The Astrophysical Journal, 889:22 (15pp), 2020 January 20 https://doi.org/10.3847/1538-4357/ab5524

Characterizing Magnetic Reconnection Regions Using Gaussian Mixture Models on Particle Velocity Distributions

Romain Dupuis[1], Martin V. Goldman[2], David L. Newman[2], Jorge Amaya[1], and Giovanni Lapenta[1]
[1] Center for Mathematical Plasma Astrophysics, KU Leuven, Celestijnenlaan 200B, bus 2400, B-3001 Leuven, Belgium; romain.dupuis@kuleuven.be, dupuis.ro@gmail.com
[2] University of Colorado, Boulder, CO 80309, USA
Received 2019 September 20; revised 2019 October 21; accepted 2019 November 5; published 2020 January 21

Abstract

We present a method based on unsupervised machine learning to identify and characterize regions of interest using particle velocity distributions as a signature pattern. An automatic density estimation technique is applied to particle distributions provided by particle-in-cell simulations to study magnetic reconnection regions. Its application to magnetic reconnection is new. The key components of the method involve (i) a Gaussian mixture model determining the presence of a given number of subpopulations within an overall population, and (ii) a model selection technique with a Bayesian information criterion to estimate the appropriate number of subpopulations. Thus, this method automatically identifies the presence of complex distributions, such as beams or other non-Maxwellian features, and can be used as a detection algorithm able to identify reconnection regions. The approach is demonstrated for a specific double Harris sheet simulation, but it can in principle be applied to any other type of simulation data on the particle distribution function.

Unified Astronomy Thesaurus concepts: Solar magnetic reconnection (1504); Gaussian mixture model (1937); Plasma astrophysics (1261)

1. Introduction

Plasmas are host to a complex mixture of interacting processes. Determining what process happens in a given location is often a challenge. When plasmas are modeled with a kinetic description, the researcher is confronted with a 6D data set. In particle-in-cell (PIC) methods, this information comes as a collection of hundreds or thousands of particles per cell with a total data size nowadays reaching a totality close to trillions of particles. Often the analysis focuses only on the electromagnetic fields and the moments of the particle distribution that are 3D manageable data sets. However, this leaves out the richest part of the simulation information: the particles. We consider here an approach based on the Gaussian mixture model (GMM) to extract information automatically from the particle distribution without requiring human intervention.

To demonstrate the approach, we consider the often-studied case of a plasma undergoing magnetic reconnection. Magnetic reconnection (Gonzalez & Parker 2016) plays a crucial role in collisionless plasmas. By breaking down the frozen-in magnetic fields, the magnetic field energy is converted into kinetic energy, thermal energy, and particle acceleration energy. This process appears fundamental in the transport mechanism, and it represents one of the most important sources of particle acceleration in space. Magnetic reconnection can occur at various scales and locations, such as in laboratory plasma (Yamada et al. 2014), in plasma turbulence (Haynes et al. 2014), or in the magnetotail (Eastwood et al. 2013). Therefore, magnetic reconnection has been studied in many ways, including in situ measurements with the *Magnetospheric Multiscale* (*MMS*) mission (Burch et al. 2016a) and numerical simulations (Hesse et al. 2014; Goldman et al. 2016).

In this paper, we are interested in automatically characterizing reconnection regions from particle distributions. We investigate this in a well-documented case, that of 2D collisionless PIC simulations. The literature of the last 20 yr has numerous examples of variations of 2D reconnection setups, from simple Harris initializations (Birn et al. 2001) to more realistic equilibria (Sitnov et al. 2013) based on analytical models (Lembege & Pellat 1982) or on global MHD simulations (Ashour-Abdalla et al. 2015). In this context, we can more readily interpret the result of the new diagnostic presented here using the knowledge base of 2D PIC simulations accrued in the recent past. Future work will then expand the use of the new diagnostics to less documented cases such as turbulent reconnection in 2D and 3D, and to distributions obtained from in situ space missions.

The particle-scale kinetic physics, and in particular the electron-scale one, has recently received renewed interest in the context of magnetic reconnection, thanks to the *MMS* mission. In particular, electron distributions have been shown to be good indicators for magnetic reconnection. For instance, exhaust electrons can give rise to highly structured anisotropy when the reconnection rate achieves its maximum (Shuster et al. 2014). Crescent-shaped distributions can be detected near the electron stagnation point for asymmetric reconnection (Burch et al. 2016b) as indication of the presence of meandering orbits (Bessho et al. 2016). Meandering orbits and crescents are observed also in other regions around reconnection regions, such as in the proximity of the separatrix of asymmetric reconnection but also in symmetric reconnection (Egedal et al. 2016; Lapenta et al. 2017). Triangular shapes have also been observed in the vicinity of the X line within the electron diffusion region (EDR) for weak guide fields (Shuster et al. 2015). In the presence of magnetic islands, specific distributions can be present for each region, such as flat-top or crescent-shaped distributions (Cazzola et al. 2016). Thus, while distributions provide richer insight into the local physics than the local fields and moments, it seems clear that a unique specific distribution cannot be used as a signature for reconnection as it does not reflect the phenomenon for all possible external conditions. For this reason, developing a detection algorithm based on machine learning techniques and able to detect non-Maxwellian features is especially desirable. Such methods can detect complex shapes

from the analyses of electron velocity distributions. Moreover, they could be coupled with other more classical detection methods based on field quantities, such as agyrotropy (Aunai et al. 2013).

Machine learning is increasingly being used in various fields related to physics, such as particle physics (Radovic et al. 2018) or space and plasma science (Camporeale et al. 2018). In particular, supervised learning algorithms try to find the relationship between some input features and labeled output. Such approaches would need a database of magnetic-reconnection-related distribution functions to be efficient. Building such a database may be inconvenient and human labor intensive, therefore other approaches can be considered. Unsupervised learning, another type of machine learning technique, extracts hidden structures and patterns from data without any corresponding prelabeled target values. There is no longer mapping between inputs and outputs, as it was the case for supervised learning. Unsupervised learning encompasses mainly dimensionality reduction, clustering, generative modeling, and density estimation. Such techniques have been used very little in space weather and plasma physics (del Castillo-Negrete et al. 2010; Heidrich-Meisner & Wimmer-Schweingruber 2018). We are particularly interested in density estimation techniques, approximating the probability density function from the data (Bishop 2006). They are especially promising in identifying specific physical regimes, such as magnetic reconnection. In particle physics, authors have proposed using density estimation to detect the presence of new physics events in the data (Albertsson et al. 2018). It gave us the idea of applying such techniques on reconnection simulations with the goal of detecting specific particle distributions, such as beams or non-Maxwellian distributions, which could be used as magnetic reconnection signatures. Therefore, the method proposed in this paper contributes to the improvement of the potential connections between machine learning and plasma physics in general by providing a relevant illustration of magnetic characterization.

This article is organized as follows: Section 2 introduces several classical reconnection signatures, Section 3 gives details on various velocity distributions related to reconnection, Section 4 presents the strategy for identifying reconnection regions, Section 5 describes the PIC simulations, and the results are discussed in Section 6.

2. Identifying Reconnection Signatures

As a specific example for untrained automatic detection of features in a plasma, we select the process of magnetic reconnection. Magnetic reconnection is notoriously difficult to detect in general 3D geometries (Biskamp 2000; Priest & Forbes 2007).

In 2D configurations, reconnection is associated with the presence of an EDR, which modifies the magnetic field, due to the generation of dissipative electric fields (Birn & Priest 2007; Hesse et al. 2011; Burch et al. 2016b). As this region is very small and localized, its precise detection is hard, especially for spacecraft measurements. In 3D configurations, the problem is even more challenging. General mathematical arguments have been proposed, based on breaking the frozen-in condition (Hesse & Schindler 1988) and on topological concepts such as null points (Lau & Finn 1990; Priest & Titov 1996), the magnetic skeleton (Haynes et al. 2007), and the so-called squashing factor (Titov 2007).

A wide range of signatures has been highlighted in the literature, using different sources of data. We refer the readers to a recent review of all the signatures of reconnection proposed in recent years (Goldman et al. 2016). These measures can be organized in two groups. The first is based on field quantities: for example, the detection of magnetic nulls (Fu et al. 2015) and magnetic skeletons (Haynes et al. 2007) or the explicit violation of magnetic flux conservation (Newcomb 1958; Vasyliunas 1975; Hesse & Schindler 1988). The second category uses the moments of the plasma species: for example, the relative drift between the plasma and the field lines (slippage) or the energy dissipation measured on the electron frame (Zenitani et al. 2011). As a specific example of this category, we will use here the second-order moment—the pressure tensor—proposed to identify reconnection sites (Scudder & Daughton 2008). Indeed, nongyrotropic velocity distributions have been recently shown by direct in situ measurements (Burch et al. 2016b) to play a key role during the magnetic reconnection process. Usually, the different methods quantify the deviations from symmetry for the pressure tensor. However, this concept has led to different definitions (Scudder & Daughton 2008; Aunai et al. 2013; Swisdak 2016) with different domains of application. In the present paper, we rely specifically on the measure of gyrotropy called Q (Swisdak 2016):

$$Q = \frac{P_{12}^2 + P_{13}^2 + P_{23}^2}{P_{\parallel} + 2P_{\perp}}, \tag{1}$$

where $P_{\parallel}$ and $P_{\perp}$ are the diagonal terms of the tensor and P_{12}, P_{13}, and P_{23} are the sub and upper diagonal terms of the symmetric tensor. The measure of gyrotropy is equal to 0 for gyrotropic tensors while Q is equal to 1 for maximal deviations. A key aspect of the measure, relevant for our present study, is that it is a scalar measure of the complexity of the particle velocity distribution: the pressure tensor is the second-order moment of this distribution and by measuring its deviation from a simple isotropic Maxwellian, Q indirectly measures the complexity of the distribution. For example, in the EDR, the measure of gyrotropy is linked to the presence of a complex velocity distribution that includes a main low-energy and isotropic population and a minority higher energy population distributed in the perpendicular velocity plane as a crescent (Bessho et al. 2016; Burch et al. 2016b). The measure of gyrotropy is large in regions where distributions present such complexities as a crescent. A more detailed view of the complexity of the velocity distribution function can be obtained looking at moments higher than two, for example using an expansion in Hermite polynomials (Loureiro et al. 2016; Servidio et al. 2017; Meyrand et al. 2019).

