

# ROHSA: Regularized Optimization for Hyper-Spectral Analysis<sup>★</sup>

## Application to phase separation of 21 cm data

Antoine Marchal<sup>1,2</sup>, Marc-Antoine Miville-Deschênes<sup>1</sup>, François Orieux<sup>2,3</sup>, Nicolas Gac<sup>3</sup>, Charles Soussen<sup>3</sup>, Marie-Jeanne Lesot<sup>4</sup>, Adrien Revault d'Allonnes<sup>4</sup>, and Quentin Salomé<sup>5</sup>

<sup>1</sup> AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, 91191 Gif-sur-Yvette, France

<sup>2</sup> Institut d'Astrophysique Spatiale, CNRS UMR 8617, Université Paris-Sud 11, Batiment 121, 91405 Orsay, France  
e-mail: antoine.marchal@ias.u-psud.fr

<sup>3</sup> Laboratoire des Signaux et Systèmes (CNRS, CentraleSupélec, University of Paris-Sud), Université Paris-Saclay, 91192 Gif-sur-Yvette, France

<sup>4</sup> LIP6, Université Pierre et Marie Curie-Paris 6, UMR7606, 4 Place Jussieu Paris cedex 05, 75252 Paris, France

<sup>5</sup> Instituto de Radioastronomía y Astrofísica, Universidad Nacional Autónoma de México, 58089 Morelia, Mexico

Received 21 February 2019 / Accepted 29 April 2019

### ABSTRACT

*Context.* Extracting the multiphase structure of the neutral interstellar medium is key to understanding star formation in galaxies. The radiative condensation of the diffuse warm neutral medium producing a thermally unstable lukewarm medium and a dense cold medium is closely related to the initial step leading the atomic-to-molecular (HI-to-H<sub>2</sub>) transition and the formation of molecular clouds. Up to now, the mapping of these phases out of 21 cm emission hyper-spectral cubes has remained elusive mostly due to the velocity blending of individual cold structures present on a given line of sight. As a result, most of the current knowledge about the HI phases rests on a small number of absorption measurements on lines of sight crossing radio sources.

*Aims.* The goal of this work is to develop a new algorithm to perform separation of diffuse sources in hyper-spectral data. Specifically the algorithm was designed in order to address the velocity blending problem by taking advantage of the spatial coherence of the individual sources. The main scientific driver of this effort was to extract the multiphase structure of the HI from 21 cm line emission only, providing a means to map each phase separately, but the algorithm developed here should be generic enough to extract diffuse structures in any hyper-spectral cube.

*Methods.* We developed a new Gaussian decomposition algorithm named ROHSA based on a multi-resolution process from coarse to fine grid. ROHSA uses a regularized nonlinear least-square criterion to take into account the spatial coherence of the emission and the multiphase nature of the gas simultaneously. In order to obtain a solution with spatially smooth parameters, the optimization is performed on the whole data cube at once. The performances of ROHSA were tested on a synthetic observation computed from numerical simulations of thermally bi-stable turbulence. We apply ROHSA to a 21 cm observation of a region of high Galactic latitude from the GHIGLS survey and present our findings.

*Results.* The evaluation of ROHSA on synthetic 21 cm observations shows that it is able to recover the multiphase nature of the HI. For each phase, the power spectra of the column density and centroid velocity are well recovered. More generally, this test reveals that a Gaussian decomposition of HI emission is able to recover physically meaningful information about the underlying three-dimensional fields (density, velocity, and temperature). The application on a real 21 cm observation of a field of high Galactic latitude produces a picture of the multiphase HI, with isolated, filamentary, and narrow ( $\sigma \sim 1\text{--}2\text{ km s}^{-1}$ ) structures, and broader ( $\sigma \sim 4\text{--}10\text{ km s}^{-1}$ ), diffuse, and space-filling components. The test-case field used here contains significant intermediate-velocity clouds that were well mapped out by the algorithm. As ROHSA is designed to extract spatially coherent components, it performs well at projecting out the noise.

*Conclusions.* In this paper we introduce ROHSA, a new algorithm that performs a separation of diffuse sources in hyper-spectral data on the basis of a Gaussian decomposition. The algorithm makes no assumption about the nature of the sources, except that each one has a similar line width. The tests we made shows that ROHSA is well suited to decomposing complex 21 cm line emission of regions of high Galactic latitude, but its design is general enough that it could be applied to any hyper-spectral data type for which a Gaussian model is relevant.

**Key words.** ISM: clouds – ISM: kinematics and dynamics – ISM: structure – methods: data analysis – methods: numerical – methods: observational

## 1. Introduction

Star formation in galaxies is strongly linked to the physical processes that govern the evolution of the interstellar medium (ISM). Stars form by gravitational collapse of dense ( $n > 10^4\text{ cm}^{-3}$ ) and cold ( $T \sim 10\text{ K}$ ) structures in molecular

clouds but the process that leads to the formation of these over-densities is still unclear. One key element seems to be related to the efficiency of the formation of cold clouds of neutral hydrogen (HI; [Ostriker et al. 2010](#)).

The current vision of the HI comes from an important legacy. Early observations of the 21 cm line showed a significant difference between emission and absorption spectra. On lines of sight crossing radio-sources, the HI appears in absorption with

<sup>★</sup> ROHSA is available in free access via the following web page: <https://github.com/antoinemarchal/ROHSA>

very narrow features (a few  $\text{km s}^{-1}$ ). In emission the 21 cm line contains these narrow features on top of much broader spectral structures ( $10\text{--}20 \text{ km s}^{-1}$ ). Clark (1965) was the first to suggest that this might be the signature of a cloud-intercloud medium in pressure equilibrium. Very rapidly Field (1965) and Field et al. (1969) introduced the concept of thermal instability and laid out the theoretical ground of a “two phase” HI model showing that, at the pressure of the ISM, the heating and cooling processes naturally lead to two thermally stable states: a dense cold neutral medium (CNM –  $T \sim 50 \text{ K}$ ,  $n \sim 50 \text{ cm}^{-3}$ ) immersed in a diffuse warm neutral medium (WNM –  $T \sim 8000 \text{ K}$ ,  $n \sim 0.3 \text{ cm}^{-3}$ ). This vision was later complemented by Wolfire et al. (1995, 2003) considering updated heating (dominated by the photoelectric effect on small dust grains) and cooling (dominated by CII:  $158 \mu\text{m}$ , OI:  $63 \mu\text{m}$ ,  $L\alpha$  and electron recombinations onto positive charged grains) processes of the ISM.

This description of the diffuse neutral gas complemented a parallel hypothesis that emerged in the 1950s (e.g., von Weizsäcker 1951; von Hoerner 1951; Chandrasekhar & Münch 1952) and considered the ISM as a multi-scale turbulent medium. In this case, the density and velocity structures are the result of a highly dynamical and out-of-equilibrium medium. In order to reconcile the “static/two-phase” and the “turbulent” hypotheses, several studies have aimed at understanding the production of the CNM in a turbulent and thermally unstable flow using numerical simulations (e.g., Hennebelle & Péroul 1999; Koyama & Inutsuka 2002; Audit & Hennebelle 2005; Hennebelle et al. 2008; Saury et al. 2014). In general, these numerical studies show that the WNM has the properties of a trans-sonic turbulent flow, while the CNM shows a much more contrasted density structure, in accordance with the cloud-intercloud picture. In addition, such studies indicate the presence of a significant fraction of the mass being in the thermally unstable regime (i.e., with a temperature mid-way between the CNM and WNM stable states). For instance, Saury et al. (2014) showed that 30% of the HI is in the thermally unstable regime. Interestingly, these latter authors also show that this lukewarm neutral medium (LNM) is spatially located around the cold structures, pointing at the transitional nature of this thermal state.

From an observational standpoint, studies combining 21 cm absorption and emission data have clearly revealed the presence of HI at intermediate/unstable temperatures, typically between 500 and 5000 K (e.g., Heiles & Troland 2003a; Kanekar et al. 2003; Roy et al. 2013a,b; Murray et al. 2015, 2018a). Based on a coherent modeling of emission and absorption spectra, Heiles & Troland (2003b), Murray et al. (2015, 2018a) estimated that about 30% of the HI is in the cold CNM phase, 20% in the thermally unstable regime, and 50% in the WNM. Nevertheless the fraction of the HI in each phase remains uncertain and large variations are observed: the fraction of the mass in the CNM ranges from  $\sim 1\%$  to more than 50% (Murray et al. 2018a).

The nature of these variations and how they relate to the dynamical conditions of the gas remains largely unexplored from the observational point of view. One main hurdle in getting access to this information is the fact that our knowledge of the multiphase nature of the HI relies on 21 cm absorption measurements that are limited to lines of sight crossing radio sources. By nature, this way of observing prevents us from mapping the HI phases. To go further, and really compare with numerical simulations that are currently under-constrained by observation, it is mandatory to map the column density structure of each phase and study the spatial variations of their centroid velocity and velocity dispersion. This calls for methods that can extract the information of each HI phase from fully sampled 21 cm emission

data only. Huge efforts have been made to map the 21 cm emission of the Galactic HI (recent examples are Taylor et al. 2003; Kalberla et al. 2005; Stil et al. 2006; McClure-Griffiths et al. 2009; Winkel et al. 2016; Peek et al. 2018) and a large amount of data is now available. The information about the multiphase and multi-scale nature of the HI contained in these large hyper-spectral data cubes has remained elusive due to the difficulty in separating the emission from the different phases on each line of sight. In this paper we propose a new method to map out the contribution of each phase to the 21 cm emission. The method is based on a decomposition of the 21 cm emission line with Gaussian profiles and constraints that favors spatially coherent parameters.

The paper is organized as follows. In Sect. 2, we describe the methodology used to develop our Gaussian decomposition algorithm. An evaluation of numerical simulation is presented in Sect. 3. In Sect. 4 we present an application to observations. The discussion and summary are presented Sects. 5 and 6.

## 2. Methodology

### 2.1. Gaussian decomposition of the 21 cm emission

Very early on after its detection, the 21 cm line was observed to be well described by a sum of a small number of Gaussian components. This was found to be true for the least confused absorption spectra (Muller 1957, 1959; Clark 1965) but also for emission spectra observed away from the Galactic plane (Heeschen 1955). Very few spectra at high Galactic latitudes do not comply with that rule, whatever the angular resolution of the data. Recently, Kalberla & Haud (2018) showed that more than 60% of the spectra over the whole sky can be described by the sum of seven Gaussian components or less. In the two decades after the detection of the 21 cm line, many studies used the Gaussian decomposition to infer physical parameters from the data (Matthews 1957; Davis 1957; Muller 1957, 1959; Dieter 1964, 1965; Lindblad 1966; Takakubo & van Woerden 1966; Mebold 1972). The fact that a small number of Gaussian components is needed to describe the signal was seen as a convenient way to describe the emission profiles with a small set of parameters (Takakubo & van Woerden 1966). It is also a very strong element in favor of the Gaussian function as a significant descriptor of the underlying physics.

Takakubo (1967) showed that the width of the 21 cm emission line could be grouped in three components ( $\sigma \leq 3 \text{ km s}^{-1}$ ;  $3 < \sigma < 7 \text{ km s}^{-1}$ ;  $\sigma \geq 7 \text{ km s}^{-1}$ ), a result confirmed later on (Mebold 1972; Haud & Kalberla 2007; Kalberla & Haud 2018). Takakubo (1967) also showed that the narrow 21 cm features are well correlated with Ca+ K line absorption measurements. These latter authors concluded that the narrow component, also seen in 21 cm absorption spectra, is likely to be isolated cold clouds (CNM) in the Solar neighborhood. They also showed that the spatial distribution of the centroid velocity, velocity width, and column density of the large feature is compatible with a warm (WNM) and diffuse disk that follows Galactic rotation. The second group of Gaussians, with a width ( $\sigma$ ) of between 3 and  $7 \text{ km s}^{-1}$ , is generally attributed to gas in the thermally unstable range, but a fraction of them could be caused by blending of narrow features.

The exact mass fraction of gas in each phase (CNM, LNM, and WNM) is still a matter of debate. In addition, because this knowledge is based on absorption measurements, there is very little information about the structure of these phases on the sky. Being able to separate the different phases on each line of sight

would allow us to study the structure and kinematics of the cold phase and its relationship with the more diffuse gas. In theory, one could expect that Gaussian decomposition of the emission spectra could provide such mapping.

## 2.2. Limitation of the Gaussian model

There are many pitfalls in the description of the 21 cm emission spectra as a sum of Gaussian components: velocity blending, ambiguities of the number of components, nonGaussian profiles, noise peaks, plurality of solutions. Another important limitation is the effect of optical depth of the 21 cm line that modifies the shape of the line. More generally, the main opposition to Gaussian decomposition of emission spectra is that any spectrum can be decomposed that way provided that enough Gaussians are used. If that is the case, how can one be sure that the Gaussian representation provides some real physical information about the emitting gas? For instance, two spatially disconnected cold structures present on the same line of sight could appear at the same projected velocity. In this case their respective emission profiles would be confused. This velocity blending affects both the emission and absorption spectra; it is present in the data, and more so at lower Galactic latitudes where the length of the line of sight is larger and the number of HI structures increases.

This line of reasoning led [Dickey & Lockman \(1990\)](#) to advise against using Gaussian decomposition to analyze 21 cm spectra. Later on, it continued to be used ([Verschuur & Schmelz 1989](#); [Verschuur & Magnani 1994](#); [Poppel et al. 1994](#); [Haud 2000](#); [Verschuur 2004](#); [Begum et al. 2010](#); [Martin et al. 2015](#); [Kalberla & Haud 2018](#)) but overall there was a loss of interest, except for the analysis of absorption spectra. Indeed by comparing nearby absorption and emission spectra one can recover the effect of optical depth on the 21 cm emission. In addition, absorption measurements are only sensitive to cold gas, limiting the velocity blending problem. For these reasons, the Gaussian decomposition continued to be used in this context ([Dickey et al. 2003](#); [Kanekar et al. 2003](#)), and especially after the seminal work of [Heiles & Troland \(2003b\)](#) who developed a dedicated formalism which has been used in several other studies since (e.g. [Stanimirović & Heiles 2005](#); [Begum et al. 2010](#); [Stanimirović et al. 2014](#); [Lee et al. 2015](#); [Murray et al. 2014, 2015, 2018b](#)). Indeed, key information about the nature of the HI came from the joint Gaussian decomposition of emission and absorption spectra.

## 2.3. Development of a new approach

Following what has been done for the comparison of emission and absorption spectra where the Gaussian decomposition is considered valid, we would like to argue that a similar decomposition could be envisaged for emission data only, at least at high Galactic latitudes where the effect of optical depth of the 21 cm line has been shown to be negligible ([Murray et al. 2018b](#)).

The fact that physical information could be obtained using a Gaussian decomposition of absorption data reveals the fact that thermal broadening has a significant effect in shaping the line profile, or in other words, that the dynamics of each HI phase is typical of sub- or trans-sonic turbulence. When the amplitude of turbulent and thermal motions are commensurate, the line appears smooth and can be represented by a small number of Gaussian components ([Miville-Deschênes et al. 2003](#)). If the HI at high Galactic latitude is indeed represented by a two-phase medium with small, cold, and trans-sonic structures immersed in a relatively low-Mach-number and warm

diffuse phase, the Gaussian representation could bear significant physical information.

