency range after fitting, since they are likely to be corrupted by edge effects. We should like to avoid discarding bins which are likely to have an interesting contribution from the CS: moving from an upper frequency limit of 180 MHz to one of 210 MHz could well be advantageous in this respect, though to some extent this will depend on the properties of the signal we aim to find.

5 CONCLUSIONS

We have argued that without a good reason to assume that the foregrounds for EoR experiments have a specific functional form, it is preferable to fit them with non-parametric methods that use their assumed smoothness directly, rather than to fit parameters of some chosen model. Unfortunately, most non-parametric methods tend to give poor quality fits compared to parametric ones that use the 'correct' model. We suggest that Wp smoothing may be an exception to this rule in the case examined in this paper.

Wp smoothing penalizes changes in curvature. In the general case it does so primarily by penalizing the existence of inflection points, but in the case that the inflection points are known or fixed, it penalizes the integrated change of curvature 'apart from inflection points'. We have drawn attention to the results of Mächler (1993, 1995), who derives a boundary value problem the solution of which is the desired smoothing function. We have sketched two algorithms which suffice to solve this problem in the case of EoR foregrounds, which we assume have no inflection points (as would be the case for a sum of several power law spectra with negative index). Our preferred algorithm is detailed in Section 3.2, while the other is outlined in Appendix A.

We have tested Wp smoothing on synthetic data cubes which include contributions from a detailed simulation of the CS from the EoR, a realistic model of the diffuse foregrounds, and the levels of noise and instrumental corruption expected for the LOFAR EoR experiment. Though Wp smoothing is considered to be non-parametric, it does require the specification of a smoothing pa-

rameter which governs the relative importance of the sum of the squared residuals and the curvature penalty function in the fitting. For the purposes of most of our tests we have adopted a value for λ which, for our dataset, provides a good compromise between over-fitting, which causes an underestimate of the variance of the CS, and under-fitting, which causes positive or negative correlations between the fitting errors and the foregrounds. Using this value of 0.5 for λ, we found that Wp smoothing easily outperforms other non-parametric methods we have tried, including the smoothing splines shown in Section 4.2[7], and is competitive with parametric fitting even when we are able to choose a parametrized functional form with advance knowledge of the foregrounds.

No scheme seems able to prevent the quality of the fit from degrading at the ends of the frequency interval used for observation. This problem can be mitigated somewhat by analysing data cubes with a high frequency resolution, though we note that high resolution is already desirable to avoid averaging away our signal, and this may be a more important criterion when deciding what resolution to use. We can make the quality of the fit marginally more uniform by increasing the weight given to data points near the ends of the frequency range. We argue, though, that the cost of doing so (in terms of increasing the noise on the fit) is too heavy for it to be worthwhile.

It may therefore be helpful to extend the range of frequencies observed. It is difficult to extend to lower frequencies (higher redshifts) because of the presence of the FM band. The increasing foreground and noise amplitude may also limit the usefulness of low frequency observations, though it is plausible that observations with the LOFAR low band antennas (which can observe at 30–80 MHz) could help constrain the shape of the foregrounds. Extending to higher frequencies is more promising. Firstly, the foregrounds and noise are smaller in amplitude. Secondly, because higher frequencies correspond to $z < 6$ we expect a negligible contribution from redshifted 21cm emission there. This helps to establish a baseline against which we can detect a higher redshift excess coming from the CS, and ensures that this excess occurs well away from the problematic edges of the frequency range. We have tested the quality of our fitting using two alternative simulations of the CS which exhibit more extended reionization, and have analysed all three simulations using a datacube which extends only to 185 MHz rather than 200 MHz. We find that away from the edges, neither change badly affects the quality of the foreground fitting.

We also note that we have concentrated primarily on the recovery of the excess variance coming from the CS as a measure of the quality of our fits. Other statistics such as the skewness may be more robust (Harker et al. 2009). It is also the case that the power from fitting errors, noise and the CS peaks at different scales, so a power spectrum analysis may improve prospects for detection of a signal, as well as giving more sensitive constraints on models than the integrated variance once a detection is made (e.g. Morales, Bowman & Hewitt 2006; Bowman, Morales & Hewitt 2007). Given this scale dependence, it is interesting to consider whether or not it may be advantageous to fit out the foregrounds in the *uv* plane. This does bring complications, for example that we must fit a complex function of frequency at each point in the *uv* plane, as opposed to a real function at each point in the image plane. It is possible, though, that by adapting the fitting according to the relative strength of the foregrounds, noise and signal at different scales we can improve sensitivity. We defer detailed study of power spectrum estimation and *uv* plane effects to future work.