In the present work, we present a third new approach to detect reconnection: directly using the distribution function and measuring its complexity via the GMM, an unsupervised machine learning approach largely used in other fields (Bishop 2006).

3. Fitting Particle Velocity Distributions

Fitting distributions to a sample of data is the process of choosing a probability distribution modeling a data set and estimating the associated parameters. The selection of a correct distribution function must take into account various parameters involving mathematical and physical arguments. Is the distribution unimodal or multimodal? Does the phenomena show symmetric or skewed behavior? Can we derive specific bounds

for the distributions? The answer to these questions will guide the choice of the model.

Various distributions have been used to fit plasma particle velocity distributions. The most common model in space plasmas is the Maxwellian and the bi-Maxwellian, taking into account temperature anisotropy. The classical Maxwellian distribution shows good results by describing velocity distribution in low-energy regions, in particular for ions (Gruntman 1992; Kasper et al. 2006). However, the plasma velocity distributions exhibit non-Maxwellianities for suprathermal regions, where the distribution is governed rather by power-law tails. Thus, the Kappa distribution (also called generalized Lorentzian) has been proposed to describe both low-energy Maxwellian cores and suprathermal tails (Vasyliunas 1968; Summers & Thorne 1991). Kappa distributions have gained an important notoriety in numerous studies in space plasmas (Hellberg & Mace 2002; Pierrard & Lazar 2010; Livadiotis & McComas 2013; Ogasawara et al. 2013; Lazar et al. 2018; Livadiotis et al. 2018). One can note that when the spectral index kappa increases toward infinity, the Kappa distribution tends to a Maxwellian.

Previous studies have shown examples of electron velocity distribution fitting using a 1D cut (Pulupa et al. 2014) or 2D distributions (Wilson et al. 2019) for the solar wind. In the latter paper, the best approximation is built as a sum of three densities for the cold dense core, the hot halo, and the beam. Each component is fit by choosing among a list of potential distributions and optimizing the associated parameters. This kind of approach allows a physical interpretation to the main distributions of each component to be provided. However, it relies on a strong physical knowledge, such as the list of potential distributions or the range of variations for all distribution parameters. Such detailed knowledge is not necessarily available or reliable for all physical phenomena or locations. Souza et al. (2018) used an automatic clustering method, called a self-organizing map, to organize pitch-angle-resolved particle flux data collected in the outer Van Allen belt region into different categories. With regard to reconnection, we previously described that various distribution shapes could be observed near reconnection sites, such as crescent shapes or triangles, and their presence may depend on various conditions. Their discovery may sound recent, and other distributions may exist but have not yet been discovered. An ideal algorithm must therefore not rely on a specific set of distributions for the detection of reconnection.

Density estimation techniques aim to build a model of a non-observable probability density function by observing a set of data points. They appear therefore as a potential candidate to automatically fit complex distribution functions. We expect to identify particle distributions with specific shapes, such as beams or non-Maxwellian features in order to relate them to reconnection sites. A growing interest for such methods has been observed in astronomy (Ivezić et al. 2014). There are two main kinds of density estimations: parametric and nonparametric methods. The first one represents the natural approach, where the distribution is estimated by fitting the parameters of a given model to the data. For instance, a Gaussian distribution can be locally approximated by a second-order polynomial. Ni et al. (2015) fit electron pitch-angle distributions using $\sin^N(\alpha)$ functions, where α is the local particle pitch angle and N the power law. A very popular method, called GMM, fits the data with a sum of Gaussian distributions (Bishop 2006). On the other hand, nonparametric methods try to make as few assumptions as possible, mainly by working with infinite-dimensional models. One of the simpler nonparametric density estimators is the histogram which splits the support of the distribution into bins and then the value of the function is defined as the number of samples falling into that bin. A very popular method in machine learning, called kernel density estimation (KDE), proposes a more general approach by convolving the data with a smooth kernel function (Sheather 2004). However, for the two methods, a specific issue arises as the width of the kernel (or the size of the bin) must be chosen. If this value is too small, a noisy function is observed as randomness in the signal is highlighted. If the value is too large, modes are smoothed out and important structures are obscured. Several strategies have been proposed to determine this parameter, such as cross-validation or plug-in methods (Heidenreich et al. 2013).

4. Detection Algorithm

In this section, a detection algorithm based on the GMM (McLachlan & Peel 2004) is presented. Parametric methods were preferred over nonparametric ones as they provide an easier interpretability. After introducing the main mathematical derivations of the GMM, the selection of the number of components is detailed, and two specific metrics based on thermal energy are defined.

4.1. Density Estimation with GMMs

The mixture model is defined as a weighted sum of given densities with unknown parameters. The most common density is the Gaussian density as it ensures a closed formalism for the determination of the parameters and limits the computation to the means and the covariances. Moreover, the second reason is that Gaussian density can be considered as a reasonable assumption for the density when no prior information is available for the probability density function. A general technique for finding the unknown parameters consists in maximizing the likelihood function with the expectation-maximization (EM) algorithm. This approach is detailed below.

With regard to the mathematical formalism, the random variable $\boldsymbol{x}$ associated with the observations is assumed to be written as a linear superposition of K multivariate Gaussians weighted by the mixing coefficients w_k:

$$p(\boldsymbol{x}|\Phi) = \sum_{k=1}^{K} w_k \, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_k}). \tag{2}$$

The normal distribution $\mathcal{N}$ is parameterized by the mean $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$ as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right]. \tag{3}$$

All of the parameters of the GMM are regrouped in the mixture parameter $\boldsymbol{\Phi} = [w_1, \cdots, w_q, \boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_K}]$. The Python package Scikit-learn is used to perform all of the computations (Pedregosa et al. 2011).

4.1.1. Unobserved Latent Variables

A K-dimensional binary random latent variable $\boldsymbol{z} \in \mathbb{R}^K$ is introduced, such as a particular component z_k of $\boldsymbol{z}$ is equal to 1 and all other elements are equal to 0, meaning that z_k satisfies $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$. In particular, the kth component is 1 if the observation of $\boldsymbol{x}$ is generated from the kth Gaussian such that the marginal distribution over $\boldsymbol{z}$ is directly related to the mixture proportion as $p(z_k = 1) = w_k$.

We want to rewrite the definition of the Gaussian mixture in Equation (2) by introducing the unobserved latent variable $\boldsymbol{z}$. The distribution of this latter can be expressed as

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} w_k^{z_k}. \tag{4}$$

Moreover, the conditional distribution of $\boldsymbol{x}$ given the latent variable $\boldsymbol{z}$ is straightforward:

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_k})^{z_k}. \tag{5}$$

The joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$ is expressed with the product rule:

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} w_k^{z_k} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_k})^{z_k}. \tag{6}$$

Finally, the marginal distribution is integrated over $\boldsymbol{z}$, and we find the same expression as in Equation (2):

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_k})^{z_k}. \tag{7}$$

This new expression of the Gaussian mixture now involves the latent variable $\boldsymbol{z}$ and will be very useful in computing all the parameters of the mixture as we can now work with the joint distribution function instead of the marginal distribution. Moreover, each observation $\boldsymbol{x_i}$ is now associated with a specific value of the latent variable z_i.

Let us now introduce the condition probability of $\boldsymbol{z}$ given the observation $\boldsymbol{x}$ and called γ:

$$\begin{aligned} \gamma(z_k) \coloneqq p(z_k = 1|\boldsymbol{x}) &= \frac{p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}{p(\boldsymbol{x})} \\ &= \frac{w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_k})}{\sum_{j=1}^{K} w_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta_j})}. \end{aligned} \tag{8}$$

It can also be viewed as the responsibility that the component k takes for "explaining" the observation $\boldsymbol{x}$ (Bishop 2006).

4.1.2. Maximum Likelihood with EM Algorithm

Let assume we have a set of observations $\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}$ from the variable $\boldsymbol{x}$. They form the matrix of the training set $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. The same procedure is applied to build the matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ associated to the latent variable values. The log likelihood l of the Gaussian mixture can be expressed from the observations as

$$l(\phi|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x_i}|\boldsymbol{\theta_k}) \right]. \tag{9}$$

Maximizing this expression appears to be a complex problem, due to the summation inside the logarithm. Two main approaches exist to solve this maximization problem: classical gradient-descent or EM algorithm (Bishop 2006). We will focus on the latter.

The mixture parameters regrouped in $\boldsymbol{\Phi}$ are estimated iteratively using an EM algorithm (Dempster et al. 1977). Let us set the gradient of the likelihood expression in Equation (9) to zero with respect to (i) the mean μ_k, (ii) the covariance matrix Σ_k, and (iii) the mixture proportion w_k (coupled with a Lagrange multiplier to take into account the constraint $\sum_{k=1}^{K} z_k = 1$). After several derivations detailed in Bishop (2006) and by writing $\gamma(z_{ik})$, the specific value of the responsibility for a given observation $\boldsymbol{x_i}$, we end up with three expressions:

$$\boldsymbol{\mu_k} = \frac{\sum_{i=1}^{n} \gamma(z_{ik}) \boldsymbol{x_i}}{\sum_{i=1}^{n} \gamma(z_{ik})}, \forall k \in [1, \cdots, K], \tag{10}$$

$$\boldsymbol{\Sigma_k} = \frac{\sum_{i=1}^{N} \gamma(z_{ik})(\boldsymbol{x_i} - \boldsymbol{\mu_k})(\boldsymbol{x_i} - \boldsymbol{\mu_k})^T}{\sum_{i=1}^{N} \gamma(z_{ik})}, \forall k \in [1, \cdots, K], \tag{11}$$

$$w_k = \frac{1}{n} \sum_{i=1}^{n} \gamma(z_{ik}), \forall k \in [1, \cdots, K]. \tag{12}$$

However, these three expressions do not provide a closed-form solution for the parameters of the mixture, due to the complex relationship between them and γ expressed in Equation (8) (Bishop 2006). For this reason, a simple iterative scheme for finding a solution to the maximum likelihood problem (EM algorithm) is used for this particular case of the GMM. The EM algorithm is split into two steps: the expectation step (also called the E step) and the maximization step (also called the M step).