The perspective of mapping the phases of the HI is so important that we ventured to explore new ways of decomposing the emission spectra that could be applicable to the high Galactic sky. The main difficulty is the effect of velocity blending for cold structures. As mentioned by [Takakubo & van Woerden \(1966\)](#), there will always be cases where a given spectrum can be fitted with a smaller number of components than its neighbors, if two or more components have similar central velocity and velocity dispersion. One way to avoid this confusion is to look for solutions that have a spatial continuity, or that have a slow spatial variation. [Poppel et al. \(1994\)](#), [Haud \(2000\)](#), [Martin et al. \(2015\)](#) and [Miville-Deschênes et al. \(2017b\)](#) have implemented Gaussian decomposition methods that use some information about their neighbors in order to favor spatially coherent solutions. Formally nevertheless, these algorithms do not force solutions to be spatially coherent; they simply provide initial guesses to the fit of a single spectrum based on the most likely solutions found in some surrounding area. The optimization is not bound to this initial guess and it can always converge to another solution that would break the spatial smoothness of the parameter space.

The novelty of the algorithm we present here is that it is the first one that imposes the spatial coherence in the determination of the parameters. In order to do that, all the spectra of the emission cube are fitted at the same time. To make sure that the recovered parameters are spatially smooth, specific regularization terms are added to the cost function with non-negativity constraints on the amplitude. This algorithm, called ROHSA, is described below.

## 2.4. ROHSA

ROHSA performs a regression analysis using a regularized non-linear least-square criterion. We formulate in this section the Gaussian model used as well as the energy terms added to the cost function to take into account the spatial coherence of the emission and the multiphase nature of the gas simultaneously. The quasi-Newton algorithm, L-BFGS-B, used to perform the optimization is then briefly described. Finally, we formulate the algorithm performed by ROHSA based on a multi-resolution process from coarse to fine grid.

### 2.4.1. Model

The data are the measured brightness temperature  $T_B(v_z, \mathbf{r})$  at a given projected velocity  $v_z$  across sky coordinates  $\mathbf{r}$ . The proposed model  $\tilde{T}_B(v_z, \boldsymbol{\theta}(\mathbf{r}))$  is a sum of  $N$  Gaussian  $G(v_z, \boldsymbol{\theta}_n(\mathbf{r}))$

$$\tilde{T}_B(v_z, \boldsymbol{\theta}(\mathbf{r})) = \sum_{n=1}^N G(v_z, \boldsymbol{\theta}_n(\mathbf{r})), \quad (1)$$

with  $\boldsymbol{\theta}(\mathbf{r}) = (\boldsymbol{\theta}_1(\mathbf{r}), \dots, \boldsymbol{\theta}_n(\mathbf{r}))$  and where

$$G(v_z, \boldsymbol{\theta}_n(\mathbf{r})) = a_n(\mathbf{r}) \exp\left(-\frac{(v_z - \boldsymbol{\mu}_n(\mathbf{r}))^2}{2\sigma_n(\mathbf{r})^2}\right) \quad (2)$$

is parametrized by  $\boldsymbol{\theta}_n = (a_n, \boldsymbol{\mu}_n, \sigma_n)$  with  $a_n \geq 0$  being the amplitude,  $\boldsymbol{\mu}_n$  the position, and  $\sigma_n$  the standard deviation 2D maps of the  $n$ th Gaussian profile across the plan of sky. The residual is

$$L(v_z, \boldsymbol{\theta}(\mathbf{r})) = \tilde{T}_B(v_z, \boldsymbol{\theta}(\mathbf{r})) - T_B(v_z, \mathbf{r}). \quad (3)$$

The estimated parameters  $\hat{\theta}$  are defined as the minimizer of a cost function that includes the sum of the squares of the residual,

$$Q(\theta) = \frac{1}{2} \left\| L(v_z, \theta) \right\|_{\Sigma}^2 = \frac{1}{2} \sum_{v_z, \mathbf{r}} \left( \frac{L(v_z, \theta(\mathbf{r}))}{\Sigma(\mathbf{r})} \right)^2, \quad (4)$$

where  $\Sigma$  is the standard deviation 2D map of the noise assumed without spatial correlation. In practice this term is estimated using a sequence of empty velocity channels of  $T_B(v_z, \mathbf{r})$ .

For each of the  $N$  Gaussians, we want to obtain a spatially coherent solution, meaning that for each parameter, the values have to be close for neighboring lines of sight. This can be done by penalizing the small-scale spatial fluctuations of each parameter, measured by the energy at high spatial frequencies. The considered high-pass filter is the second-order differences, that is the Laplacian filtering, defined by the 2D convolution kernel,

$$d = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \quad (5)$$

The following regularization term, itself containing energy terms, is added to the cost function given in Eq. (4),

$$R(\theta) = \frac{1}{2} \sum_{n=1}^N \lambda_a \| \mathbf{D} \mathbf{a}_n \|_2^2 + \lambda_\mu \| \mathbf{D} \mu_n \|_2^2 + \lambda_\sigma \| \mathbf{D} \sigma_n \|_2^2, \quad (6)$$

where  $\mathbf{D}$  is a matrix performing the 2D convolution using the kernel  $d$  and  $\lambda_a$ ,  $\lambda_\mu$ , and  $\lambda_\sigma$  are hyper-parameters that tune the balance between the different terms.

These terms ensure a positive spatial correlation of the model parameters for neighboring pixels. However, each term is free to have large variation across the field at larger scale. Since  $\sigma_n$  contains information about the gas thermodynamics, we design an additional term in the cost function to group Gaussians with similar  $\sigma_n$ . This is implemented in order to favor any solution that would produce components ascribable to each of the phases (WNM, LNM or CNM). In order to do that we add another term,  $\lambda'_\sigma \| \sigma_n - m_n \|_2^2$ , which constrains  $\sigma_n$  to be close to an unknown scalar value  $m_n$ . The full regularization term is then

$$R(\theta, \mathbf{m}) = \frac{1}{2} \sum_{n=1}^N \lambda_a \| \mathbf{D} \mathbf{a}_n \|_2^2 + \lambda_\mu \| \mathbf{D} \mu_n \|_2^2 + \lambda_\sigma \| \mathbf{D} \sigma_n \|_2^2 + \lambda'_\sigma \| \sigma_n - m_n \|_2^2, \quad (7)$$

with  $\mathbf{m} = (m_1, \dots, m_N)$  and  $\mathbf{a}_n \geq 0, \forall n \in [1, \dots, N]$ . The last two terms in Eq. (7), representative of a joined constraint imposed on  $\sigma_n$ , allow us to interpret a posteriori and simultaneously the morphology and the thermodynamical state of each component extracted from the data. The full cost function is then

$$J(\theta, \mathbf{m}) = Q(\theta) + R(\theta, \mathbf{m}). \quad (8)$$

#### 2.4.2. Optimization algorithm

Unlike  $Q(\theta)$ , each energy term proposed in Eq. (6) involves linear dependences on the parameter  $\theta$ . The cost function in Eq. (8) is therefore a regularized nonlinear least-square criterion. The minimizer,

$$[\hat{\theta}, \hat{\mathbf{m}}] = \underset{\theta, \mathbf{m}}{\operatorname{argmin}} J(\theta, \mathbf{m}), \text{ wrt. } \mathbf{a}_n \geq 0, \forall n \in [1, \dots, N], \quad (9)$$

**Algorithm 1** ROHSA based on a multi-resolution process from coarse to fine grid where  $\langle T_B \rangle_i$  is the averaged data at  $i$  scale.

---

**Require:**  $T_B(v_z, \mathbf{r}), \theta^{(0)}, \mathbf{m}^{(0)} = \mathbf{0}, N, \lambda_a, \lambda_\mu, \lambda_\sigma, \lambda'_\sigma$   
1: **for**  $i = 1$  to  $I$  **do**  
2:      $[\theta^{(i)}, \mathbf{m}^{(i)}] \leftarrow \underset{\theta, \mathbf{m}}{\operatorname{argmin}} J(\theta^{(i-1)}, \mathbf{m}^{(i-1)}, \langle T_B \rangle_i)$ .  
3: **end for**  
**return**  $\theta^{(I)}, \mathbf{m}^{(I)}$

---

has no closed form expression and is not directly tractable because of the complexity of the model  $\tilde{T}_B$  and the size of the unknown and data. The proposed solution relies instead on an iterative optimization algorithm that uses the gradient

$$\nabla J(\theta, \mathbf{m}) = \begin{bmatrix} \nabla L(\theta) \times L(\theta) \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \nabla_\theta R(\theta, \mathbf{m}) \\ \nabla_{\mathbf{m}} R(\theta, \mathbf{m}) \end{bmatrix}, \quad (10)$$

which is tractable since it involves the residual, the Jacobian of the residual  $\nabla L(\theta)$ , and 2D convolutions with the kernel  $d$  for  $\mathbf{D}$  and  $\mathbf{D}^t$ . The gradient  $\nabla R^t(\theta, \mathbf{m}) = [\nabla_\theta R^t(\theta, \mathbf{m}), \nabla_{\mathbf{m}} R^t(\theta, \mathbf{m})]$  and  $\nabla J(\theta)$  are detailed in Appendix A.

For the optimization, ROHSA relies on L-BFGS-B (for Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds), a quasi-Newton iterative algorithm described by Zhu et al. (1997) which allows for the positivity constraints of the amplitudes to be taken into account. In this algorithm, after an initialization  $\theta_{(0)}$ , the solution is approached iteratively by

$$\theta_{(k+1)} = \theta_{(k)} - \alpha_{(k)} \mathbf{H}_{(k)}^{-1} \nabla J(\theta_{(k)}, \mathbf{m}_{(k)}), \quad (11)$$

where  $\mathbf{H}^{-1} \nabla J(\theta, \mathbf{m})$  is approximated with the L-BFGS formula. The iterations are repeated until one of the two following criteria is met: (1) the total number of evaluations of  $J(\theta, \mathbf{m})$  and  $\nabla J(\theta, \mathbf{m})$  exceeds a maximum number of iterations defined by the user; (2) the projected gradient is sufficiently small (i.e.,  $|\operatorname{proj} \nabla J(\theta, \mathbf{m})| / (1 + |J(\theta, \mathbf{m})|) < 10^{-10}$ ).

Due to its nonlinearity, the least-square criterion  $J(\theta, \mathbf{m})$  described in Eq. (8) is likely to include local minimizers. Therefore, the L-BFGS-B algorithm used by ROHSA is also likely to converge toward one of these local minima, making the solution highly dependent on the initialization  $\theta_{(0)}^{(I)}$ . In order to overcome this difficulty, we based the design of ROHSA on an iterative multi-resolution process from coarse to fine grid (described in the following section) to automatically choose  $\theta_{(0)}^{(I)}$  and to converge towards a satisfactory local minimum.

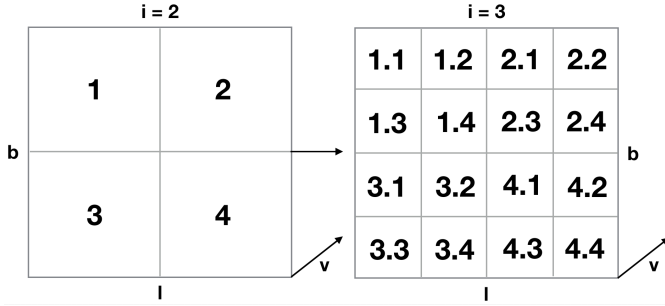
#### 2.4.3. ROHSA algorithm

ROHSA is based on an iterative algorithm using a multi-resolution process from coarse to fine grid presented in Algorithm 1. The number of iterations  $I$  depends on the size of the fine grid  $S$  and is defined by the relation  $2^I = S$ . For example, a grid of size  $S^2 = 256^2$  requires  $I = 8$  iterations. Each iteration is made of three steps.

1. Data are averaged at scale  $i$  as

$$\langle T_B \rangle_i = \frac{1}{K_i} \sum_{k \in \mathcal{V}_i} T_B(v_z, \mathbf{r}_k), \quad (12)$$

where  $\mathcal{V}_i$  defines the neighborhood at scale  $i$ , as described in Fig. 1, and  $K_i$  is the number of positions in that neighborhood. For  $i = 1$ , all the spatial information is compressed into a single spectrum:  $\langle T_B \rangle_1 = \langle T_B(v_z, \mathbf{r}) \rangle$ .



**Fig. 1.** Graphic visualization of neighborhoods  $\mathcal{V}_2$  and  $\mathcal{V}_3$  used to obtain the spatially averaged data versions  $\langle T_B \rangle_2$  and  $\langle T_B \rangle_3$ .

2. The parameters  $\theta^{(i)}$  and  $m^{(i)}$  are estimated on that spatially averaged data version  $\langle T_B \rangle_i$  by minimizing the cost function given in Eq. (8).

The minimization (line 2 of Algorithm 1) is made using L-BFGS, described in the previous section. We note that for scale  $i = 1$ , there is no spatial information  $\langle T_B \rangle_1$  and the result does not depend on the regularization.

3. Parameters  $\theta^{(i)}$  are spatially interpolated at nearest neighborhood to serve as initialization for the next scale  $i + 1$ .

The free hyper-parameters  $\lambda_A$ ,  $\lambda_\mu$ ,  $\lambda_\sigma$ ,  $\lambda'_\sigma$  remain constant during the iterations.

### 3. Evaluation on numerical simulation

To evaluate the performance of ROHSA, we applied it to synthetic 21 cm observations computed from a numerical simulation of thermally bi-stable turbulence flow. This allowed us to directly compare the solution given by ROHSA to the properties of the gas present in the simulation. That direct comparison with numerical reality is an essential test to evaluate the performances of a source-separation algorithm like ROHSA.

#### 3.1. Numerical simulation

To test ROHSA we used the hydrodynamical simulation of thermally bi-stable turbulence performed by Saury et al. (2014). We used their 1024N01 simulation ( $1024^3$  pixels and a physical size of the box of 40 pc) characterized by (1) an initial density  $n_0 = 0.1 \text{ cm}^{-3}$ , (2) a large-scale velocity  $v_S = 12.5 \text{ km s}^{-1}$  and (3) a spectral weight  $\zeta = 0.2$ . The initial density corresponds to the typical density of the WNM before condensation, the large-scale velocity represents the amplitude given to the field that generates large-scale turbulent motions in the box, and finally the spectral weight controls the modes of the turbulent mixing (here a majority of compressible modes). The Mach number of this simulation has been evaluated to be around  $M = 0.85$  for  $T > 200 \text{ K}$ .

In order to explore the performances of ROHSA we use only a subset of this simulation. We concentrate our analysis on a region of  $256 \times 256 \times 1024$  pixels with a moderate CNM fraction in order to limit the effect of HI self-absorption (see Sect. 3.2.3).

#### 3.2. 21 cm line synthetic observations

The synthetic 21 cm observations were computed using the formalism described by Miville-Deschênes & Martin (2007).

##### 3.2.1. Distribution of velocity fluctuations

In the 3D spatial space, the neutral hydrogen can be described by three 3D fields: the temperature  $T(\mathbf{x})$ , the density  $\rho(\mathbf{x})$ , and the

$z$ -component of the turbulent velocity field  $v_z(\mathbf{x})$ . Here the 3D spatial positions are denoted by the vector  $\mathbf{x}$  while the 2D vector expressing the line-of-sight is denoted by  $\mathbf{r}$ . The  $z$ -axis is taken along the line of sight.

Information about the velocity field is inevitably lost because of the projection along  $z$ -axis. This makes this description of HI a nonexhaustive one. For each position  $\mathbf{x}$ , we assume that the velocity dispersion of a given cell is dominated by thermal motions. This is a fair approximation as the turbulent velocity dispersion at the cell size ( $0.04 \text{ pc}$ ) is  $\sigma_{\text{turb}} \sim 0.3 \text{ km s}^{-1}$ , which is smaller than the thermal broadening everywhere in the simulation: the smallest thermal broadening for the coldest gas found in the simulation ( $T = 20 \text{ K}$ ) is  $\sigma_{\text{therm}} = 0.4 \text{ km s}^{-1}$ . The distribution function of the  $z$ -component of the velocity  $v_z(\mathbf{x})$  of a given cell is then given by  $\phi_{v_z}(\mathbf{x})$ , a Maxwellian shifted by  $v_z(\mathbf{x})$ ,

$$\phi_{v_z}(\mathbf{x}) dv'_z = \frac{1}{\sqrt{2\pi}\Delta(\mathbf{x})} \times \exp\left(-\frac{(v'_z - v_z(\mathbf{x}))^2}{2\Delta^2(\mathbf{x})}\right) dv'_z, \quad (13)$$

where  $\Delta(\mathbf{x}) = \sqrt{k_B T(\mathbf{x})/m_H}$ , which is the thermal broadening of the 21 cm line,  $m_H$  is the hydrogen atom mass, and  $k_B$  the Boltzmann constant.