Our results suggest that by paying close attention to the method used in fitting the foregrounds for EoR experiments, the sensitivity of these experiments can be increased, and we may have greater confidence that a detection of the signal is not affected too severely by foreground contamination. Foreground subtraction is very unlikely to be a bottleneck in the data processing and analysis pipeline, and so it is reasonable to consider relatively sophisticated and computationally expensive fitting methods if they provide a benefit. We have argued that Wp smoothing does seem to provide such a benefit, and will continue to test its performance as more elaborate models of the foregrounds and the instrument become available.

[7] Local regression and wavelet denoising, for example, overfit so severely that it becomes difficult to show them conveniently in the same figures.

ACKNOWLEDGMENTS

GH is supported by a grant from the Netherlands Organisation for Scientific Research (NWO). As LOFAR members, the authors are partially funded by the European Union, European Regional Development Fund, and by 'Samenwerkingsverband Noord-Nederland', EZ/KOMPAS.

REFERENCES

Ascher U., Russell R. D., 1981, SIAM Review, 23, 238
Bowman J. D., Morales M. F., Hewitt J. N., 2007, ApJ, 661, 1
Di Matteo T., Perna R., Abel T., Rees M. J., 2002, ApJ, 564, 576
Harker G. J. A. et al., 2009, MNRAS, 393, 1449
Iliev I. T., Mellema G., Pen U.-L., Bond J. R., Shapiro P. R., 2008, MNRAS, 384, 863
Jelić V. et al., 2008, MNRAS, 389, 1319
Kierzenka J., Shampine L. F., 2001, ACM Trans. Math. Softw., 27, 299
Labropoulos P. et al., 2009, MNRAS, submitted (arXiv:0901.3359)
Mächler M., 1989, PhD thesis, ETH Zürich
Mächler M., 1993, Research report 71, Very smooth nonparametric curve estimation by penalizing change of curvature. Seminar für Statistik ETH Zürich
Mächler M., 1995, Annals of Statistics, 23, 1496
Mellema G., Iliev I. T., Pen U.-L., Shapiro P. R., 2006, MNRAS, 372, 679
Morales M. F., Bowman J. D., Hewitt J. N., 2006, ApJ, 648, 767
Oh S. P., Mack K. J., 2003, MNRAS, 346, 871
Shaver P. A., Windhorst R. A., Madau P., de Bruyn A. G., 1999, A&A, 345, 380
Thomas R. M., Zaroubi S., 2008, MNRAS, 384, 1080
Thomas R. M. et al., 2009, MNRAS, 393, 32
Zaldarriaga M., Furlanetto S. R., Hernquist L., 2004, ApJ, 608, 622

APPENDIX A: ALTERNATIVE SOLUTION METHODS

It is possible to rewrite Equations 2, 7, 9 and 10 in a convenient form to solve them using a standard BVP solver. We have implemented Wp smoothing in this manner to test our finite difference scheme, and present the equations in appropriate form here for completeness.

At first sight the boundary conditions of Equation 10 look awkward, since they use the value of the function at points which are not at the ends of the interval. Solvers for such 'multi-boundary' problems are available, however. Moreover, by reexpressing the sums as integrals, we can take care of the boundary conditions by adding two more differential equations to the system, in line with the elegant trick suggested in section 5 of Ascher & Russell (1981).

This is promising, but doesn't help with the dependence on $f(x_i)$ for all i on the right-hand side of Equation 7, so we use a different trick. We start by rewriting Equations 2 and 7 as coupled first-order equations, as is commonly done:

$$h'(x) = g(x)\,; \tag{A1}$$

$$g'(x) = p_{\mathbf{w}}(x)\mathrm{e}^{h(x)}\left[-\frac{1}{2\lambda}\sum_{i=1}^{n}(x-x_i)_+\psi_i(y_i - f(x_i))\right]\,; \tag{A2}$$

$$f'(x) = k(x)\,; \tag{A3}$$

$$k'(x) = p_{\mathbf{w}}(x)\mathrm{e}^{h(x)}\,. \tag{A4}$$

Equations A1 and A3 define our new functions g and k respectively, and the boundary condition of (9) becomes $g(x_1) = g(x_n) = 0$.