In the E step, the posterior probability of $\boldsymbol{z}$ (the responsibility) is computed from the current value of the parameters. Then, in the M step, all of the parameters are re-estimated by using the previously computed posterior probability. The algorithm can be written with the following steps (Bishop 2006):

1. Initialize the parameters of the mixture $\boldsymbol{\Phi}$: means $\boldsymbol{\mu_k}$, covariances $\boldsymbol{\Sigma_k}$, and mixing coefficients w_k.
2. E step: compute the responsibility $\gamma(z_{ik})$ from Equation (8).
3. M step: re-estimate the parameters $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$, and w_k from, respectively, Equations (10)–(12) using the current responsibilities $\gamma(z_{ik})$.
4. Check the convergence. If not, return to 1.

One can note that a sample weight q_i can also be associated to each sample $\boldsymbol{x_i}$. In the case detailed above, all the weights are equal to 1. The modified algorithm taking into account sample weights changes the EM algorithm and the computation of the log likelihood. The new responsibility γ can be rewritten as (Colomé et al. 2017):

$$\gamma(z_{ik}) \coloneqq q_i \gamma(z_{ik}).$$

4.2. Model Selection

The number of Gaussians K usually acts as an input to the GMM algorithm. This value may be specified by the user at the beginning of the algorithm, or it can be estimated by analyzing the data. Several methods are proposed in the literature: cross-validation, elbow method, information criterion, etc. (Bishop 2006). The general idea is to define an estimator related to the relative quality of the Gaussian mixture for a given set of data. Information criteria represent good candidates as they give a trade-off between the goodness of fit and the complexity of the model. The two main estimators are called the Akaike information criterion (AIC) and Bayesian information criterion

(BIC; Anderson 2002):

$$\begin{aligned} \mathrm{AIC} &= 2k - 2\ln(L) \\ \mathrm{BIC} &= \ln(n)k - 2\ln(L), \end{aligned} \tag{13}$$

where k is the number of parameters to estimate in the model and L the likelihood. In cases of weighted particles, the number of particles n corresponds to the weighted number of particles. BIC penalizes the model complexity more than AIC. However, AIC and BIC performances depend on the nature of the data generating the model: sample size, complexity of the model, whether the true model is contained in the model set or not, etc. (Anderson 2002). As data from simulations may be noisy and the number of particles is significant, BIC has been preferred in this work to automatically select the number of components of the mixture. Special attention should be paid to the number of particles n, which can be arbitrarily large for PIC simulations. On the one hand, a very small number of particles would lead to noisy distributions and a BIC parameter with a weak penalization for complex models. In this case, some components may only fit noise. On the other hand, a very large number of particles may overpenalize models with several components. From the authors' experience, typical numbers of particles between 1000 and 10,000 seem to be acceptable. It may be interesting to compare these numbers of particles with missions such as *MMS* or Cluster.

Nevertheless, the physical meaning of the number of components K and the parameters associated with each Gaussian must be analyzed carefully as they must not be necessarily interpreted as specific beams or electron populations. Indeed, if the data show complex shapes or are not near Gaussian, the number of components K does not correspond to the number of different populations (Ivezić et al. 2014). For instance, a flat-top distribution is approximated by several Gaussians, but each component is needed to approach the broad mode of the distribution. A Kappa distribution can also be represented by a central Gaussian centered around the mode plus another Gaussian with a very large width to fit the wide tail, thus two Gaussians are needed for a single population. Moreover, as presented previously, BIC is sensitive to various parameters: the data themselves and the sample size. For instance, if the source of the data does not change but the number of samples increased, the resulting number of components may also change. However, BIC is still an efficient criterion for providing a statistical analysis based on the underlying properties of the data. It can help detect important variations in the distribution. Another strategy consists of fixing the number of components to a high value in order to improve the fit for very complex distributions, which can show poor results for a small number of components. In this case, GMM is very close to a nonparametric density estimation method, such as KDE. Such strategy is illustrated in Appendix C.

4.3. Thermal Energy Variation

As the particle distributions are approximated by sums of Gaussians instead of a single Maxwellian, it is interesting to analyze the variation of the thermal velocity for these two representations. The thermal energy for a single velocity distribution is given by its variance. The straight measure of thermal energy based on the moment of the whole distribution is

$$E_{\mathrm{thermal}} = \frac{1}{N_p}\sum_{i=1}^{3}\left[\sum_{p}(\boldsymbol{V_p} - \langle \boldsymbol{V_p}\rangle)^2\right]_i, \text{ with } \langle \boldsymbol{V_p}\rangle = \sum_{p}\frac{\boldsymbol{V_p}}{N_p}. \tag{14}$$

The variance $(\sigma^2)^{(K)}$ for K multiple Maxwellians is given by

$$(\sigma^2)^{(K)} = \sum_{i=1}^{3}\left[\sum_{k=1}^{K} w_k^2(\boldsymbol{\sigma_k})^2 + \sum_{k=1}^{K} w_k(\boldsymbol{\mu_k})^2 - \left(\sum_{k=1}^{K} w_k(\boldsymbol{\mu_k})\right)^2\right]_i. \tag{15}$$

The first term can be interpreted as the mixture of the variances and is related to the thermal energy per unit mass of the mixture. Therefore, it is written as the thermal energy (per unit mass) of the K multiple Maxwellians:

$$E_{\mathrm{thermal}}^{(K)} = \frac{1}{2}\sum_{i=1}^{3}\sum_{k=1}^{K} w_k^2[\boldsymbol{\sigma_k^2}]_i. \tag{16}$$

The thermal energy ratio E_{drop} is derived to compute the reduction in thermal speed for the particles, aiming to distinguish heating from accelerating particles into beams. It measures the ratio between the mixture of the variance and the variance of the velocity distribution:

$$E_{\mathrm{drop}} = \frac{E_{\mathrm{thermal}}^{(K)}}{E_{\mathrm{thermal}}}. \tag{17}$$

This metric is defined to always be below 1. Low values indicate that the thermal energy of the mixture is much smaller than the thermal velocity computed directly from the definition, suggesting that the second-order moment of the overall distribution is not a good indicator of the conditions present. An extreme example is that of two cold beams which individually have zero thermal spread and only a relative mean velocity but when taken together appear as a broad thermal spread. This measure identifies these conditions, spotting distributions characterized by interpenetrating beams.

The last two terms of Equation (15) can be read as the deviation of each mean compared to the overall mixture mean:

$$E_{\mathrm{dev}}^{(K)} = \sum_{i=1}^{3}\left[\sum_{k=1}^{K} w_k(\boldsymbol{\mu_k})^2 - \left(\sum_{k=1}^{K} w_k(\boldsymbol{\mu_k})\right)^2\right]_i. \tag{18}$$

This deviation is always positive as it corresponds to a weighted variance. This is the thermal energy of the center of all beams, measuring the distance between them. A second metric E_{dev}, called the thermal velocity deviation, defines the ratio between the velocity deviation for the mixture and the classical thermal velocity of the distribution:

$$E_{\mathrm{dev}} = \frac{E_{\mathrm{dev}}^{(K)}}{E_{\mathrm{thermal}}}. \tag{19}$$

This strictly positive quantity allows the different mixtures to be interpreted. High values mean the components are widely separated and presumably have a distinct identity and perhaps origin (Eastwood et al. 2015). Small values point to mixtures of components close to each other and perhaps carry less meaningful separation.

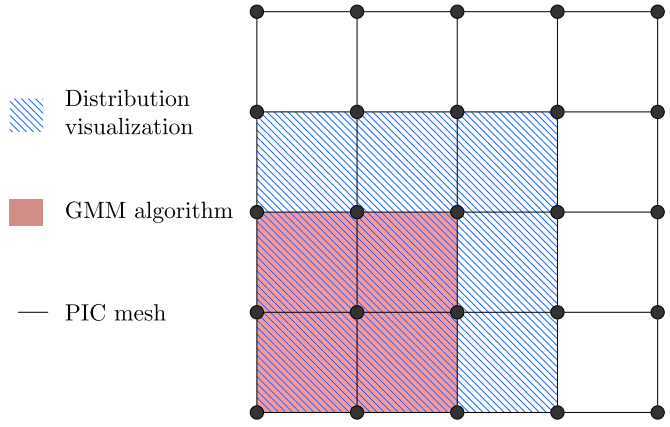

Figure 1. Illustration of the different levels of granularity between the PIC cells, the GMM algorithm cells, and the distribution visualization regions. They may be at different levels due to the few number of particles.

4.4. General Procedure

The detection algorithm has been performed on a square window with a size of r by r cells. This set of cells is merged in each direction to build the computational domain. Using large windows provides enough particles to the algorithm. It ensures that the model selection by BIC is less sensitive to the number of particles and does not necessarily favor very simple models with few numbers of components with regard to the real complexity of the data. Figure 1 illustrates the different levels between this square window, the PIC mesh, and the visualization regions for the distributions. We are limiting the maximum number of components to six for the BIC optimization. The general procedure is given by the following algorithm:

Algorithm 1. Detection algorithm

Data: particle coordinates $\boldsymbol{X}$ and velocities $\boldsymbol{V}$, optional sample weights $\boldsymbol{q}$, mesh Ω with n_x cells by n_y cells
Result: distribution function f_i, E_{drop}, and E_{dev} for each subset i
begin
  merge cells to increase the number of particles by subset, new mesh Ω' is of size $\frac{n_x}{r} \times \frac{n_y}{r}$;
  for $i \in \Omega'$ **do**
    for $k = 1$ to $k = 6$ **do**
      $GMM_k = GMM(\boldsymbol{V}_i; k)$;
      $BIC_k = BIC(GMM_k)$;
    end
    $K = \arg\max_{k\in[1,6]} BIC_k$;
    $f_i = GMM_K$;
    compute E_{drop} and E_{dev};
  end
end

5. PIC Simulations

The simulations are performed with the fully kinetic massively parallel implicit moment method PIC code iPic3D (Markidis & Lapenta et al. 2010; Innocenti et al. 2017). We are particularly interested in a double Harris sheet case where two well-separated reconnection sites can be identified. Two different guide field values have been tested: 0.1 and 1.0. The simulation is 2.5D, thus all vectors are considered to be 3D, but their spatial variation is limited to a 2D plan independent of the dawn–dusk (Z) direction. All quantities are presented in normalized form.