##### 3.2.2. Brightness temperature: general case

The general case for the computation of the 21 cm brightness temperature  $T_B(v_z, \mathbf{r})$  is based on the following radiative transfer equation.

$$T_B(v_z, \mathbf{r}) = \sum_z T(\mathbf{r}, z) \left[ 1 - e^{-\tau(v_z, \mathbf{r}, z)} \right] e^{-\sum_{z' < z} \tau(v_z, \mathbf{r}, z')}, \quad (14)$$

where  $\tau(v_z, \mathbf{r}, z)$  is the optical depth of the 21 cm line defined as

$$\tau(v_z, \mathbf{r}, z) = \frac{1}{C} \frac{\rho(\mathbf{r}, z) \phi_{v_z}(\mathbf{r}, z)}{T(\mathbf{r}, z)} dz, \quad (15)$$

and  $C = 1.82243 \times 10^{18} \text{ cm}^{-2} (\text{K km s}^{-1})^{-1}$ . In this representation, a gas cell at position  $z'$  absorbs emission from the cell located behind it along the line of sight, i.e., at  $z > z'$ .

##### 3.2.3. Optically thin limit

In the optically thin limit, in cases where the self-absorption is negligible (i.e.,  $\tau(v_z, \mathbf{r}, z) \ll 1$  everywhere), the 21 cm brightness temperature is proportional to the density  $\rho$ :

$$T_B^{\text{thin}}(v_z, \mathbf{r}) dv'_z = B(\mathbf{r}) \otimes \frac{1}{C} \int_0^H dz \rho(\mathbf{r}, z) \phi_{v_z}(\mathbf{r}, z) dv'_z, \quad (16)$$

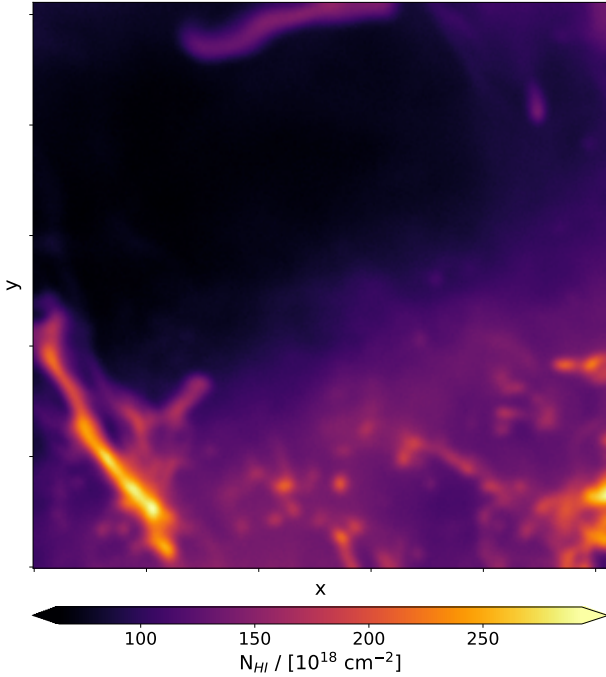
where  $H$  is the depth of the cloud and  $\otimes$  the spatial convolution. Here we consider the case that includes spatial smoothing by a telescope beam  $B(\mathbf{r})$ .

The integrated column density  $N_{\text{HI}}^{\text{thin}}(\mathbf{r})$  and the centroid velocity  $\langle v_z(\mathbf{r}) \rangle$  of the 21 cm line can be obtained directly by integrating  $T_B^{\text{thin}}(v_z, \mathbf{r})$  along the velocity axis:

$$N_{\text{HI}}^{\text{thin}}(\mathbf{r}) = C \int_{-\infty}^{+\infty} T_B(v_z, \mathbf{r}) dv_z, \quad (17)$$

and

$$\langle v_z(\mathbf{r}) \rangle = \frac{\int_{-\infty}^{+\infty} v_z T_B(v_z, \mathbf{r}) dv_z}{\int_{-\infty}^{+\infty} T_B(v_z, \mathbf{r}) dv_z}. \quad (18)$$



**Fig. 2.** Integrated column density  $N_{\text{HI}}$  (optically thin approximation) of the 21 cm synthetic observation computed from the thermally bi-stable numerical simulation of Saury et al. (2014).

### 3.2.4. Synthetic observation

We computed the synthetic position-position-velocity (PPV) data cube in the general case using Eq. (14). Each spectrum has an effective velocity resolution of  $0.8 \text{ km s}^{-1}$  and covers  $-40 < v_z < 40 \text{ km s}^{-1}$ . We considered the beam  $B(\mathbf{r})$  of the instrument by convolving the synthetic PPV cube with a Gaussian kernel characterized by standard deviations of two pixels along the spatial axis. We then added a homogeneous Gaussian noise of  $0.05 \text{ K}$  to each spectrum.

In order to mimic observation, integrated column density maps shown in the rest of the paper are computed using the optically thin limit presented in Eq. (17). The integrated column density map of the synthetic PPV cube is shown in Fig. 2.

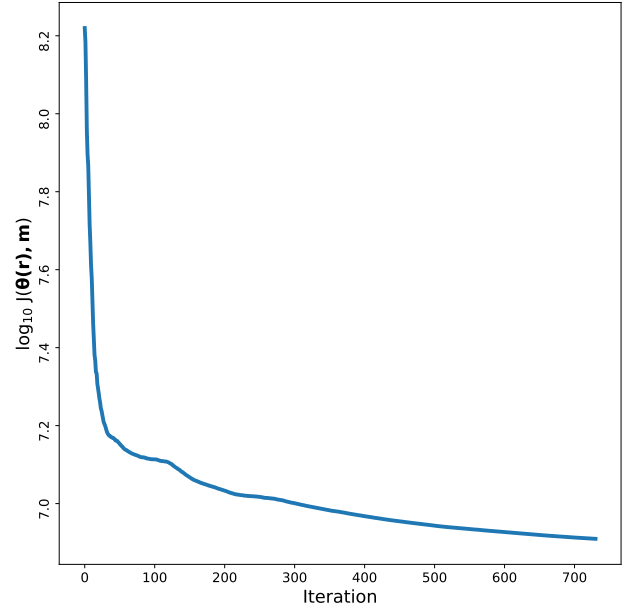
## 3.3. Results

ROHSA was then applied to decompose the synthetic observation computed in Sect. 3.2. In this section we discuss the choice of the free parameters of ROHSA, the global properties of the Gaussian sample, and the properties of individual components identified by the decomposition. Subsequently, the mapping of a three-phase coherent model is presented with direct comparisons to the phases extracted directly from the simulation.

### 3.3.1. Choosing the free parameters of ROHSA

ROHSA has six free parameters : the number of Gaussian components  $N$ , four hyper-parameters  $\lambda_i$  and, the maximum number of iterations of the LBFGS algorithm.

The most important parameter is  $N$ . This has to be sufficiently high to ensure a complete encoding of the signal, that is, to ensure that the residual is dominated by noise. As we discuss in Sect. 3.3.3, a given number of Gaussian components does not imply that all components are used to describe the signal along every line of sight. Components are allowed to have



**Fig. 3.** Evolution of the cost function  $J(\theta(\mathbf{r}), \mathbf{m})$  as a function of the number of iterations performed by ROHSA on the synthetic observation computed in Sect. 3.2.

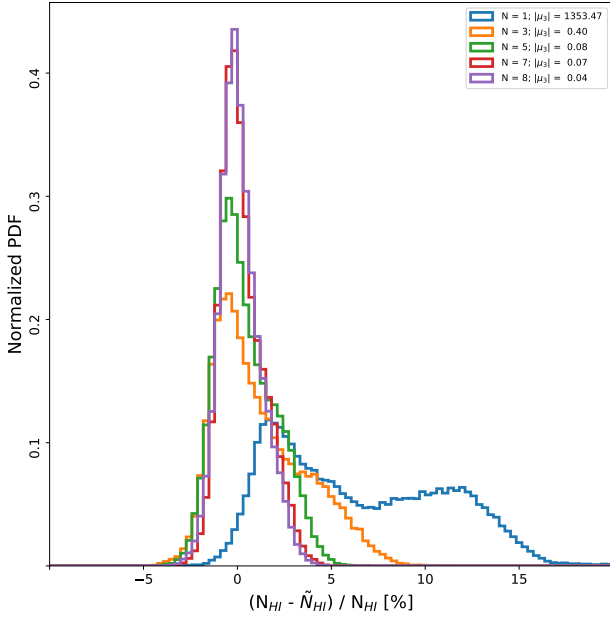
an amplitude of zero at any position. This is especially relevant for components encoding cold features. Since the CNM clouds occupy a small fraction of the total volume (see also Sect. 3.3.3), we expect the associated amplitude fields to have a large fraction close to zero. This is ensured by the energy term  $\|\sigma_n - m_n\|_2^2$  which minimizes the variance of the dispersion velocity of each component. Amplitudes are brought to zero if there is no need for a Gaussian to describe the signal at some location, instead of encoding another phase like the WNM for example. Avoiding phase mixing allows for overfitting to be prevented.

As with  $N$ , different values of the hyper-parameters  $\lambda_a$ ,  $\lambda_\mu$ ,  $\lambda_\sigma$ ,  $\lambda'_\sigma$  can be tested to obtain a satisfactory solution. If the hyper-parameters are null (i.e., no regularization) the signal could be fully encoded, but no spatial coherence will appear in the solution  $\theta^{(I)}$ . On the other hand, if the hyper-parameters are too high, the solution will tend towards a solution that could be too spatially coherent, or even flat, wiping out small-scale fluctuations and providing a bad fit to the data. A spatially coherent solution that describes the data well with the smallest value of  $N$  is a good criterion to select the values of the hyper-parameters.

The last criterion set by the user is the maximum number of iterations of the LBFGS algorithm computing Eq. (11); see Sect. 2.4.2. That parameter must be large enough to ensure the convergence of the solution. The convergence of ROHSA from a numerical perspective is presented in Fig. 3, which shows the evolution of the cost function  $J(\theta(\mathbf{r}), \mathbf{m})$  for 730 iterations.

The decomposition of the synthetic observations presented in this section converges to a satisfactory solution with  $N = 8$ ,  $\lambda_a = 10\,000$ ,  $\lambda_\mu = 10\,000$ ,  $\lambda_\sigma = 10\,000$ , and  $\lambda'_\sigma = 1000$ . As previously recommended, these values are empirically found to converge towards a noise-dominated residual and a signal that is encoded with a minimum number of Gaussian components. To illustrate this, Fig. 4 shows the normalized probability distribution function of the relative difference  $(N_{\text{HI}} - \tilde{N}_{\text{HI}})/N_{\text{HI}}$  between the solution  $\tilde{N}_{\text{HI}}$  inferred with ROHSA and the data  $N_{\text{HI}}$  for different values of  $N$ .

As an a posteriori assessment, the skewness  $\mu_3$  of the residual is shown to quantify the quality of the encoding. As the



**Fig. 4.** Normalized probability distribution function of the relative difference  $(N_{\text{HI}} - \tilde{N}_{\text{HI}})/N_{\text{HI}}$  between the solution  $\tilde{N}_{\text{HI}}$  inferred with ROHSA and the data  $N_{\text{HI}}$  for different numbers of Gaussian components  $N = [1, 3, 5, 7, 8]$ . The norm of the skewness  $|\mu_3|$  is shown in the legend to quantify the quality of the encoding.

emission is only positive and noise has a symmetric distribution centered on zero, the skewness of the residual can be used as a way to evaluate if the emission is well estimated or if the algorithm has over-fitted the data and included some noise in the solution. A positive skewness is usually an indication that some emission is left in the residual. If the model has not enough freedom to encode the emission fully; more components are needed or the regularization terms should be lowered. On the other hand, a negative skewness indicates that the decomposition is over-fitting the data; positive noise fluctuations are included in the model leaving more negative noise fluctuations than positive ones in the residual. This is usually an indication that the regularization coefficients ( $\lambda_i$ ) should be larger to increase the spatial smoothness of the solution. A skewness of the residual close to zero is an indication of a valid solution, one that is not distorted by the regularization and that does not overfit the signal. For the case of the synthetic observations presented here,  $N = 8$  fully encodes the signal with a relatively low skewness  $|\mu_3| = 0.04$ .

The computation time of ROHSA depends on the maximum number of iterations, the dimension of the PPV cube, and the number of Gaussian components. For each step of the multi-resolution process from coarse to fine grid, the computation time used by ROHSA for this particular case (purple line) is presented in Appendix B with a lower number of Gaussian components for the sake of comparison. The evaluation performed here requires nearly two hours of computation time on a single CPU, which makes it difficult to explore a large range of hyper-parameters and number of Gaussian components. To overcome this difficulty, a GPU implementation of ROHSA is under development.

### 3.3.2. Global properties of the Gaussian sample

ROHSA recovered the total emission of the synthetic observation with a relative variation of 0.3%. An example of the Gaussian decomposition for a representative  $4 \times 4$  mosaic of the simulation is shown Fig. 5. The spatial coherence of the solution can

be seen over the mosaic with a smooth variation of the amplitude, the central velocity, and the velocity dispersion of each Gaussian. It is already possible to distinguish in those spectra the convergence of the decomposition toward different velocity dispersions, that is, different temperatures/phases of the gas due to the energy term  $\lambda'_r \|\sigma_n - m_n\|_2^2$ . To have a clear view of these different components, let us take a look at the probability distribution function  $\sigma$  weighted by the fraction of total emission of each Gaussian  $\sqrt{2\pi} a_n \sigma_n / \sum_r N_{\text{HI}}(\mathbf{r})$  presented in Fig. 6. This diagram shows the amount of gas in a given range of velocity dispersion (i.e., indirectly a certain range of temperature). It is clear that ROHSA, in this case, converges toward a three-phase model with typical velocity dispersion close to the expected values in the CNM ( $\sigma < 2 \text{ km s}^{-1}$ ), LNM ( $\sigma \sim 6 \text{ km s}^{-1}$ ), and WNM ( $\sigma \sim 8 \text{ km s}^{-1}$ ). We see that a similar behavior is also present in the application to an observation of high Galactic latitude presented in Sect. 4. We note also that since eight Gaussian components have been used by ROHSA, some phases are encoded by several components. Association of the different components to characterize a three-phase model is presented in Sect. 3.3.4.