Now, again following Ascher & Russell (1981), we split the domain of solution into $n-1$ intervals, $[x_1, x_2]$, $[x_2, x_3]$, …, $[x_{n-1}, x_n]$. In each interval we change variables, letting

$$t = \frac{x - x_m}{x_{m+1} - x_m} \quad \text{for} \quad x_m \leqslant x \leqslant x_{m+1} \tag{A5}$$

which maps each interval onto the unit interval, $[0, 1]$. Then, on this interval, we define functions $f_m(t)$, $g_m(t)$, $h_m(t)$, $k_m(t)$, $p_{\mathbf{w},m}(t)$ for $m = 1, 2, \ldots, n-1$ such that, for $x_m \leqslant x \leqslant x_{m+1}$, $f_m(t) = f(x)$, $g_m(t) = g(x)$, $h_m(t) = h(x)$, $k_m(t) = k(x)$ and $p_{\mathbf{w},m}(t) = p_{\mathbf{w}}(x)$. We further define the functions $q_m(t)$ for $m = 1, \ldots, n$ where $q_m(t) = f_m(0)$ for $m = 1, \ldots, n-1$ and $q_n(t) = f_{n-1}(1)$. Our system of four equations (A1–A4) then becomes the following system of $5n-4$ equations (where dashes now indicate differentiation with respect to t):

$$f'_m(t) = (x_{m+1} - x_m)k_m(t) \tag{A6}$$

$$k'_m(t) = (x_{m+1} - x_m)p_{\mathbf{w},m}(t)\mathrm{e}^{h_m(t)} \tag{A7}$$

$$h'_m(t) = (x_{m+1} - x_m)g_m(t) \tag{A8}$$

$$g'_m(t) = (x_{m+1} - x_m)p_{\mathbf{w},m}(t)\mathrm{e}^{h_m(t)} \times \left\{\frac{-1}{2\lambda}\sum_{i=1}^{m}[x_m + (x_{m+1} - x_m)t]\psi_i(y_i - q_i(t))\right\} \tag{A9}$$

$$q'_j(t) = 0 \tag{A10}$$

where the index m runs from 1 to $n-1$ and j runs from 1 to n. The functions q_j carry the value of f at the data points, $f(x_i)$, to the interior of the intervals, a property which is imposed with the boundary conditions

$$q_m(0) = f_m(0) \quad \text{for} \quad m = 1, \ldots, n-1; \tag{A11}$$

$$q_n(0) = f_{n-1}(1)\,. \tag{A12}$$

Our original boundary conditions become

$$g_1(0) = g_{n-1}(1) = 0\,; \tag{A13}$$

$$\sum_{i=1}^{n}\psi_i(y_i - q_i(0)) = \sum_{i=1}^{n}x_i\psi_i(y_i - q_i(0)) = 0\,. \tag{A14}$$

The remaining $4(n-2)$ boundary conditions come from imposing continuity on the functions $f(x)$, $g(x)$, $h(x)$ and $k(x)$:

$$f_m(1) = f_{m+1}(0)\,; \tag{A15}$$

$$g_m(1) = g_{m+1}(0)\,; \tag{A16}$$

$$h_m(1) = h_{m+1}(0)\,; \tag{A17}$$

$$k_m(1) = k_{m+1}(0)\,; \tag{A18}$$

where here the index m runs from 1 to $n-2$.

Note that the boundary conditions only involve the value of functions at $t = 0$ and $t = 1$, and that to calculate the derivatives given by Eqns. A6–A10 at a given value of t only requires the evaluation of functions at the same value of t. The system is therefore suitable for solution using the MATLAB routine 'bvp4c' (Kierzenka & Shampine 2001), a BVP solver that uses a collocation method. We call it with an initial mesh of five evenly spaced points, and with initial conditions calculated in a similar fashion to those used for the finite difference scheme in the main text. The system of equations is greatly expanded from the four with which we started since the special form of the problem is not exploited, and typically takes several seconds to solve on our test machines.