PIC simulations of 2D Harris sheet (Harris 1962) reconnection are paradigmatic in reconnection and are therefore our choice in determining how the diagnostic presented here performs in this classic well-known problem, often also referred to as the GEM Challenge (Birn et al. 2001). We consider here specifically the double Harris sheet case defined by

$$B_x(y) = B_0(-1 + \tanh(y - y_1) - \tanh(y - y_2)), \quad (20)$$

with the location of the two current layers at $y_1 = L_y/4$ and $y_1 = 3L_y/4$ (Wu et al. 2011). Pressure balance is maintained by a uniform temperature but a nonuniform density:

$$n_s(y) = n_0(-1 + \operatorname{sech}(y - y_1)^2 + \operatorname{sech}(y - y_2)^2) + n_b, \quad (21)$$

where a background density equal to $n_b = n_0/10$ is added. The equilibrium is defined by the thickness $L/d_i = 0.5$ and with the parameters $m_i/m_e = 256$, $v_{\text{the}}/c = 0.045$, and $T_i/T_e = 5$. With these choices, the asymptotic in-plane field B_0 is set by the ratio $\omega_{ci}/\omega_{pi} = 0.0097$ and the peak Harris density $n_0 = 1$ is imposed by the normalization used that results in the ion plasma frequency and ion inertial length to be unitary. The coordinates are chosen with the initial Harris magnetic field along x with size $L_x = 30d_i$ and the initial gradients along y with $L_y = 40d_i$. The third dimension, where the initial current and guide field are directed, is invariant. Periodicity is assumed in all directions. The Cartesian mesh has a size of 769 × 1025, and about 196,600,000 particles with varying weights are injected in the computational domain, representing approximately 250 particles by cell. The particle distributions are analyzed in a frame of reference driven by the local magnetic field, in addition to the Cartesian system, as suggested in Goldman et al. (2016). The B-field-aligned basis is defined by the following three vectors:

$$\begin{aligned} \boldsymbol{e}_{\parallel} &:= \widehat{\boldsymbol{B}}, \quad \text{where } \widehat{\boldsymbol{B}} = \frac{\boldsymbol{B}}{\|\boldsymbol{B}\|} \\ \boldsymbol{e}_{\perp 1} &:= \widehat{\boldsymbol{B}} \times \boldsymbol{e}_z \\ \boldsymbol{e}_{\perp 2} &:= \widehat{\boldsymbol{B}} \times \boldsymbol{e}_{\perp 1} = -\widehat{B}^2 \boldsymbol{e}_z + (\boldsymbol{e}_z.\widehat{\boldsymbol{B}})\widehat{\boldsymbol{B}}. \end{aligned} \quad (22)$$

Therefore, $\boldsymbol{e}_{\parallel}$ is parallel to the total magnetic field, $\boldsymbol{e}_{\perp 1}$ is in the reconnection x–y plane, perpendicular to in-plane magnetic field lines, and $\boldsymbol{e}_{\perp 2}$ is in the $-z$-direction for a magnetic field with small z-component.

6. Results

The automatic detection algorithm presented above can be used to spot magnetic reconnection and regions of interest. From the point of view of plasma physics, the decomposition in several Gaussians by the GMM algorithms can be considered as reasonable only if results are in agreement with other classical methods. For this reason, outcomes of the detection algorithm are compared with measures of gyrotropy considering the guide field value of 0.1. The second case with a guide field of 1.0 is described in Appendix A. As the double Harris sheet case shows very similar behavior for the two layers, the results are only presented for the bottom layer. Moreover, as a significant number of particles is most suitable to train the density estimation model, the algorithm is performed on a coarser resolution. Each group of 4 by 4 cells ($r = 4$) are merged into a square window. Appendix B gives details of the sensitivity to this number of cells. Finally, only electron distributions with

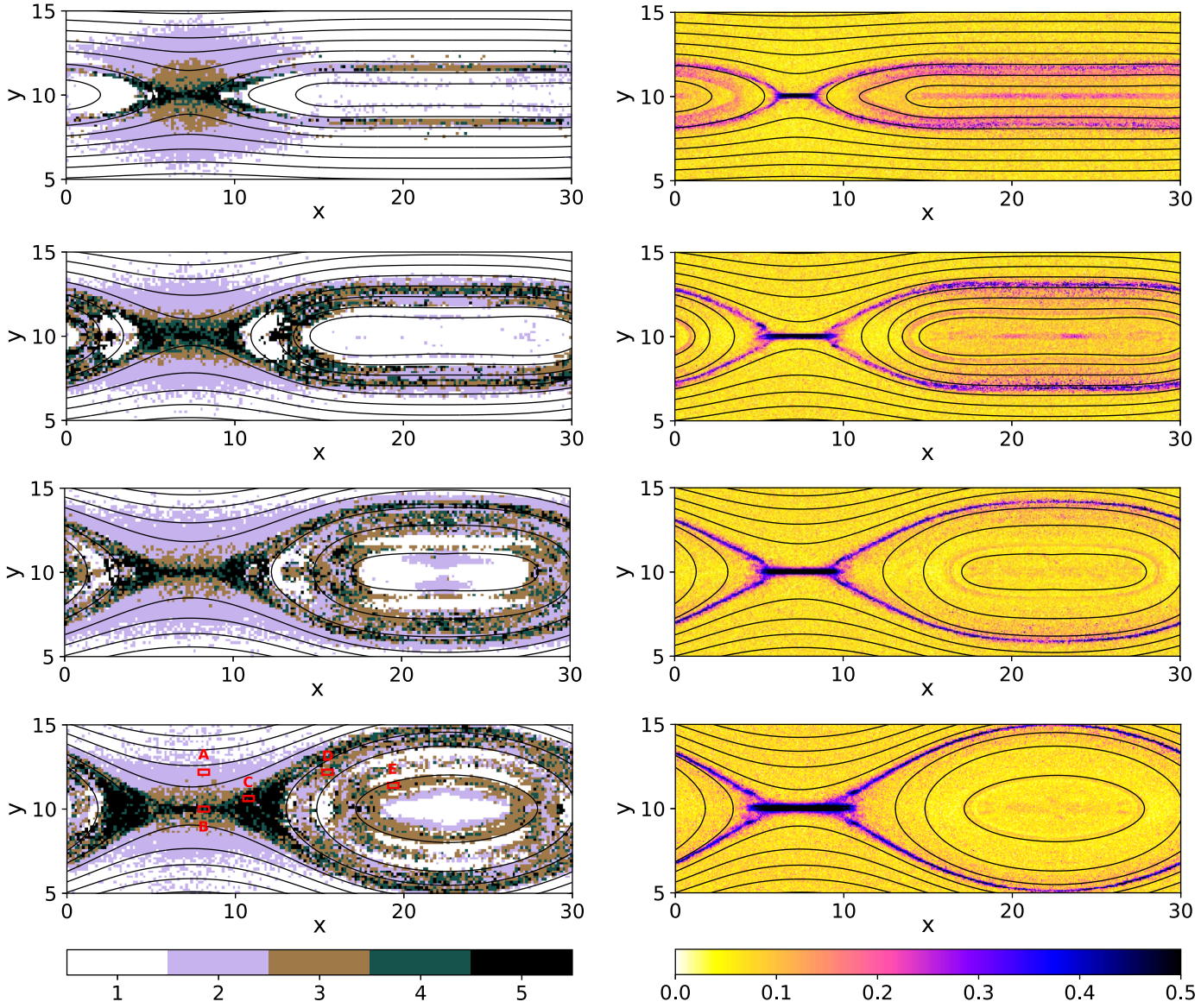

Figure 2. Magnetic reconnection detection for the double Harris sheet case with a guide field value of 0.1 at four different time steps, from top to bottom: $t = 8000$, $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$. The left column presents the number of components provided by the BIC optimization, and the right column shows the measure of gyrotropy $\sqrt{Q}$ defined in Equation (1). The red rectangles indicates the location where specific distributions are observed. They merge four GMM cells in the x-direction and two GMM cells in the y-direction.

varying weights are investigated in this paper. Therefore, the algorithm takes into account the weight of each electron.

Figure 2 compares the number of components identified by the detection algorithm in the left column with the measure of gyrotropy in the right column for various time steps. The objective is to highlight the behavior of the two quantities when the reconnection grows. Considering first the number of components, different structures are observed. Indeed, it can be clearly stated that it is not only the EDR, identified by the peak of Q, that is detected but also a much wider panel of different regions, which are symmetric with respect to the central plane $y = 10$. The GMM algorithm seems to locate inflows, ion diffusion region and EDRs, outflow, and separatrix boundaries. Another striking result of Figure 2 is the capability of the algorithm to detect regions where the influence of the reconnection seems to be weak, such as far upstream of the X line and near the O point. The noise of the PIC simulations is filtered out and unique distributions are successfully recognized for distributions with a single component. Starting with the first time step $t = 8000$, a large background tagged with two components extends from $y \approx 7$ to $y \approx 13$ and surrounds the EDR located at $x \approx 7$. This region may correspond to the ion diffusion region. The EDR is mainly composed of mixtures with five and four components, highlighting complex velocity distributions. The GMM analysis in this region is expected to be related to the results provided by Swisdak's measure of the gyrotropy. Both methods focus on the non-Maxwellianity and complexity of the distributions, through the moments for the measure of gyrotropy and directly by estimating the underlying probability density function for the GMM method. Downstream from the EDR in the outflow, a C-shape structure is noticeable on each side, characterized by distributions with four and five components connecting the EDR with the separatrix region. The latter is mainly composed of distributions with two and three components.