### 3.3.3. Properties of individual components

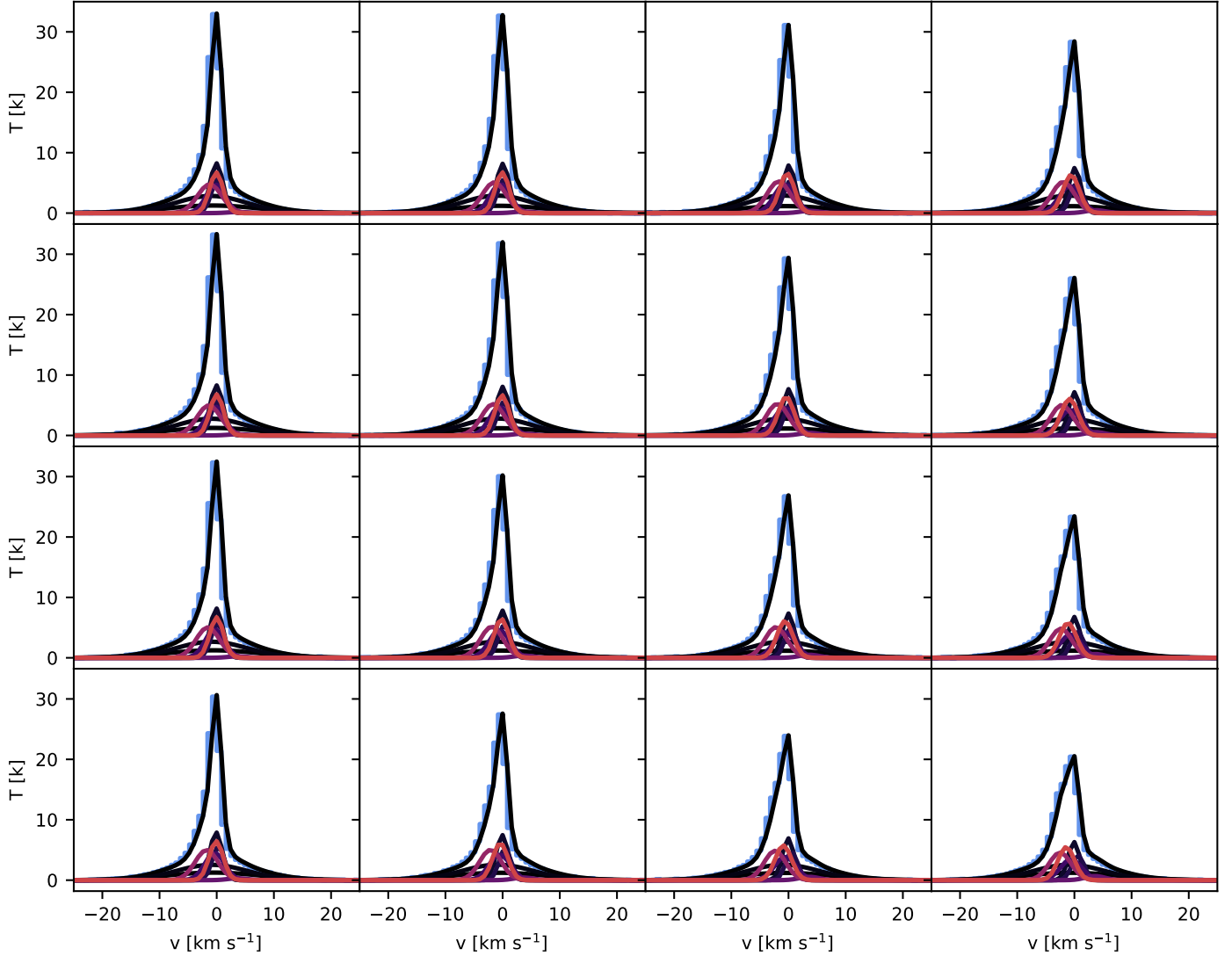
Integrated column density maps of each component obtained with ROHSA for  $N = 8$  Gaussian are presented in Fig. 7. For each component  $G_n$ , the mean velocity  $\langle \mu_n \rangle$  and mean velocity dispersion  $\langle \sigma_n \rangle$  averaged over the field are presented in Table 1. The surface filling factor appears to vary considerably between the eight components. The components with low values of  $\langle \sigma_n \rangle$  are sparsely present, while the component with the largest velocity dispersion is present everywhere on the field.

We recall that the numerical simulations used here were designed to reproduce the WNM–CNM condensation process of the HI through the thermal instability. The factor 100 difference in density between the two phases, and the fact that the mass fraction in each one is about 50%, implies that the cold phase fills only a few percent of the volume (Saury et al. 2014). This translates directly in the column density maps recovered by ROHSA; the narrow components, corresponding to colder structures, fill only a fraction of the projected field of view, while the larger component is present everywhere.

The eight velocity fields and velocity dispersion fields are presented in Figs. 8 and 9, respectively. At some location of the fields, when there is no need for a Gaussian to describe the signal over several pixels, the amplitude goes to zero. The corresponding velocity and velocity dispersion fields then have no reason to fluctuate. It turns out that where  $a_n$  goes to zero,  $\mu_n$  and  $\sigma_n$  are flat. This explains the apparent variation of spatial resolution as seen for example in Fig. 8 (bottom right). We notice that the first component  $G_1$  (top left) and the second component  $G_2$  (top right), encoding the WNM and the LNM, respectively, are defined everywhere, meaning that  $\mu_1$  and  $\sigma_1$  also have fluctuations everywhere. The implications of this are discussed in Sect. 5.

### 3.3.4. Mapping the three-phase neutral ISM

In order to compare the result of the decomposition with the reality given by the numerical simulations, we grouped the eight components into three fields corresponding to the WNM, the LNM, and the CNM. The comparison with the numerical simulations requires that we identify ranges in temperature that demarcate these three phases. As with most numerical simulations that include the classical heating and cooling processes



**Fig. 5.** Example of the Gaussian decomposition obtained by ROHSA for a random  $4 \times 4$  mosaic of the synthetic observation. The original signal is shown in blue and the total brightness temperature encoded by ROHSA is shown in black. The other lines show the individual Gaussian components. The spatial coherence of the solution can be seen over the mosaic with a smooth variation of the amplitude, the central velocity and the dispersion velocity of each component.

of the ISM (Wolfire et al. 1995), the simulation of Saury et al. (2014) we use here shows a continuum of temperature, with no clear separation and with a significant fraction of the gas present at temperatures corresponding to the thermally unstable regime (see their Figs. 14 and 15). To facilitate the comparison with previous studies, we decided to use the canonical values  $T_{k \text{ lim, CNM/LNM}} = 500 \text{ K}$  and  $T_{k \text{ lim, LNM/WNM}} = 5000 \text{ K}$  (Heiles & Troland 2003b) to separate the simulation in three components. The integrated column density maps associated to each phase are computed following the methodology described in Sect. 3.2.

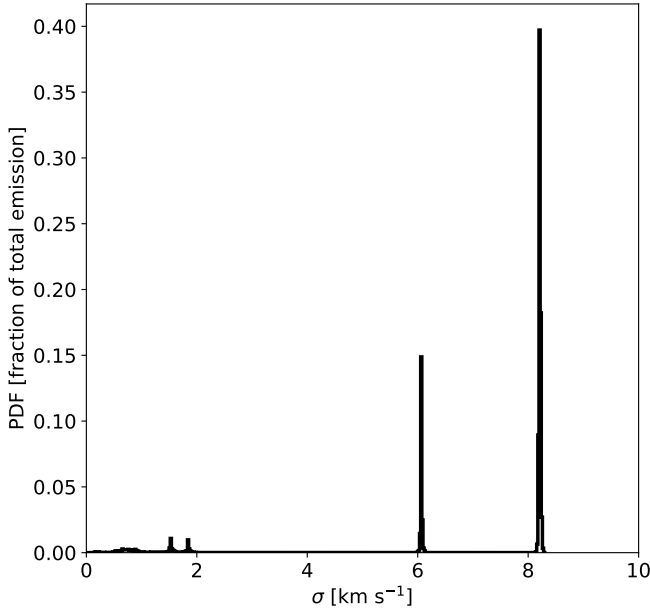
The comparison between the integrated column density maps recovered with ROHSA and those inferred directly from the simulation is presented Fig. 10. The intensity and the morphology of each phase is well recovered. It is nevertheless possible to see some leakages between the phases, in particular between WNM and LNM. This is due partly to the poorly defined temperature thresholds used to separate the phases. It is also due to small confusions during the Gaussian decomposition where the intensity, the dispersion velocity, and velocity centroid of each component in the PPV space are close to each other. In other words, for similar central velocities, the scales of fluctuations on the velocity

axis characterizing each component are too close to one another (see Figs. 5 and 6).

One way to evaluate the quality of the reconstruction is to compare the statistical properties of the cloud and inter-cloud components. Three different fields are used: the integrated column density field of the cloud medium (CNM), the integrated column density field of the inter-cloud medium (LNM + WNM), and, because it is fully sampled in the plan of sky, the centroid velocity field of the inter-cloud medium. Figure 11 presents the integrated column density field and the centroid velocity field, computed using Eq. (18), of the inter-cloud medium obtained combining the LNM and the WNM inferred with ROHSA and those obtained directly from the simulation.

In order to compare the estimates from ROHSA to the ones obtained from the simulation, we compute the spatial power spectrum (SPS) of each image. The SPSs of the integrated column density of the cloud and inter-cloud medium are presented Fig. 12 and the SPS of the centroid velocity field of the inter-cloud medium is presented in Fig. 13. In each case, the statistics recovered by ROHSA is consistent with the numerical simulation over all scales. The shape of these power spectra is interesting in





**Fig. 6.** Probability distribution function  $\sigma$  weighted by the fraction of total emission of each Gaussian  $\sqrt{2\pi}a_n\sigma_n/\sum_r N_{\text{HI}}(\mathbf{r})$  of the simulated field. ROHSA converges toward three distinguishable phases associated to the WNM, LNM, and CNM.

**Table 1.** Mean velocity  $\langle\mu_n\rangle$  and mean velocity dispersion  $\langle\sigma_n\rangle$  of the eight Gaussian components  $G_n$  inferred by ROHSA on the 21 cm synthetic observation of the numerical simulation described in Sect. 3.2.

	$G_1$	$G_2$	$G_3$	$G_4$	$G_5$	$G_6$	$G_7$	$G_8$
$\langle\mu_n\rangle$ (km s $^{-1}$ )	0.2	-1.7	0.5	2.3	-0.1	5.0	-3.7	-2.5
$\langle\sigma_n\rangle$ (km s $^{-1}$ )	8.2	6.1	0.5	0.6	1.5	1.5	1.8	0.9

itself; the inter-cloud medium is featureless with an almost constant power law, as the cloud phase is more structured, with a break at about 20 pixels, showing a typical scale linked to the condensation process. Interestingly, ROHSA is able to capture all these features very well.

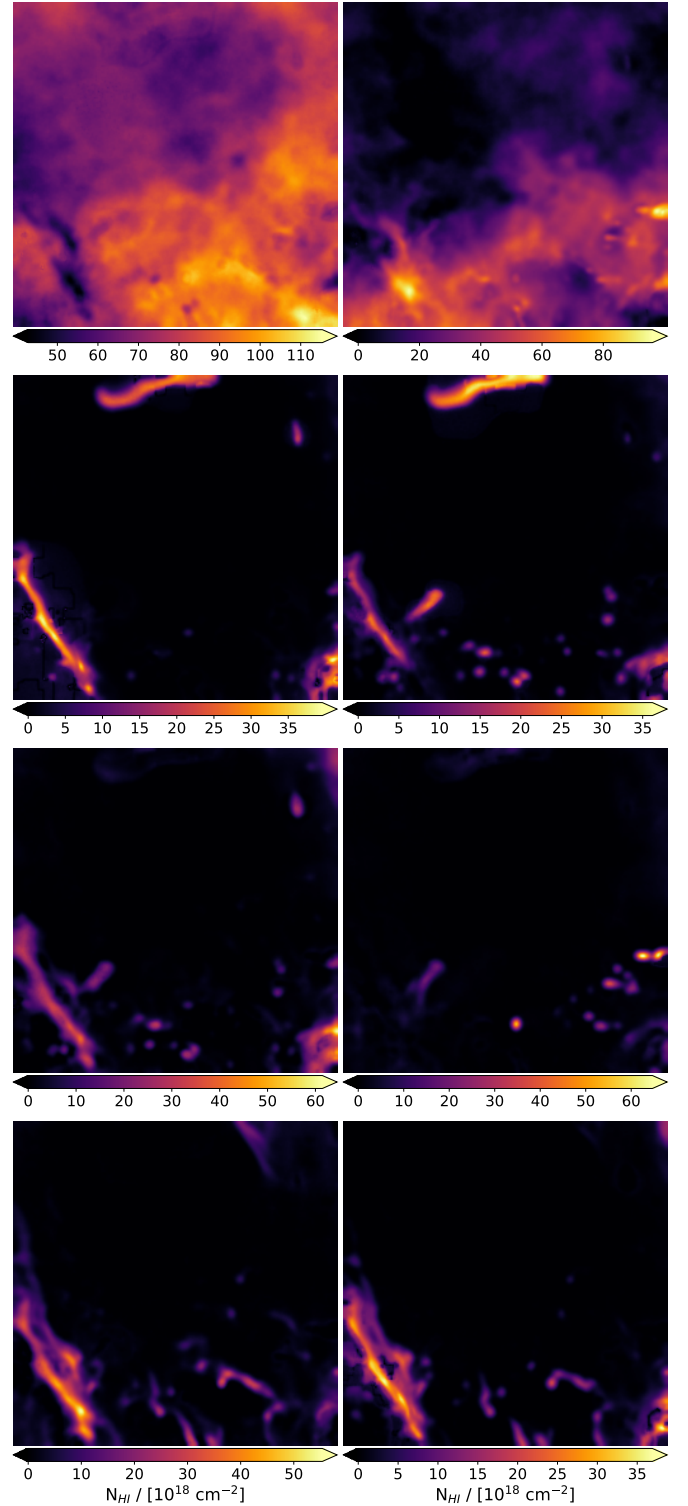
## 4. Application on high-latitude HI gas

After validating the identification of the HI phases on numerical simulations, in this section we present the application of ROHSA on a 21 cm observation of a region with high Galactic latitude.

### 4.1. North ecliptic pole

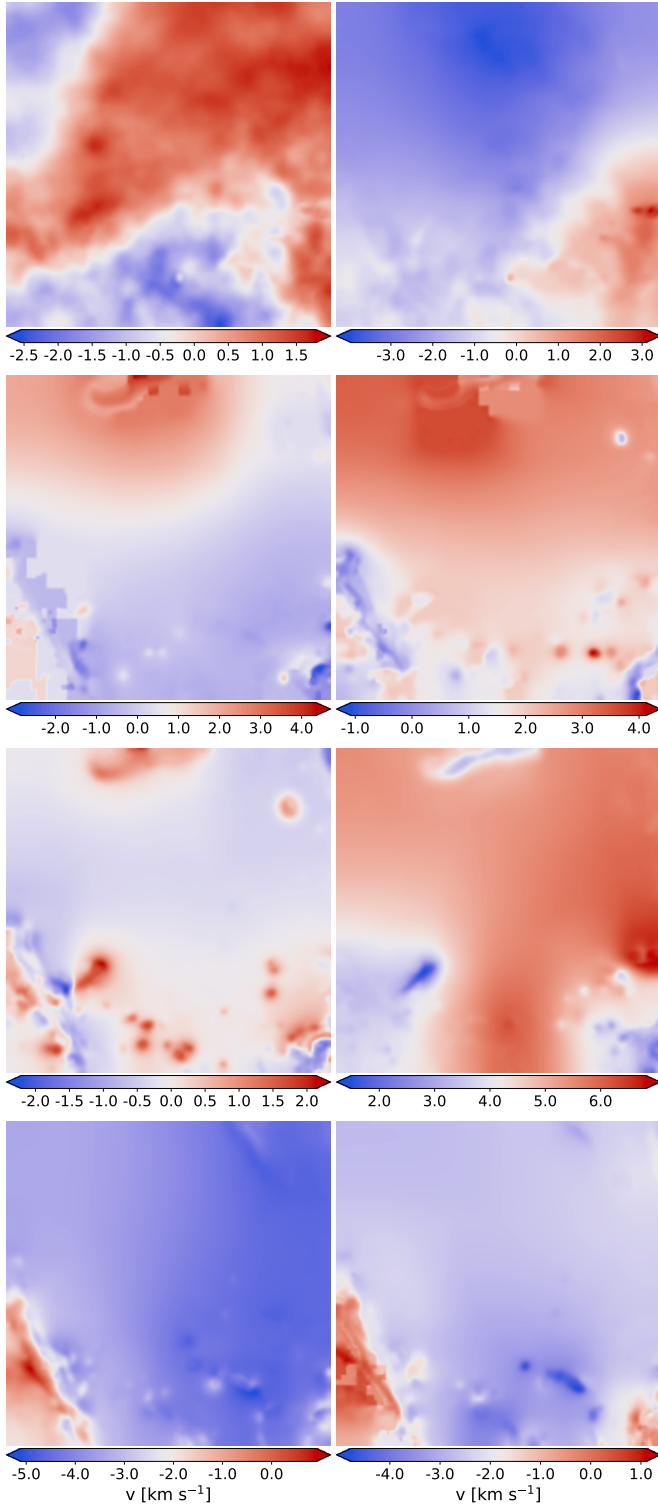
To avoid the complication of low-latitude observations, where the 21 cm emission is significantly affected by velocity crowding and self-absorption, we chose to apply ROHSA to one of the high Galactic latitude fields of the GHIGLS<sup>1</sup> survey (Martin et al. 2015). We chose the North ecliptic pole (NEP) field, a  $12^\circ \times 12^\circ$  region centered on  $l = 96^\circ.40$ ,  $b = 30^\circ.03$  observed with the Green Bank Telescope, providing a  $9''.55$  spatial resolution. The HI spectra have an effective velocity resolution of  $0.807 \text{ km s}^{-1}$  and cover  $-200 < v \text{ (km s}^{-1}\text{)} < 50$ . The integrated column density map computed using Eq. (17) is shown in Fig. 14 and a mosaic of representative emission spectra is shown in Fig. 15.