With regard to the three other time steps $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$, they show very similar structures and behaviors. The size of the EDR tends to slightly increase in

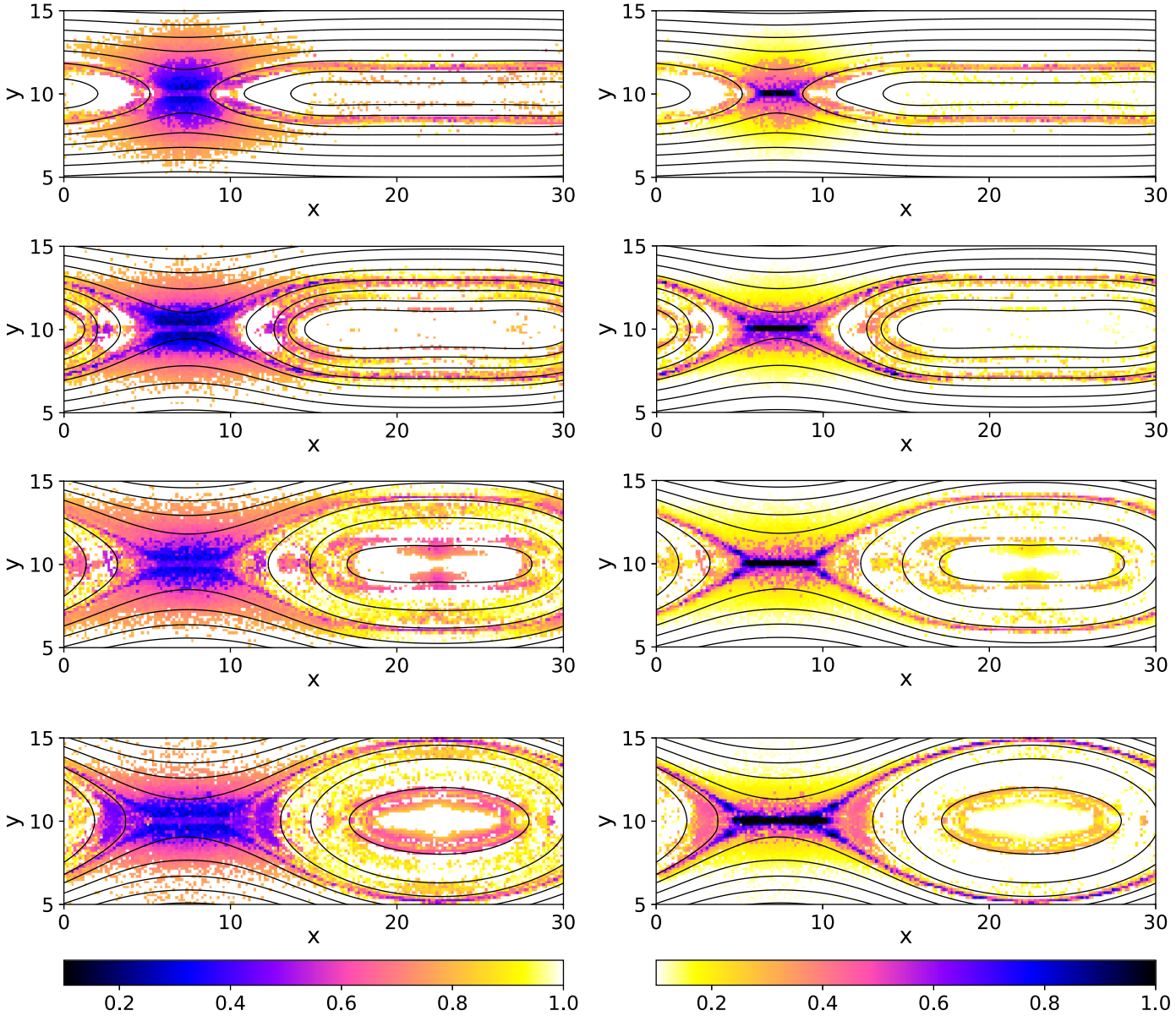

Figure 3. The left column highlights the energy drop $E_{\rm drop}$ defined by Equation (17), and the right column depicts the energy deviation $E_{\rm dev}$ given by Equation (19). Both quantities are presented at four different time steps, from top to bottom: $t = 8000$, $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$.

the x-direction over the time while the extent of the ion diffusion region remains steady. The outflow region is still clearly identified, and its location remains quite steady. The reconnection gives rise to a clear magnetic island on the right side of the figure at these time steps. The thickness of the region around the O point tends to increase dramatically in the y-direction when the reconnection grows. Several different distribution types can be observed, leading to a rather noisy mix with a background with two components and some with three and four components. Moreover, secondary structures gradually appear near the O point, creating a link between the bottom and the top layers of the island. Finally, two concentric ellipses can be observed at $t = 20{,}000$. They are composed of three components for the outer ellipse and two components for the inner ellipse. It is important to note that no spatial constraints or correlations are imposed on the detection algorithm, thus all the structures identified by the BIC minimization may exist in the distributions.

All of the results provided by the detection algorithm are then compared to the values of Q depicted in the right column in Figure 2 for the same time steps. A few similarities are observed: the measure of gyrotropy clearly highlights the EDR for all time steps with peak values observed above 0.5, and topological boundaries of the reconnection are also mapped, almost coinciding with the boundaries of the GMM algorithm with slight differences. However, different behaviors compared to the detection algorithm are exhibited. For instance, the region surrounding the EDR is not diagnosed by the measure of gyrotropy as well as the outflow and inner structures around the O point. Small artifacts seem to be present within the topological boundaries, but the background noise prevents them from being clearly identified. Indeed, the measure of gyrotropy is not exactly zero for regions far away from the reconnection with a background noise around 0.1, while the detection algorithm clearly identifies single distributions.

Figure 3 displays $E_{\rm drop}$ and $E_{\rm dev}$ in order to support the analysis of the number of components, helping to make

distinctions between the different distributions. First, the result for $E_{\rm drop}$ in the left column maps a large region around inflow and the EDR with respectively low values about 0.7 and 0.5, reflecting potential particle accelerations. Moreover, a very narrow region matching the EDR definition of the measure of gyrotropy is spotted by high values of $E_{\rm dev}$ illustration in the right column. Maximum values are truncated by the color bar, but they exceed 3.0. Thus, the algorithm clearly spots a specific region related to the EDR, similar to the measure of gyrotropy results. In particular, distributions in the EDR are expected to be very complex with mixtures distributed all over the velocity space, while clustered mixtures can be predicted for the ion diffusion region. For the two regions, mixtures with a small extent relative to the second statistical moment of the overall distribution are expected.

The energy deviation and energy drop also provide valuable information on distributions around the O point. In particular, the very low $E_{\rm dev}$ values (around 0.0) in this region suggest that a significant number of distributions tagged with two and three components should not be interpreted as beams but rather as a single distribution deviating from a Maxwellian. The high-energy drop ratio (about 0.9) supports this assumption as these clustered distributions show a thermal energy very close to the thermal energy of a Maxwellian. With regard to the outflow, quite small $E_{\rm drop}$ values about 0.6 are observed coupled with significant $E_{\rm dev}$ around 0.8. We can therefore assume that the mixtures have a pattern similar to the EDR but to a lesser extent. The two concentric ellipses near the O point observed in Figure 2 show very similar behaviors, with typical $E_{\rm drop}$ values of 0.7 and $E_{\rm dev}$ around 0.3, suggesting a deviation from a Maxwellian. Finally, the topological boundaries of the reconnection, identified by the measure of gyrotropy in Figure 2 seems also to be highlighted by the energy deviation, showing high values near the boundaries.

Figure 4 illustrates the different distributions associated with the five red rectangles displayed in Figure 2. These rectangles depict distributions by considering four square windows of the algorithm in the x-direction and two square windows in the y-direction. Several observations can be made:

1. Inflow region (box "A"): the distributions show a strong anisotropy demonstrating the heating along the parallel direction $v_{\parallel}$. This behavior is characteristic for the inflow region (Egedal et al. 2016). However, the GMM algorithm does not explain the data with a unique anisotropic Gaussian but uses two components to fit the broad mode with the short tail. It may suggest that the electron distribution already deviates from a Maxwellian within the inflow region. The short tail approximated by two components leads to high $E_{\rm drop}$ values as each mixture is relatively small compared to the overall second moment. The other directions perpendicular to the magnetic field show a Gaussian shape.
2. Electron distribution region (box "B"): a crescent shape seems to be observed in the $v_{\perp 1}-v_{\perp 2}$ marginal distribution, justifying the high number of components identified by the algorithm. As expected from the $E_{\rm drop}$ and $E_{\rm dev}$ analysis, all of the mixtures are spread over the distribution with a small width compared to the second moment computed over the whole distribution, which cannot properly fit such a complex distribution. Triangular shapes are depicted in the two other projections $v_{\parallel}-v_{\perp 1}$ and $v_{\parallel}-v_{\perp 2}$.
3. Outflow region (box "C"): a crescent shape is also observed, justifying why this region has a number of components similar to the EDR. A narrow Gaussian associated with a high weight fits the zero-centered mode of the distribution in the $v_{\perp 1}-v_{\perp 2}$ projection while scattered Gaussians with low weights are dedicated to the crescent shape. The two other projections are quite complex, due to the crescent shape. From the algorithm point of view, the crescent shape is hard to approximate, that is why five components are needed: one for the core and four others for the crescent.
4. Intermediate region (box "D"): the crescent shape has disappeared. The distribution deviates only slightly from a Maxwellian. Thin core distributions associated with strong weights are supplemented by wider distributions with smaller weights. This pair of distribution (core and tail) improves the overall fitting, in particular the long tail. This observation validates the results from $E_{\rm drop}$ and $E_{\rm dev}$: the mixture of the GMM variances is close to the second order ($E_{\rm drop}$ around 0.9) while the different components are almost all zero-centered, leading to $E_{\rm dev}$ values close to zero.
5. Outer ellipse near the O point (box "E"): the distribution shows a strong anisotropy along the parallel direction similar to the inflow. However, the distribution has a significant mode around zero coupled to a very broad tail. It explains the three components used by the algorithm: a first one approximates the mode while two others fit the tail on each side along the parallel direction.

7. Conclusion

We have proposed to automatically identify magnetic reconnection from velocity particle distributions using a density estimation technique called GMM. This approach has been able to identify various different regions around reconnection sites provided by a PIC simulation with a guide field value of 0.1, but also for a guide field value of 1.0. Analyzing the thermal heating and the distributions of the different components of the Gaussian mixtures gives a physical interpretation beyond the pure statistical properties of the GMM, helping to distinguish between heating and accelerating particles into beams but also between unimodal distributions and complex distributions. This method represents one of the first applications of machine learning algorithms to particle distributions in plasma physics. Nevertheless, it does not represent a unique solution to detect magnetic reconnection unambiguously. A central part of the algorithm is based on the BIC, which is sensitive to the number of particles. Therefore, the algorithm may require calibration to properly set the resolution with regard to the number of particles and available data.

For the moment, only 2.5D simulations with weighted particles have been investigated, but there is no reason to doubt that the approach cannot be applied to other types of simulations. Thus, testing the algorithm with 3D simulations represents a mandatory next step. Other fields of application could also be proposed, such as turbulence analysis. Moreover, simulations have access to the complete description of the plasma over all the spatial grid, while in situ observations are restricted to a a small set of measurements at specific spacecraft locations. Therefore, further works will also focus on

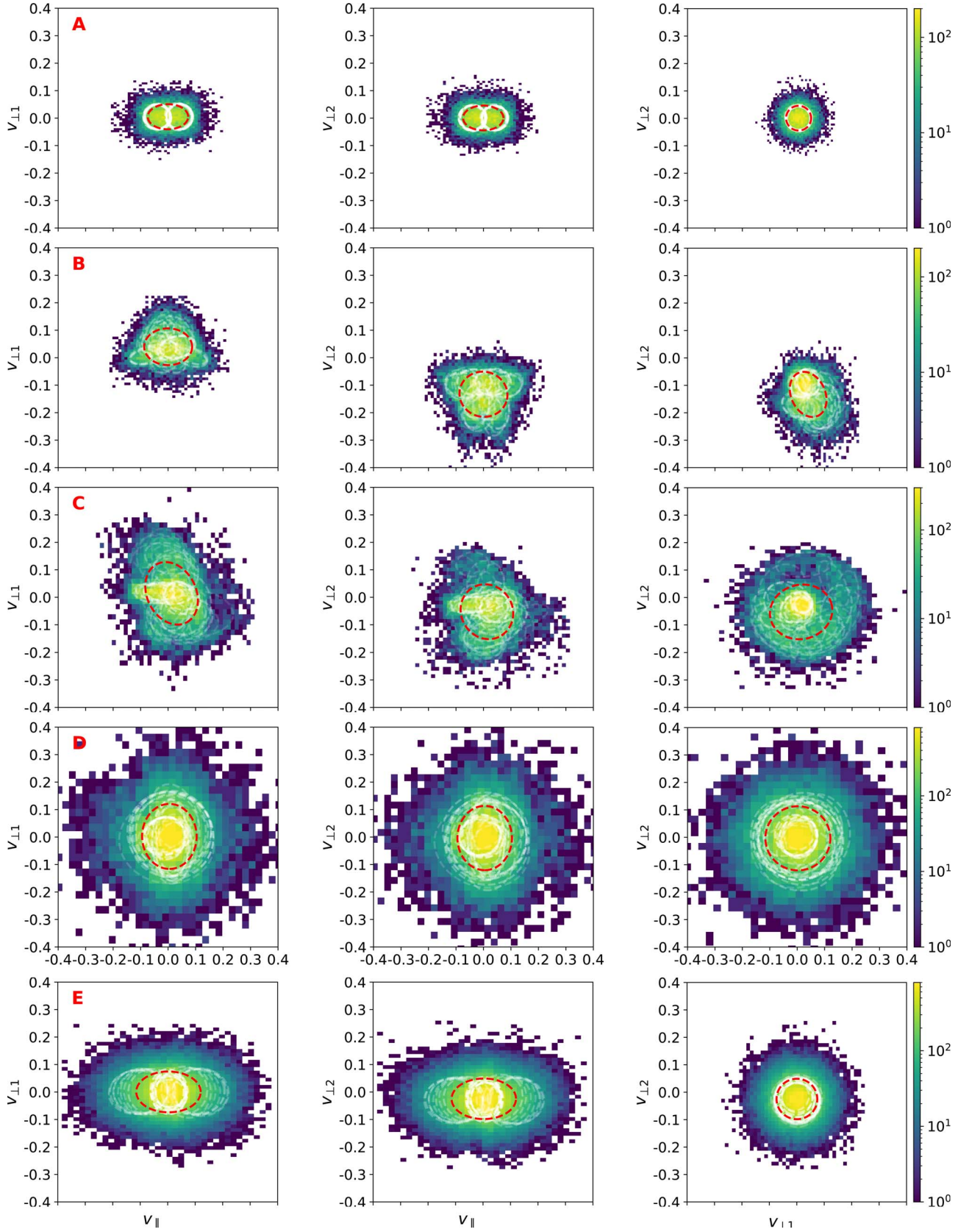

Figure 4. Electron velocity distribution for the double Harris sheet case at $t = 20{,}000$. Each row corresponds to one of the five red rectangles depicted in Figure 2. Three 2D marginal distributions are presented: $v_{\parallel}-v_{\perp 1}$, $v_{\parallel}-v_{\perp 2}$, and $v_{\perp 1}-v_{\perp 2}$. The white ellipses illustrate the different Gaussians of the mixtures in each distribution. The transparency is determined by the weight of each Gaussian: no transparency for a weight of 1 and a full transparency for a zero weight. The red ellipses give the mean and variance for a single distribution.

the application of this detection algorithm to local observational data of particle distribution function.

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 776262 (AIDA, www.aida-space.eu). This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under contract No. DE-AC02-05CH11231. Additional computing has been provided by NASA NAS and NCCS High Performance Computing, by the Flemish Supercomputing Center (VSC), and by PRACE Tier-0 allocations. The paper has received funding from Intern Fondsen KU Leuven / Internal Funds KU Leuven.

Appendix A
Guide Field Value of 1.0

As a second case, we consider the same double Harris sheet simulation but with a guide field value of 1.0. Figure 5 presents the same quantities as Figure 2 but for the guide field value of 1.0: the number of components in the left column and the measure of gyrotropy in the right column. The regions identified by the algorithm are very different. The inflow has almost vanished while the EDR is very hard to identify. The outflow seems to be present, in particular at $t = 12{,}000$, but even this region tends to disappear for the other time steps, leading to a quite noisy distribution of the number of components. As in the previous guide field case (value of 0.1), the measure of gyrotropy shares roughly the same topological boundaries, although the detection algorithm identifies wider regions, in particular around the EDR.

Figure 6 presents the results for $E_{\rm drop}$ and $E_{\rm dev}$, and is less informative for the guide field value of 1.0. For instance, low $E_{\rm drop}$ values around 0.4 are observed for almost all distributions with two or more components at $t = 20{,}000$, meaning the presence of beams is more likely. Nevertheless, a broad region is highlighted for $E_{\rm drop}$ at the first time step $t = 8000$, which is not the case for the measure of gyrotropy. With regard to the energy deviation $E_{\rm dev}$, high values above 0.8 are observed from $t = 12{,}000$ near the topological boundaries around the EDR, the outflow, and the separatrix. Thus, complex distributions are expected in these regions, which have also been highlighted by the measure of gyrotropy in Figure 5.

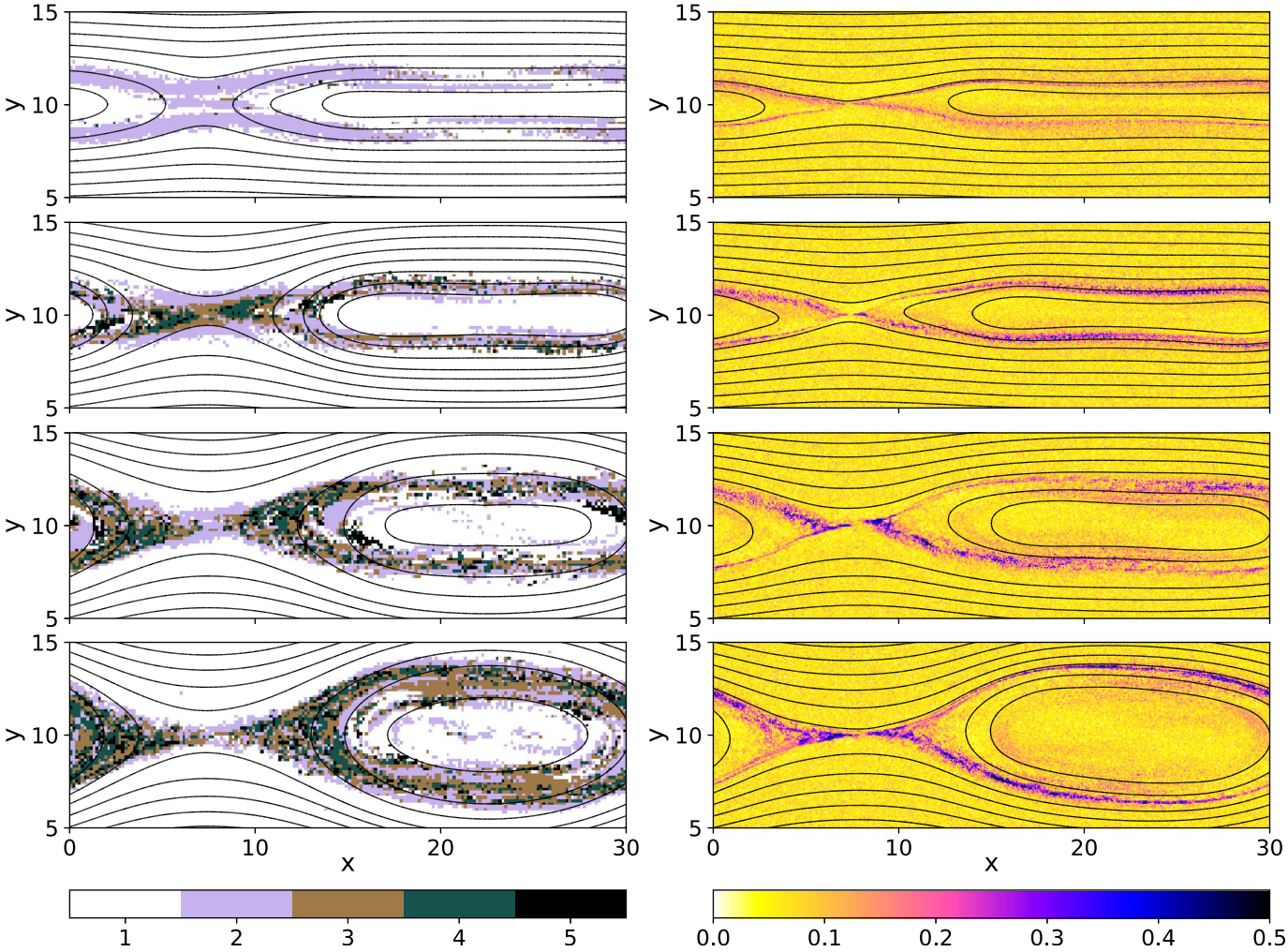

Figure 5. Magnetic reconnection detection for the double Harris sheet case with a guide field value of 1.0 at four different time steps, from top to bottom: $t = 8000$, $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$. The left column presents the number of components provided by the BIC optimization, and the right column shows the measure of gyrotropy $\sqrt{Q}$ defined in Equation (1). The red rectangles indicates the location where specific distributions are observed.