<sup>1</sup> <http://www.cita.utoronto.ca/GHIGLS/>



**Fig. 7.** Integrated column density maps (left:  $G_1, G_3, G_5, G_7$ ); (right:  $G_2, G_4, G_6, G_8$ ) obtained by ROHSA on the synthetic observation computed in Sect. 3.2. Mean velocity  $\langle\mu_n\rangle$  and mean velocity dispersion  $\langle\sigma_n\rangle$  are presented in Table 1. The surface filling factor varies considerably between components, depending on their  $\langle\sigma_n\rangle$  value.

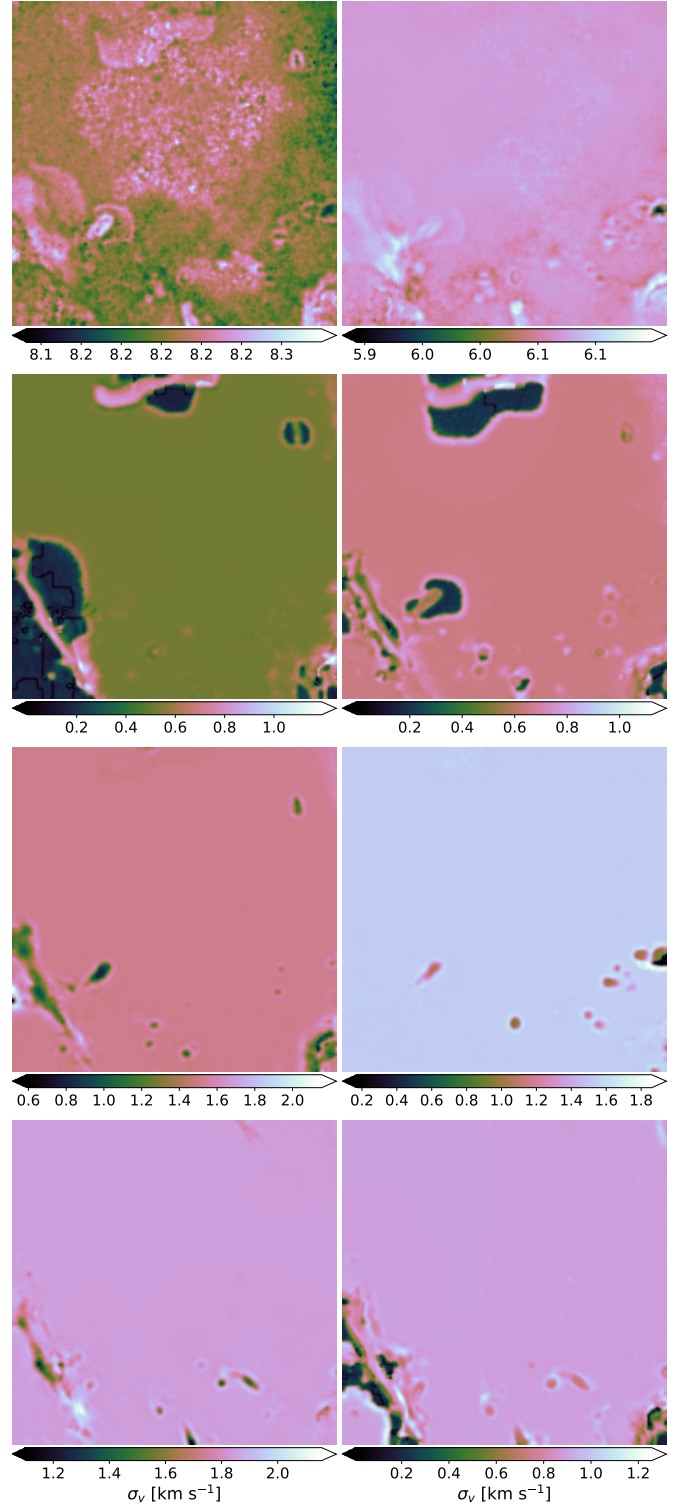
As Fig. 15 shows, high-latitude spectra of HI are more complex than the synthetic observations computed from the numerical simulations of Saury et al. (2014). This is caused by a combination of the longer line of sight in the observation (about 200 pc at  $b = 30^\circ$  compared to the 40 pc box of the simulation) and to the presence of nonlocal velocity components. The 21 cm



**Fig. 8.** Centroid velocity fields  $\mu$  (left:  $\mu_1, \mu_3, \mu_5, \mu_7$ ); (right:  $\mu_2, \mu_4, \mu_6, \mu_8$ ) obtained by ROHSA using the synthetic observation computed in Sect. 3.2.

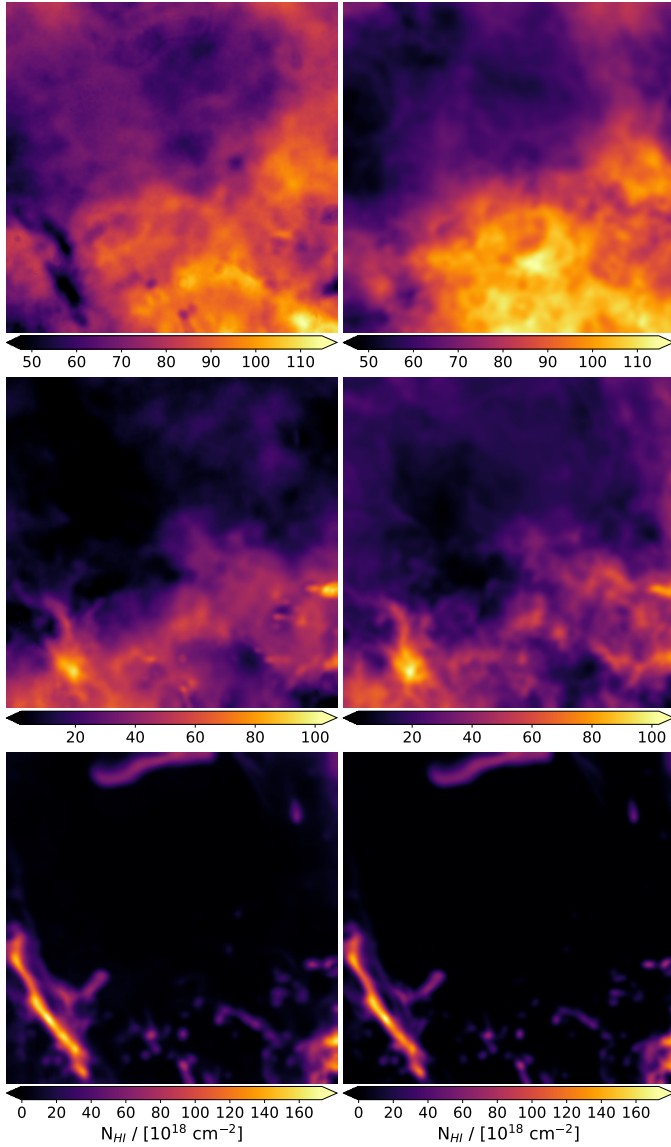
emission in NEP indeed exhibits significant emission in the intermediate velocity cloud (IVC) and high velocity cloud (HVC) ranges. In this paper we do not consider the velocity channels with HVC emission; we focus on the phase separation of the local velocity cloud (LVC) and IVC components.

The LVC range between  $-20$  and  $+20$   $\text{km s}^{-1}$  shows relatively smooth emission profiles, with a narrow peak around  $-3$   $\text{km s}^{-1}$  on top of a broader feature (see Fig. 15). The latter is rather



**Fig. 9.** Dispersion velocity fields  $\sigma$  (left:  $\sigma_1, \sigma_3, \sigma_5, \sigma_7$ ); (right:  $\sigma_2, \sigma_4, \sigma_6, \sigma_8$ ) obtained by ROHSA using the synthetic observation computed in Sect. 3.2.

smooth but when inspected in detail it shows faint spectral structures on all scales along the velocity axis. The sensitivity of the GHIGLS data is such that these fluctuations of the emission profiles are not due to noise. In fact, they can be followed from one spectrum to the next quite easily. These fluctuations of the emission spectra at scales of a few kilometres per second reveal the presence of CNM and LNM features on a range of velocities.

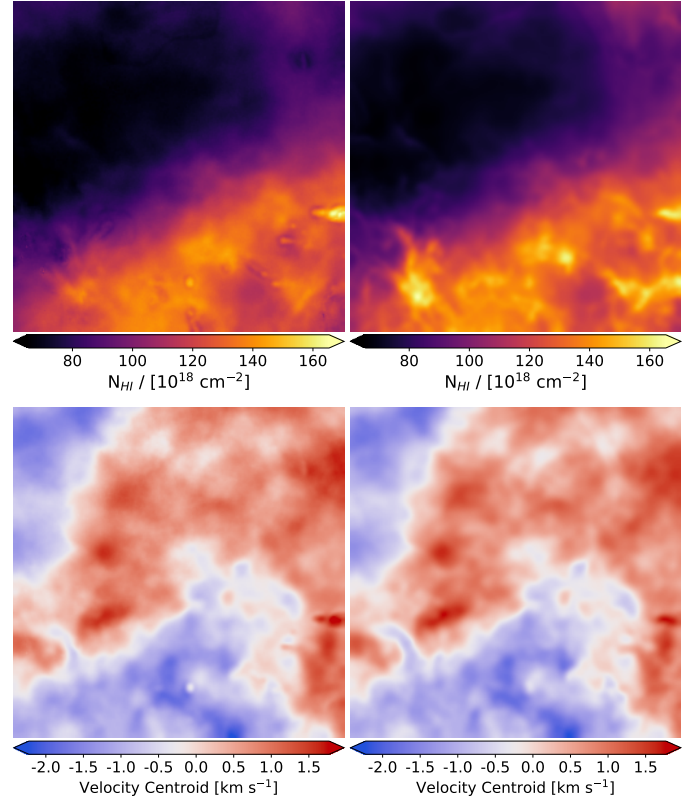


**Fig. 10.** *Left:* integrated column density maps of the three-phase model extracted by ROHSA. *Right:* integrated column density maps of the three-phase model inferred directly from the simulation using the canonical values  $T_{k\text{lim,CNM/LNM}} = 500$  and  $T_{k\text{lim,LNM/WNM}} = 5000$  K. The phases WNM, LNM, and CNM are presented from top to bottom.

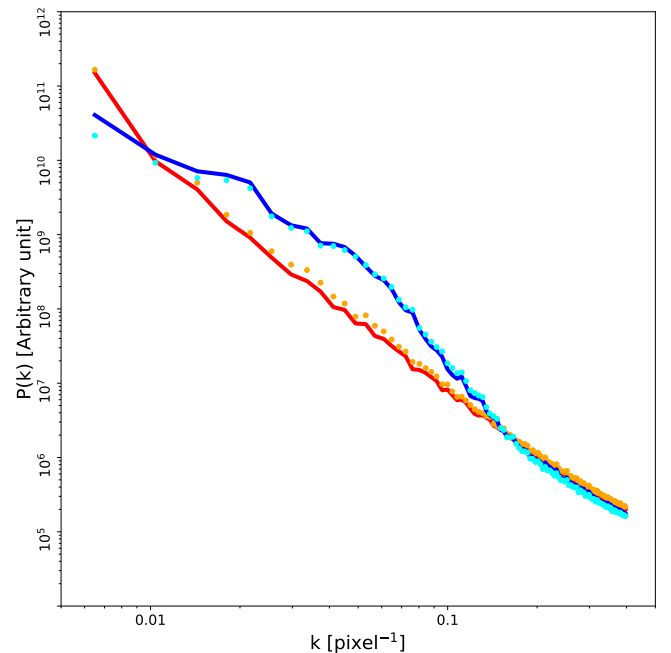
This field was selected because of its representative 21 cm emission for Galactic latitudes of  $b \sim 30^\circ$ . The emission features are not particularly complex, nor are they especially simple. In addition, a first Gaussian decomposition of NEP 21 cm data was performed by [Martin et al. \(2015\)](#) which provides an interesting point of comparison. Unlike ROHSA, [Martin et al. \(2015\)](#) used a method similar to the one described by [Haud \(2000\)](#) that considers only the term  $\|L(v_z, \mathbf{r}, \boldsymbol{\theta}(\mathbf{r}))\|_2^2$  in the parameter optimization. We highlight some qualitative comparisons in the following sections.

#### 4.2. Results

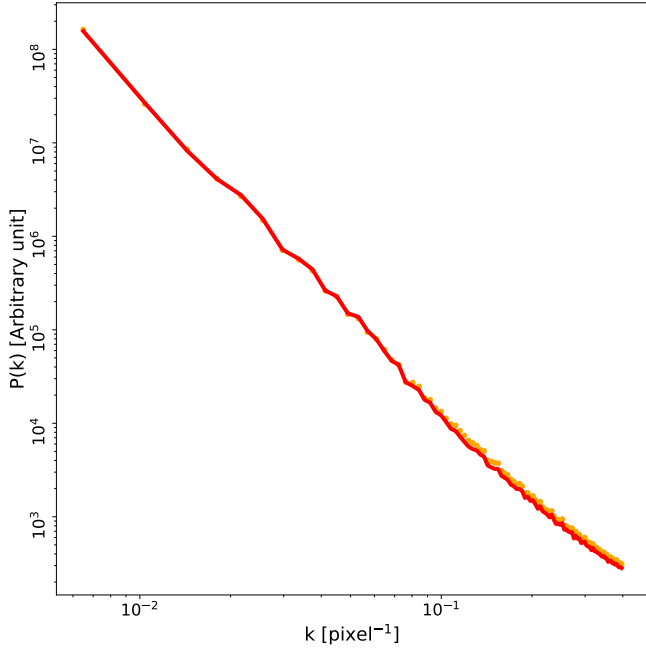
In order to decompose the 21 cm emission of the NEP field, we used ROHSA with  $N = 12$  Gaussian components and each hyper-parameter has been set to 1000. As for the previous case, these values have been chosen empirically following the same methodology as described in Sect. 3.3.1, allowing us to converge



**Fig. 11.** Synthetic observation of the integrated column density field (*top*) and the centroid velocity field (*bottom*) associated to the inter-cloud medium (WNM + LNM). *Left:* inferred with ROHSA. *Right:* inferred directly from the simulation using all cells with  $T_{k\text{lim,LNM/WNM}} > 500$  K. For statistical comparison, the spatial power spectra of each one are shown in Figs. 12 and 13.