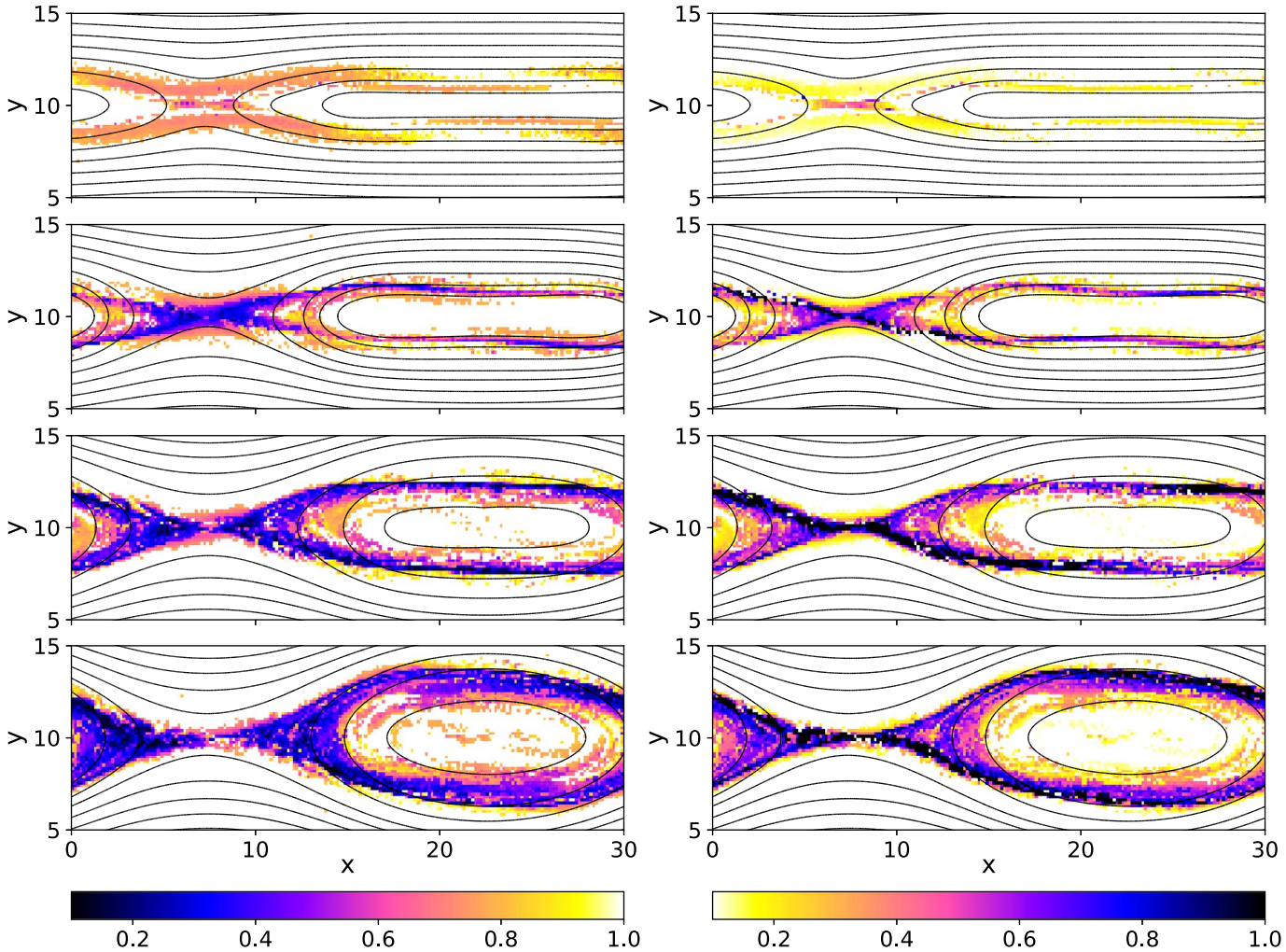

Figure 6. The left column highlights the energy drop $E_{\rm drop}$ defined by Equation (17), and the right column depicts the energy deviation $E_{\rm dev}$ given by Equation (19). Both quantities are presented at four different time steps, from top to bottom: $t = 8000$, $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$.

Appendix B
Sensitivity to the Resolution

Determining the number of components for the detection algorithm represents a not trivial model selection problem. In this paper, minimizing BIC has been chosen as the reference method. As stated in Equation (13), the sample size directly influences BIC as the latter depends on ln n. Therefore, the detection algorithm is sensitive to the number of particles provided to the GMM, encoded by the resolution chosen for the window defined in the Section 4.4. The latter must be selected carefully to find a good trade-off: a very broad window may mix several different particle populations, missing important physical scales, while a very small window cannot reach sufficient statistical convergence, due to a very low number of particles.

For the double Harris sheet case with a guide field value of 0.1, a window of four by four cells has been selected, ensuring more than 2000 particles per window over the whole domain. Figure 7 depicts the impact of the window length on the number of components provided by BIC minimization at $t = 20{,}000$. Four lengths are investigated: one, two, four, and eight cells. The smaller window ($r = 1$ cell) barely detects the EDR and the topological boundaries and shows some noise. About a hundred particles are used to train each model, thus the mixture can miss important shapes and the underlying distributions are not necessarily recovered. Typical structures start to be identified for $r = 2$, where the EDR, outflow, O point, and separatrix regions show a significant size with a spatial correlation in terms of the number of components. Only the background region around the EDR is filtered out. Finally, the shapes identified for $r = 4$ and $r = 8$ look very similar, and

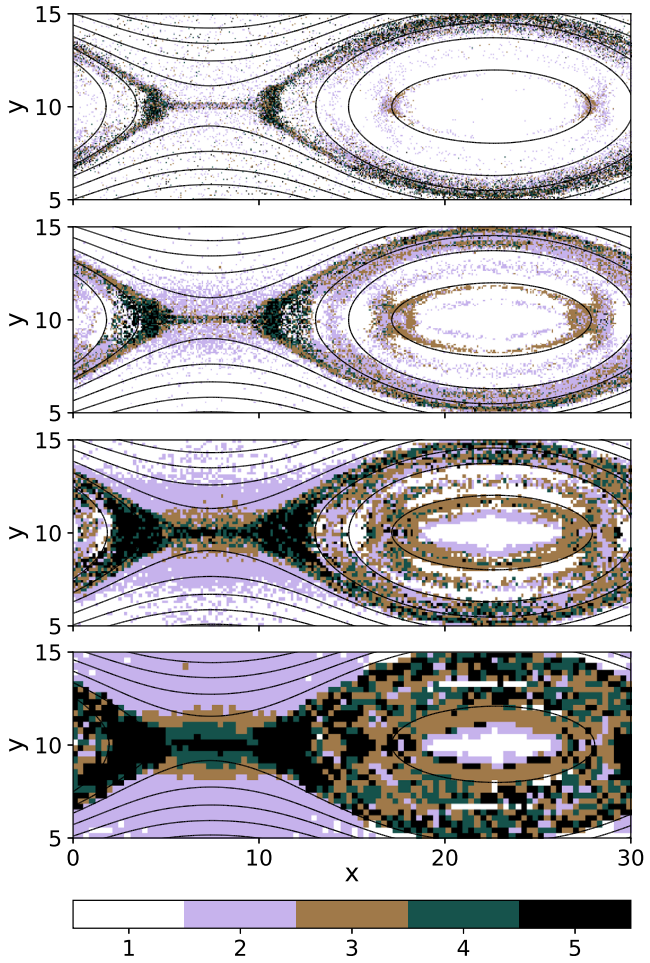

Figure 7. Impact of the number of particles on the BIC optimization. Four different window length are depicted at $t = 20{,}000$ from top to bottom: one, two, four, and eight cells.

only a few new distributions are observed, such as a wider region with five components around the outflow or a wider inflow. Therefore, Figure 7 illustrates perfectly the sensitivity of the algorithm and the BIC minimization to the window length and the number of particles. Even if characteristic structures of the reconnection are identified for each window length, from one to eight cells, a minimum number of particles (here about 2000) is needed to ensure a proper statistical convergence while making sure different populations are not merged.

Appendix C
Fixed Number of Components

Here, a fixed number of eight components is imposed for the mixture models. It ensures a greater flexibility for the GMM as very complex distributions can be more easily described, reproducing nonparametric density estimation methods. Figure 8 shows the energy drop $E_{\rm drop}$ and the energy deviation $E_{\rm dev}$. The structures identified by the algorithm are very close to the ones found with the BIC minimization depicted in Figure 3. For instance, the large background region surrounding the EDR is identified in both cases by the energy drop as well as the peak value of the EDR highlighted by $E_{\rm drop}$. Specific common features are also observed around the 0 point. The only significant difference is observed for regions tagged with a single component by the BIC minimization. Distributions far upstream from the EDR have a noise level around 0.5 for $E_{\rm drop}$ and around 0.4 for $E_{\rm dev}$ with a fixed number of components, while BIC minimization provides values close to 1.0 for $E_{\rm drop}$ and almost zero values for $E_{\rm dev}$. Therefore, BIC appears to be a good criterion to identify relevant statistical patterns in the data and to filtering out the noise in the distributions.