**Fig. 12.** Spatial power spectrum of the column density. The inter-cloud medium (WNM + LNM) is represented by the orange dotted line (simulation) and the red line (ROHSA). The CNM is shown as a cyan dotted line (simulation) and blue line (ROHSA).



**Fig. 13.** Spatial power spectrum of the centroid velocity field for the inter-cloud medium (WNM+LNM). The orange dotted line indicates the inter-cloud medium inferred directly from the simulation using all cells with  $T_{k, \text{lim}, \text{LNM}/\text{WNM}} > 500$  K (bottom-right panel of Fig. 11). The red line shows the inter-cloud medium inferred with ROHSA (bottom-left panel of Fig. 11).

toward a noise-dominated residual with a minimum number of Gaussian components. We note that the hyper-parameter values are not the same as for the first application presented in Sect. 2. The complexity of the underlying signal structure and its signal-to-noise ratio (S/N) are the main causes of these differences. However, a detailed understanding of the behavior of these hyper-parameters would require testing different values over a large number of observations. Such exploration is currently complicated by computations limitations. A GPU version of the code is under development to overcome this main limitation.

The combination of 12 Gaussian components produces a solution that recovers 99% of the total emission with spatially coherent components. The total integrated column density encoded by ROHSA and the residual between our model and the data are shown in Fig. 14 (middle and right panels).

#### 4.2.1. Global properties of the Gaussian sample

Like for the application on the synthetic observations, the Gaussian parameters recovered for the NEP field have a strong spatial coherence; ROHSA converges towards a solution with smooth variations of the Gaussian parameters across the field. It turns out that ROHSA converges toward a multiphase model with Gaussian components of various widths, very similar to the application to numerical simulations presented in Sect. 3.3 but more complex due to the presence of an IVC component in the data.

To have a global view of the thermal state of the gas as a function of velocity, it is useful to look at the two-dimensional dispersion-velocity diagram  $\sigma - v$  weighted by the fraction of total emission of each Gaussian  $\sqrt{2\pi}a_n\sigma_n / \sum_r N_{\text{HI}}(\mathbf{r})$  shown in Fig. 16. This diagram shows isolated complexes of Gaussian components in the  $\sigma - v$  space. This is a direct result of the

way the parameter optimization is done in ROHSA, with a regularization term that favors the minimum variance of  $\sigma$ . Figure 16 highlights the fact that the 21 cm emission in NEP is mainly composed of negative-velocity components. A clear trend is visible in the velocity range  $-60 < v < 0$  km s<sup>-1</sup> with  $\sigma$  decreasing from 10 to 1 km s<sup>-1</sup> going from negative to positive velocities. This likely reflects the radiative condensation of warm intermediate velocity clouds into the local velocity component of the neutral ISM.

At this point it is interesting to compare the results of ROHSA with the Gaussian decomposition of the same data performed by Martin et al. (2015). The  $\sigma - v$  diagram of Martin et al. (2015, see their Fig. 7) shows a continuous distribution with arches that bridge together the LVC and IVC gas, an effect that is not observed in our results. Similarly to what we have done here, Martin et al. (2015) used numerical simulations to evaluate the performances of their Gaussian decomposition algorithm. Their tests revealed that such arches in the  $\sigma - v$  diagram are unphysical; they are the result of LVC and IVC gas components that overlap in velocity.

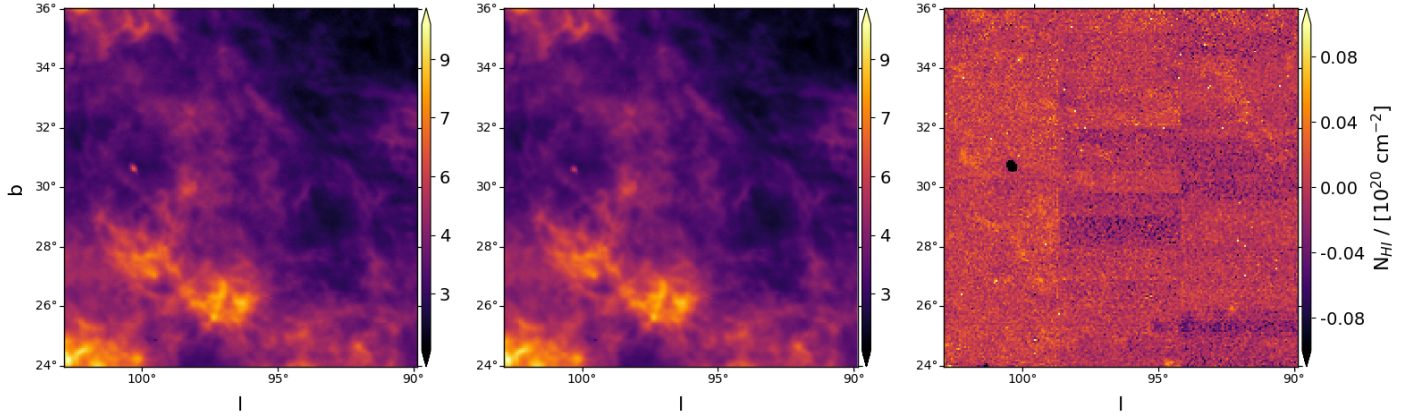
It is important to point out that both solutions provide as good a representation of the same dataset. The significant difference between the two solutions highlights the challenge of extracting a physically meaningful representation of the data. We recall that Martin et al. (2015) used an algorithm similar to the ones used by Haud (2000); Miville-Deschênes et al. (2017a); Kalberla & Haud (2018) where the spatial coherence of the solution is not enforced through regularization terms in the cost function. In these previous studies spatial coherence is attempted by providing spatially coherent initial guesses. Each spectrum is then fitted independently and no spatial coherence in the solution is enforced. In practice, this method is rather effective for relatively sparse emission data like CO (Miville-Deschênes et al. 2017a), but in the case of the more confused 21 cm data it produces parameter maps that are more affected by small-scale noise due to the degeneracy of the solution.

A Gaussian decomposition algorithm that fits each spectrum individually is easily fooled by components that overlap in velocity. In this specific case, such an algorithm would find a solution with a smaller number of components but with larger values of  $\sigma$ . The main innovation in ROHSA is that it is able to cluster different phases even if they are close in velocity. The four energy terms added to the cost function  $J(\theta(\mathbf{r}))$  allow ROHSA to find a spatially coherent solution while avoiding the mix of components due to the high confusion present in the emission.

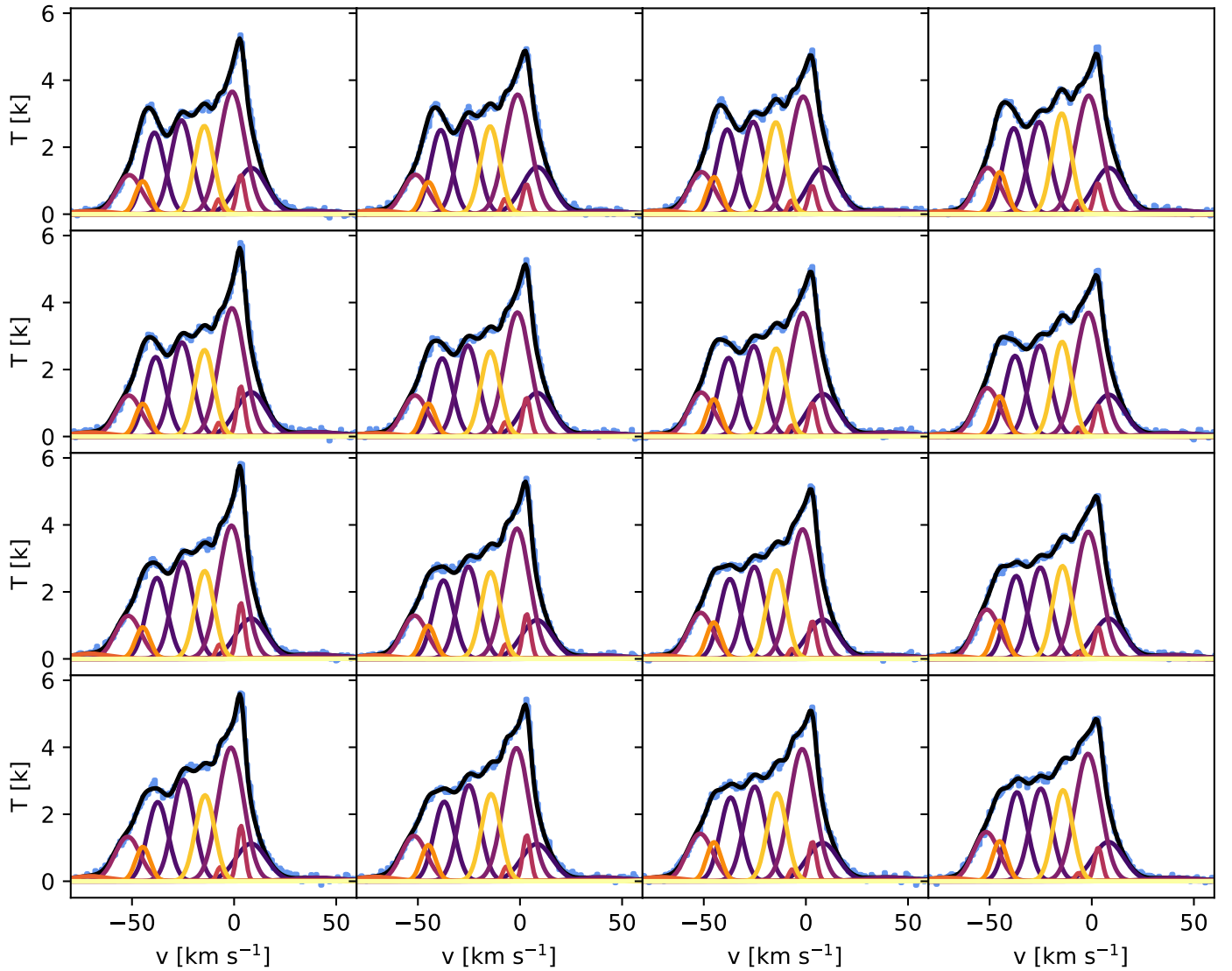
#### 4.2.2. A cloud/inter-cloud medium vision of the North ecliptic pole

Integrated column density maps, centroid velocity fields, and dispersion velocity fields obtained with ROHSA are presented in Figs. 17–19, respectively. Mean velocities  $\langle \mu_n \rangle$  and mean velocity dispersions  $\langle \sigma_n \rangle$  of the 12 Gaussian components  $G_n$  are presented in Table 2. In this section we focus on building a coherent cloud/inter-cloud medium vision considering the local component of the emission identified previously. Two of the four components of the local gas,  $G_7$  and  $G_9$ , are associated with the CNM with  $\langle \sigma_7 \rangle = 1.6$  km s<sup>-1</sup> and  $\langle \sigma_9 \rangle = 1.9$  km s<sup>-1</sup>. The other components are used to build the inter-cloud medium. Integrated column density fields and centroid velocity fields of the cloud medium and inter-cloud medium are presented in Fig. 20.

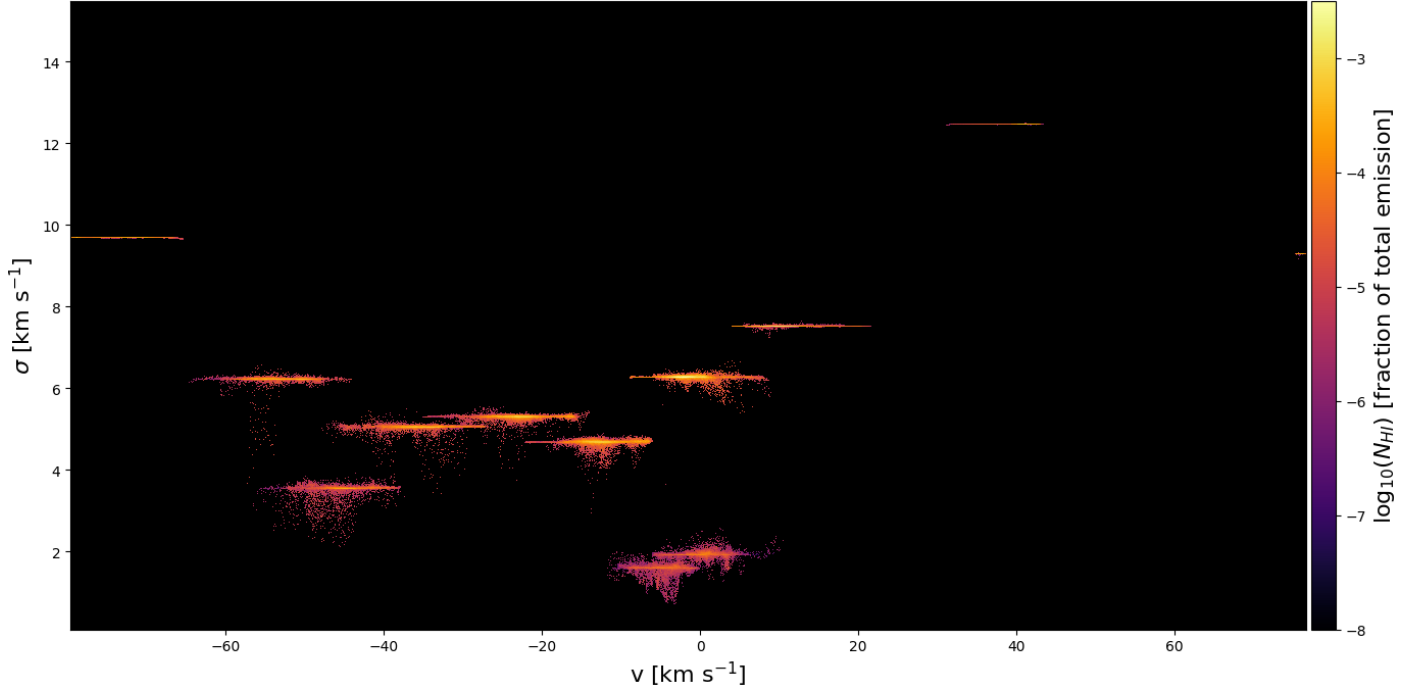
As noted by Martin et al. (2015) in their two-phase decomposition of the local component, filamentary structures are observed in the narrow component. The associated velocity



**Fig. 14.** *Left:* integrated column density  $N_{\text{HI}}$  of the NEP field which is part of the GHIGLS survey.  $N_{\text{HI}}$  was computed in the optically thin approximation (see Eq. (17)). *Middle:* integrated column density  $\hat{N}_{\text{HI}}$  of NEP inferred with ROHSA. *Right:* residual  $\hat{N}_{\text{HI}} - N_{\text{HI}}$  between the integrated column density field inferred with ROHSA and the original integrated column density field computed with the data.



**Fig. 15.** Example of the Gaussian decomposition obtained by ROHSA (colored line) for a random  $4 \times 4$  mosaic of NEP. The original signal is shown by the blue histogram and the total brightness temperature encoded by ROHSA is shown in black. The other lines detail the components of the Gaussian model. The spatial coherence of the solution can be seen over the mosaic with a smooth variation of the amplitude, central velocity, and velocity dispersion of each Gaussian component.



**Fig. 16.** Two-dimensional probability distribution function  $\sigma$ - $v$  weighted by the fraction of total emission of each Gaussian  $\sqrt{2\pi}a_n\sigma_n/\sum_r N_{\text{HI}}(r)$  of NEP. The NEP is mainly composed of negative intermediate velocity components.

**Table 2.** Mean velocity  $\langle\mu_n\rangle$  and mean velocity dispersion  $\langle\sigma_n\rangle$  of the 12 Gaussian components  $G_n$  inferred by ROHSA in NEP.

	$G_1$	$G_2$	$G_3$	$G_4$	$G_5$	$G_6$	$G_7$	$G_8$	$G_9$	$G_{10}$	$G_{11}$	$G_{12}$
$\langle\mu_n\rangle$ (km s $^{-1}$ )	-74.1	-53.9	-44.7	-35.0	-22.9	-12.6	-4.8	-1.3	0.2	10.9	40.8	75.9
$\langle\sigma_n\rangle$ (km s $^{-1}$ )	9.7	6.2	3.5	5.1	5.3	4.7	1.6	6.3	1.9	7.5	12.5	9.3

dispersion fields (see Fig. 19, component  $G_7$  and  $G_9$ ) coherent fluctuating values over a large part of the field. The core of these filamentary structures appears narrower than the envelop with velocity dispersion reaching about  $0.87 \text{ km s}^{-1}$  (the spectral resolution) in their centers.

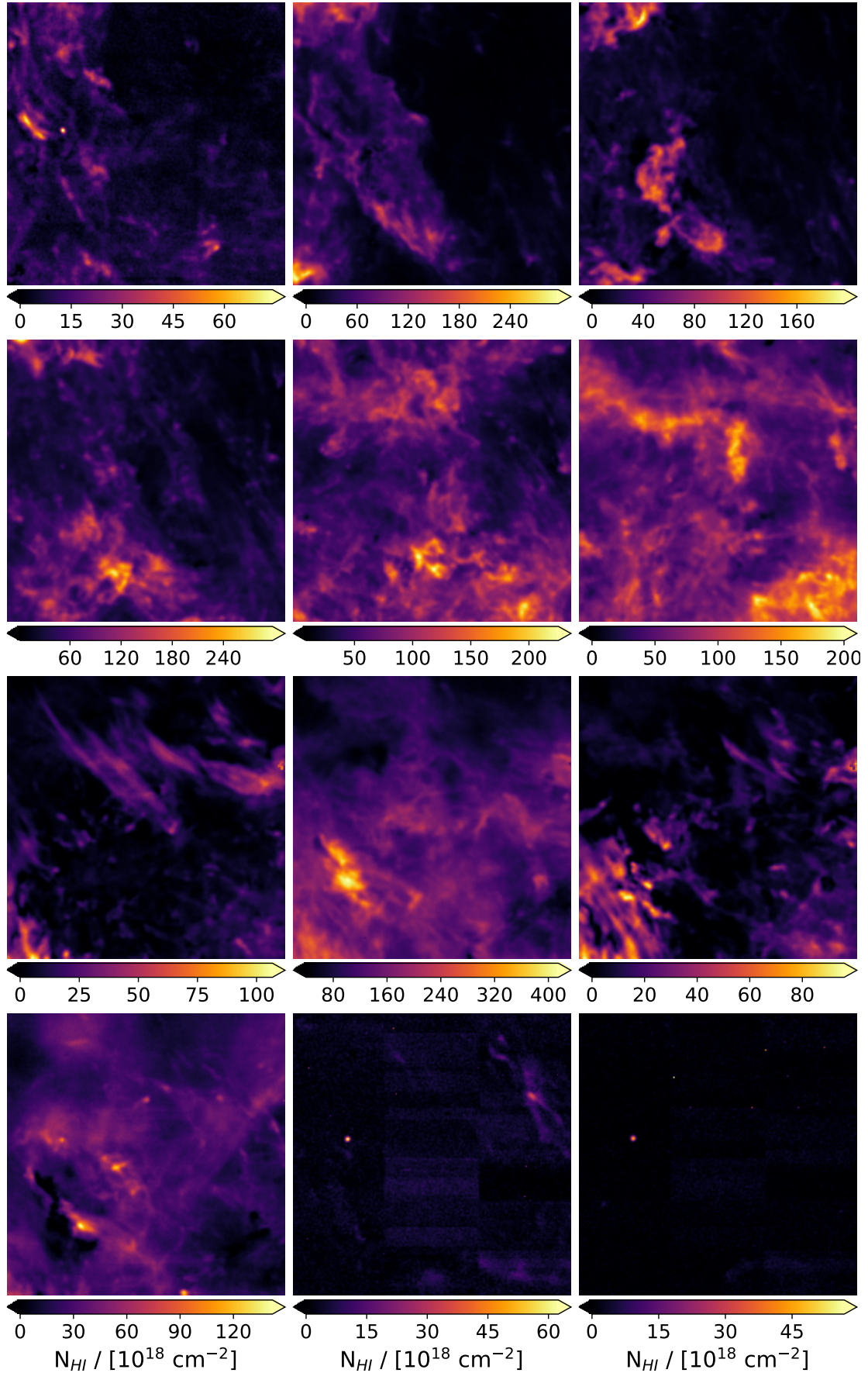
The broader component has an integrated column density field with no particular structure like filaments (see Fig. 20, top-right). Like for the numerical simulation, the sum of the broad components is likely to represent a phase that fills a large fraction of the volume, as would an inter-cloud medium. One interesting aspect of the ROHSA decomposition is that it then allows to extract the velocity field of this volume-filling component (Fig. 20, bottom-right), enabling the characterization of the turbulent cascade in a mixture of lukewarm phase and warm phase.

## 5. Discussion

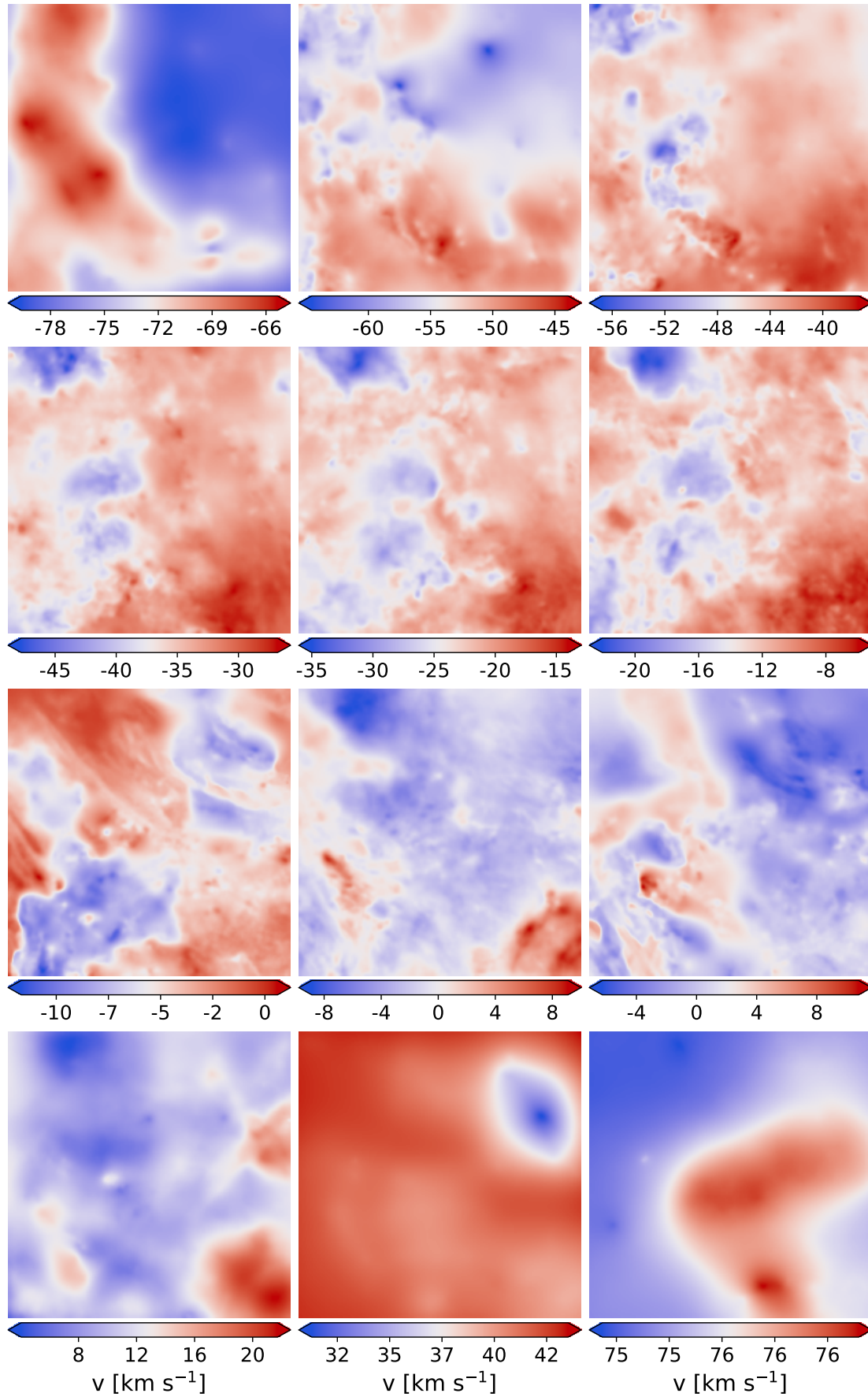
Historically, a large number of studies used a Gaussian basis to model 21 cm data. Different algorithms have been developed; all of them are fitting each spectrum individually, with or without information from the neighboring solutions to initialize the fit. To further constrain the degeneracy of the fit, solutions with the smallest number of Gaussian components have often been favored (e.g., Lindner et al. 2015). Because of velocity blending, the solution with the smallest number of components is not necessarily the best one. In some cases, narrow features overlap in velocity, making it impossible to separate them if the environment is not considered. Usually, this confusion breaks apart

a few beams away and more components can be recovered. The fundamental idea behind ROHSA is that we are trying to extract diffuse components that have column density, centroid velocity, and velocity dispersion with smooth spatial variations. The optimization scheme has been designed with that concept at its core. In order to achieve this, ROHSA fits the whole data cube at once.

The application of ROHSA on both synthetic observations from numerical simulations and observational data converges naturally toward a multiphase model of the neutral ISM. The ability of ROHSA to extract the multiphase nature of the neutral ISM opens a totally new perspective on the study of the nature of the condensation process acting in the ISM. It is clear from a numerical point of view that the formation of cold clouds is the result of the condensation of the warm and diffuse gas through the thermal instability coupled with turbulence. From an observational point of view, spatial correlations between the different phases can now be made in order to more precisely quantify how the CNM emerges from this condensation process. This separation also opens the possibility to describe the properties of the very specific multiphase turbulence of the HI. ROHSA appears to be efficient at clustering different structures in PPV space, even when there is a high level of confusion. In particular, the separation of the LVC and IVC is known to be particularly challenging as the CNM and WNM of both components overlap significantly in velocity (Martin et al. 2015). As shown in Fig. 16, ROHSA significantly limits the ‘‘Arch effect’’ typical of this confusion (see Sect. 4). Globally, the performance in ROHSA on regions of high

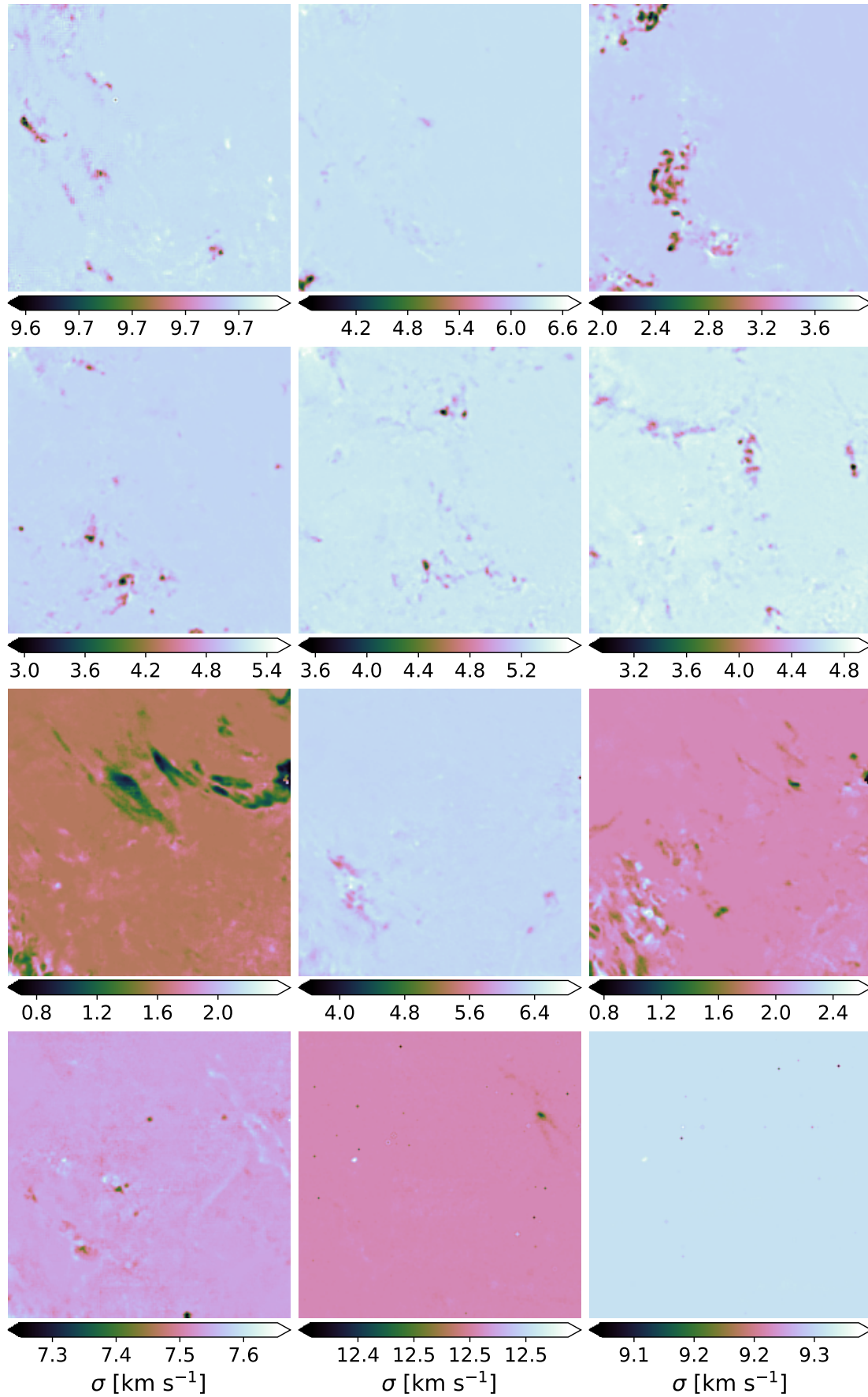


**Fig. 17.** Integrated column density maps (left:  $G_1, G_4, G_7, G_{10}$ ; middle:  $G_2, G_5, G_8, G_{11}$ ; right:  $G_3, G_6, G_9, G_{12}$ ) obtained by ROHSA applied on NEP. Mean velocity  $\langle \mu_n \rangle$  and mean velocity dispersion  $\langle \sigma_n \rangle$  are presented in Table 2.

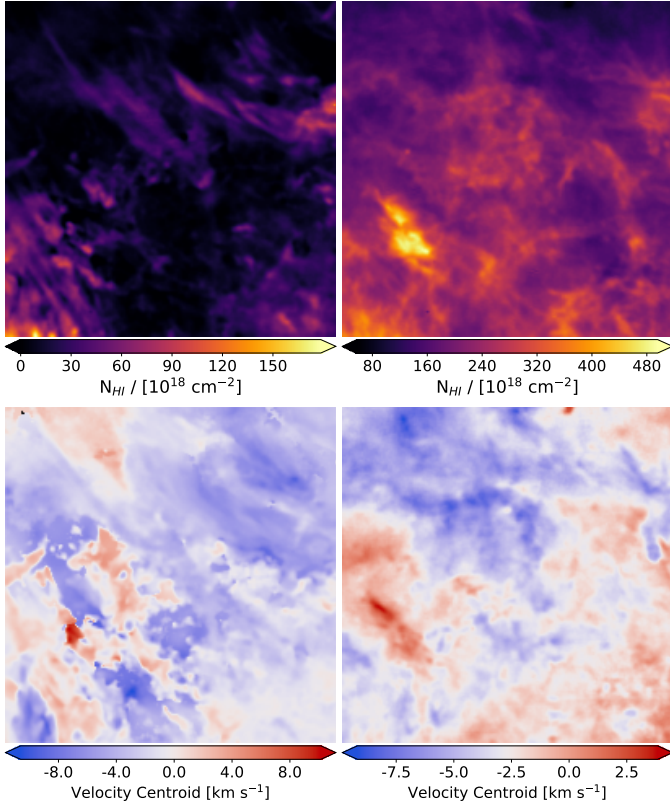


**Fig. 18.** Centroid velocity fields  $\mu$  (left:  $\mu_1, \mu_4, \mu_7, \mu_{10}$ ; middle:  $\mu_2, \mu_5, \mu_8, \mu_{11}$ ; right:  $\mu_3, \mu_6, \mu_9, \mu_{12}$ ) obtained by ROHSA applied on NEP.





**Fig. 19.** Velocity dispersion maps  $\sigma$  (left:  $\sigma_1, \sigma_4, \sigma_7, \sigma_{10}$ ; middle:  $\sigma_2, \sigma_5, \sigma_8, \sigma_{11}$ ; right:  $\sigma_3, \sigma_6, \sigma_9, \sigma_{12}$ ) obtained by ROHSA applied on NEP.



**Fig. 20.** *Left:* CNM. *Right:* inter-cloud medium (WNM+LNM) in NEP inferred with ROHSA. *Top and bottom:* column density and centroid velocity fields, respectively.

Galactic latitude opens a large range of possibilities regarding the study of infalling neutral clouds from the galactic halo.

We would also like to point out that no a priori information about the number of phases present in the neutral ISM is provided to ROHSA. The algorithm rests only on the hypothesis of the existence of components with similar line width through the energy term  $\lambda'_\sigma \|\sigma_n - m_n\|_2^2$ . In that respect, ROHSA is perfectly adapted to decomposing hyper-spectral observations of any type, not only 21 cm emission.

At this time the main limitations of ROHSA are computational. First, as the whole PPV cube is fitted at once, the use of ROHSA is limited to cubes that can fit in memory. Second, the current computation time of ROHSA is not negligible (e.g., about two hours on a single CPU for a  $256 \times 256 \times 100$  PPV cube with 8 Gaussians). This limits the possibility to make a deep exploration of the hyper-parameters  $\lambda_i$ . In particular it would be interesting to explore various weights of the hyper-parameters for the deduced quantities ( $a$ ,  $\mu$  and  $\sigma$ ). It is expected that the amplitude of the spatial variations of these quantities are not the same. For instance, in a multi-phase medium like the HI, the density field (represented by  $a$ ) might vary more strongly on smaller scales than the velocity field. This might require different values of  $\lambda_a$  compared to  $\lambda_\mu$ ,  $\lambda_\sigma$  and  $\lambda'_\sigma$ . A GPU version of the code is under development that would allow such an exploration.

## 6. Summary

Here we present a new Gaussian decomposition algorithm named ROHSA. Energy terms have been added to the classical cost function to take into account the spatial coherence of the emission and the multiphase nature of the gas simultaneously. In

order to identify a solution with spatially smooth parameters, the fit is performed on the whole hyper-spectral cube at once.

The performance of ROHSA has been evaluated using a synthetic 21 cm observation computed from a numerical simulation of thermally bi-stable turbulence. It was then tested on a 21 cm observation of a field of high Galactic latitude observed with the GBT. The main conclusions are as follows.

1. ROHSA is able to naturally highlight the physics of any multiphase medium without a priori information regarding the number of phases.
2. Evaluation on numerical simulation of thermally bi-stable turbulence shows that the sum of Gaussian components is a good approximation to model the multiphase nature of the neutral ISM.
3. The multiphase model inferred with ROHSA provides a spatially coherent vision of the integrated column density map, the centroid velocity field, and the velocity dispersion field of each component.
4. The power spectra of the integrated column density and centroid velocity fields are well recovered with ROHSA. Statistical properties of turbulence in the multiphase neutral ISM now become accessible.
5. The decomposition of a high-latitude HI gas observation shows the wide range of applications enabled with ROHSA, for instance to study the radiative condensation of the WNM and the nature of the ISM at the disk-halo interface.

*Acknowledgements.* Part of this work was supported by Hyperstars, a project funded by the MASTODONS initiative of the CNRS mission for interdisciplinarity and by the Programme National “Physique et Chimie du Milieu Interstellaire” (PCMI) of CNRS/INSU with INC/INP co-funded by CEA and CNES. This work took part under the program Milky-Way-Gaia of the PS12 project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02. We gratefully acknowledge Lucie Riu for enlightening conversations. We thank the anonymous referee whose comments and suggestions have improved this manuscript.

## References

- Audit, E., & Hennebelle, P. 2005, *A&A*, 433, 1
- Begum, A., Stanimirovic, S., Goss, W. M., et al. 2010, *ApJ*, 725, 1779
- Chandrasekhar, S., & Münch, G. 1952, *ApJ*, 115, 103
- Clark, B. G. 1965, *ApJ*, 142, 1398
- Davis, R. J. 1957, *ApJ*, 125, 391
- Dickey, J. M., & Lockman, F. J. 1990, *ARA&A*, 28, 215
- Dickey, J. M., McClure-Griffiths, N. M., Gaensler, B. M., & Green, A. J. 2003, *ApJ*, 585, 801
- Dieter, N. H. 1964, *AJ*, 69, 288
- Dieter, N. H. 1965, *AJ*, 70, 552
- Field, G. B. 1965, *ApJ*, 142, 531
- Field, G. B., Goldsmith, D. W., & Habing, H. J. 1969, *ApJ*, 155, L149
- Haud, U. 2000, *A&A*, 364, 83
- Haud, U., & Kalberla, P. M. W. 2007, *A&A*, 466, 555
- Heeschen, D. S. 1955, *ApJ*, 121, 569
- Heiles, C., & Troland, T. H. 2003a, *ApJS*, 145, 329
- Heiles, C., & Troland, T. H. 2003b, *ApJ*, 586, 1067
- Hennebelle, P., & Péroul, M. 1999, *A&A*, 351, 309
- Hennebelle, P., Banerjee, R., Vázquez-Semadeni, E., Klessen, R. S., & Audit, E. 2008, *A&A*, 486, L43
- Kalberla, P. M. W., & Haud, U. 2018, *A&A*, 619, A58
- Kalberla, P. M. W., Burton, W. B., Hartmann, D., et al. 2005, *A&A*, 440, 775
- Kanekar, N., Subrahmanyan, R., Chengalur, J. N., & Sáfouris, V. 2003, *MNRAS*, 346, L57
- Koyama, H., & Inutsuka, S.-i. 2002, *ApJ*, 564, L97
- Lee, M.-Y., Stanimirovic, S., Murray, C. E., Heiles, C., & Miller, J. 2015, *ApJ*, 809, 56
- Lindblad, P. O. 1966, *BAN Suppl.*, 1, 177
- Lindner, R. R., Vera-Ciro, C., Murray, C. E., et al. 2015, *AJ*, 149, 138
- Martin, P. G., Blagrove, K. P. M., Lockman, F. J., et al. 2015, *ApJ*, 809, 153
- Matthews, T. A. 1957, *AJ*, 62, 25
- McClure-Griffiths, N. M., Pisano, D. J., Calabretta, M. R., et al. 2009, *ApJS*, 181, 398

- Mebold, U. 1972, *A&A*, **19**, 13
- Miville-Deschênes, M.-A., & Martin, P. G. 2007, *A&A*, **469**, 189
- Miville-Deschênes, M.-A., Levrier, F., & Falgarone, E. 2003, *ApJ*, **593**, 831
- Miville-Deschênes, M.-A., Murray, N., & Lee, E. J. 2017a, *ApJ*, **834**, 57
- Miville-Deschênes, M.-A., Salomé, Q., Martin, P. G., et al. 2017b, *A&A*, **599**, A109
- Muller, C. A. 1957, *ApJ*, **125**, 830
- Muller, C. A. 1959, *IAU Symp.*, **9**, 360
- Murray, C. E., Lindner, R. R., Stanimirović, S., et al. 2014, *ApJ*, **781**, L41
- Murray, C. E., Stanimirović, S., Goss, W. M., et al. 2015, *ApJ*, **804**, 89
- Murray, C. E., Stanimirović, S., Goss, W. M., et al. 2018a, *ApJS*, **238**, 14
- Murray, C. E., Peek, J. E. G., Lee, M.-Y., & Stanimirović, S. 2018b, *ApJ*, **862**, 131
- Ostriker, E. C., McKee, C. F., & Leroy, A. K. 2010, *ApJ*, **721**, 975
- Peek, J. E. G., Babler, B. L., Zheng, Y., et al. 2018, *ApJS*, **234**, 2
- Poppel, W. G. L., Marronetti, P., & Benaglia, P. 1994, *A&A*, **287**, 601
- Roy, N., Kanekar, N., Braun, R., & Chengalur, J. N. 2013a, *MNRAS*, **436**, 2352
- Roy, N., Kanekar, N., & Chengalur, J. N. 2013b, *MNRAS*, **436**, 2366
- Saury, E., Miville-Deschênes, M.-A., Hennebelle, P., Audit, E., & Schmidt, W. 2014, *A&A*, **567**, A16
- Stanimirović, S., & Heiles, C. 2005, *ApJ*, **631**, 371
- Stanimirović, S., Murray, C. E., Lee, M.-Y., Heiles, C., & Miller, J. 2014, *ApJ*, **793**, 132
- Stil, J. M., Taylor, A. R., Dickey, J. M., et al. 2006, *AJ*, **132**, 1158
- Takakubo, K. 1967, *BAN*, **19**, 125
- Takakubo, K., & van Woerden, H. 1966, *BAN*, **18**, 488
- Taylor, A. R., Gibson, S. J., Peracaula, M., et al. 2003, *AJ*, **125**, 3145
- Verschuur, G. L. 2004, *AJ*, **127**, 394
- Verschuur, G. L., & Schmelz, J. T. 1989, *AJ*, **98**, 267
- Verschuur, G. L., & Magnani, L. 1994, *AJ*, **107**, 287
- von Hoerner, S. 1951, *Zeitschrift für Astrophysik*, **30**, 17
- von Weizsäcker, C. F. 1951, *ApJ*, **114**, 165
- Winkel, B., Kerp, J., Flöer, L., et al. 2016, *A&A*, **585**, A41
- Wolfire, M. G., Hollenbach, D., McKee, C. F., Tielens, A. G. G. M., & Bakes, E. L. O. 1995, *ApJ*, **443**, 152
- Wolfire, M. G., McKee, C. F., Hollenbach, D., & Tielens, A. G. G. M. 2003, *ApJ*, **587**, 278
- Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. 1997, *ACM Trans. Math. Softw.*, **23**, 550

## Appendix A: Optimization algorithm

Terms used to compute the gradient  $\nabla J(\boldsymbol{\theta}, \mathbf{m})$  of the cost function  $J(\boldsymbol{\theta}, \mathbf{m})$  are detailed in this appendix. Incorporated are the Jacobian of the residual  $\nabla L(\boldsymbol{\theta})$  and the gradients of the regularization term  $\nabla R^i(\boldsymbol{\theta}, \mathbf{m}) = [\nabla_{\boldsymbol{\theta}} R^i(\boldsymbol{\theta}, \mathbf{m}), \nabla_{\mathbf{m}} R^i(\boldsymbol{\theta}, \mathbf{m})]$

$$\nabla L(v_z, \boldsymbol{\theta}(r)) = \begin{bmatrix} \exp\left(-\frac{(v_z - \mu_1(r))^2}{2\sigma_1(r)^2}\right) \\ \frac{a_1(v_z - \mu_1(r))}{\sigma_1^2} \exp\left(-\frac{(v_z - \mu_1(r))^2}{2\sigma_1(r)^2}\right) \\ \frac{a_1(v_z - \mu_1(r))^2}{\sigma_1^3} \exp\left(-\frac{(v_z - \mu_1(r))^2}{2\sigma_1(r)^2}\right) \\ \vdots \\ \frac{a_N(v_z - \mu_N(r))^2}{\sigma_N^3} \exp\left(-\frac{(v_z - \mu_N(r))^2}{2\sigma_1(r)^2}\right) \end{bmatrix} \quad (\text{A.1})$$

$$\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \mathbf{m}) = \begin{bmatrix} \lambda_a \mathbf{D}^t \mathbf{D} a_1 \\ \lambda_{\mu} \mathbf{D}^t \mathbf{D} \mu_1 \\ \lambda_{\sigma} \mathbf{D}^t \mathbf{D} \sigma_1 \\ \lambda'_{\sigma}(\sigma_1 - m_1) \\ \vdots \\ \lambda'_{\sigma}(\sigma_N - m_N) \end{bmatrix} \quad (\text{A.2})$$

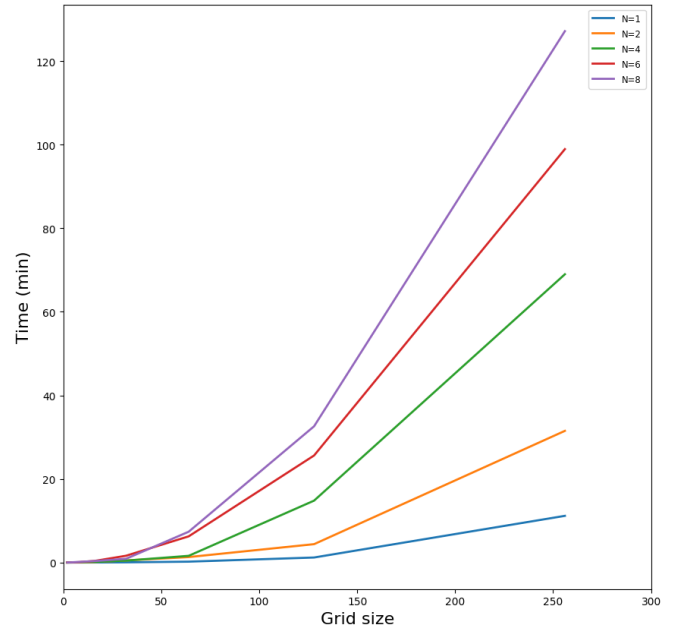
$$\nabla_{\mathbf{m}} R(\boldsymbol{\theta}, \mathbf{m}) = \begin{bmatrix} -\sum_r \lambda'_{\sigma}(\sigma_1 - m_1) \\ \vdots \\ -\sum_r \lambda'_{\sigma}(\sigma_N - m_N) \end{bmatrix}. \quad (\text{A.3})$$

## Appendix B: Computation time

The computation time needed to perform the Gaussian decomposition of the synthetic PPV cube presented in Sect. 3 is described in Fig. B.1. The computation time depends on the number of Gaussian components  $N$ , the size of the cube (number of spectra and number of velocity channels), and the maximum number of iterations used in the optimization.

For a given  $N$ , the computation time scales linearly with the number of spectra and the number of velocity channels. Therefore, as for each step of the multi-resolution process, from coarse to fine grid, the size of the grid is multiplied by a factor of four; the computation time also increases by a factor four at each step. We note that each step has the same maximum number of iterations (here 800) that also linearly impact the computation time.

Finally, we observed that the computation time depends nonlinearly on  $N$ , as seen in Fig. B.1.



**Fig. B.1.** Computation time used by one CPU to perform the Gaussian decomposition of the simulated PPV cube used in Sect. 3, for  $N=1, 2, 4, 6,$  and  $8$ , as function of the size grid. The maximum number of iterations in each case has been set to 800.