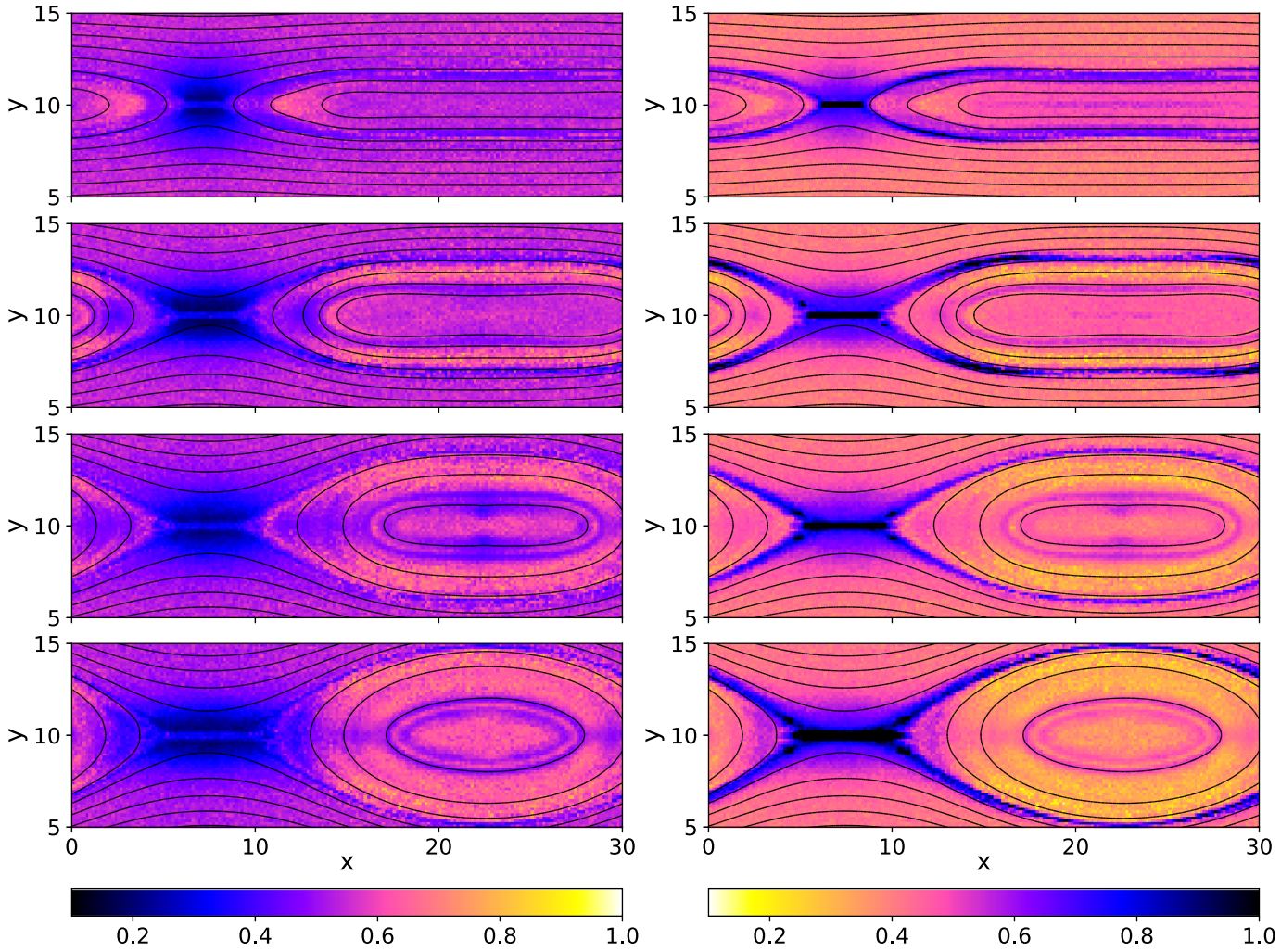

Figure 8. Detection algorithm with a fixed number of eight components for the weak guide field case. The left column highlights the energy drop $E_{\rm drop}$ defined by Equation (17), and the right column depicts the energy deviation $E_{\rm dev}$ given by Equation (19). Both quantities are presented at four time steps, from top to bottom: $t = 8000$, $t = 12{,}000$, $t = 16{,}000$, and $t = 20{,}000$.

ORCID iDs

Romain Dupuis https://orcid.org/0000-0002-7976-1034
Jorge Amaya https://orcid.org/0000-0003-1320-8428
Giovanni Lapenta https://orcid.org/0000-0002-3123-4024

References

Albertsson, K., Altoe, P., Anderson, D., et al. 2018, JPhCS, 1085, 022008
Anderson, D. 2002, Model Selection and Multi-model Inference—A Practical Information-Theoretic Approach (New York: Springer)
Ashour-Abdalla, M., Lapenta, G., Walker, R. J., El-Alaoui, M., & Liang, H. 2015, JGRA, 120, 4784
Aunai, N., Hesse, M., & Kuznetsova, M. 2013, PhPl, 20, 092903
Bessho, N., Chen, L.-J., & Hesse, M. 2016, GeoRL, 43, 1828
Birn, J., Drake, J., Shay, M., et al. 2001, JGRA, 106, 3715
Birn, J., & Priest, E. R. 2007, Reconnection of Magnetic Fields: Magnetohydrodynamics and Collisionless Theory and Observations (Cambridge: Cambridge Univ. Press)
Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Secaucus, NJ: Springer)
Biskamp, D. 2000, Magnetic Reconnection in Plasmas (Cambridge: Cambridge Univ. Press)
Burch, J., Moore, T., Torbert, R., & Giles, B. 2016a, SSRv, 199, 5
Burch, J., Torbert, R., Phan, T., et al. 2016b, Sci, 352, 2939
Camporeale, E., Wing, S., & Johnson, J. 2018, Machine Learning Techniques for Space Weather (Amsterdam: Elsevier)
Cazzola, E., Innocenti, M. E., Goldman, M. V., et al. 2016, GeoRL, 43, 7840
Colomé, A., Foix, S., Alenyà, G., & Torras, C. 2017, in ROBOT 2017: Third Iberian Robotics Conf., ed. A. Ollero et al. (Cham: Springer), 141
del Castillo-Negrete, D., Schneider, K., Farge, M., Chen, G., et al. 2010, JCoPh, 229, 2821
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. R. Stat. Soc. B, 39, 1
Eastwood, J., Goldman, M., Hietala, H., et al. 2015, JGRA, 120, 511
Eastwood, J., Phan, T., Drake, J., et al. 2013, PhRvL, 110, 225001
Egedal, J., Le, A., Daughton, W., et al. 2016, PhRvL, 117, 185101
Fu, H., Vaivads, A., Khotyaintsev, Y. V., et al. 2015, JGRA, 120, 3758
Goldman, M., Newman, D., & Lapenta, G. 2016, SSRv, 199, 651
Gonzalez, W., & Parker, E. 2016, ASSL, 427, 10
Gruntman, M. A. 1992, P&SS, 40, 439
Harris, E. G. 1962, NCim, 23, 115
Haynes, A. L., Parnell, C. E., Galsgaard, K., & Priest, E. R. 2007, RSPSA, 463, 1097
Haynes, C. T., Burgess, D., & Camporeale, E. 2014, ApJ, 783, 38
Heidenreich, N.-B., Schindler, A., & Sperlich, S. 2013, Adv. Stat. Anal., 97, 403
Heidrich-Meisner, V., & Wimmer-Schweingruber, R. F. 2018, Machine Learning Techniques for Space Weather (Amsterdam: Elsevier), 397
Hellberg, M., & Mace, R. 2002, PhPl, 9, 1495
Hesse, M., Aunai, N., Sibeck, D., & Birn, J. 2014, GeoRL, 41, 8673
Hesse, M., Neukirch, T., Schindler, K., Kuznetsova, M., & Zenitani, S. 2011, SSRv, 160, 3
Hesse, M., & Schindler, K. 1988, JGRA, 93, 5559
Innocenti, M. E., Johnson, A., Markidis, S., et al. 2017, Adv. Eng. Softw., 111, 3

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, Vol. 1 (Princeton, NJ: Princeton Univ. Press)
Kasper, J., Lazarus, A., Steinberg, J., Ogilvie, K., & Szabo, A. 2006, JGRA, 111, A03105
Lapenta, G., Berchem, J., Zhou, M., et al. 2017, JGRA, 122, 2024
Lau, Y.-T., & Finn, J. M. 1990, ApJ, 350, 672
Lazar, M., Shaaban, S., Fichtner, H., & Poedts, S. 2018, PhPl, 25, 022902
Lembege, B., & Pellat, R. 1982, PhFl, 25, 1995
Livadiotis, G., Desai, M., & Wilson, L., III 2018, ApJ, 853, 142
Livadiotis, G., & McComas, D. 2013, SSRv, 175, 183
Loureiro, N. F., Dorland, W., Fazendeiro, L., et al. 2016, CoPhC, 206, 45
Markidis, S., Lapenta, G., et al. 2010, Math. Comput. Simul., 80, 1509
McLachlan, G., & Peel, D. 2004, Finite Mixture Models (New York: Wiley)
Meyrand, R., Kanekar, A., Dorland, W., & Schekochihin, A. A. 2019, PNAS, 116, 1185
Newcomb, W. A. 1958, AnPhy, 3, 347
Ni, B., Zou, Z., Gu, X., et al. 2015, JGRA, 120, 4863
Ogasawara, K., Angelopoulos, V., Dayeh, M., et al. 2013, JGRA, 118, 3126
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pierrard, V., & Lazar, M. 2010, SoPh, 267, 153
Priest, E., & Forbes, T. 2007, in Magnetic Reconnection, ed. E. Priest & T. Forbes (Cambridge: Cambridge Univ. Press)
Priest, E. R., & Titov, V. 1996, RSPTA, 354, 2951
Pulupa, M., Bale, S., Salem, C., & Horaites, K. 2014, JGRA, 119, 647
Radovic, A., Williams, M., Rousseau, D., et al. 2018, Natur, 560, 41
Scudder, J., & Daughton, W. 2008, JGRA, 113, A06222
Servidio, S., Chasapis, A., Matthaeus, W., et al. 2017, PhRvL, 119, 205101
Sheather, S. J. 2004, StaSc, 19, 588
Shuster, J., Chen, L.-J., Daughton, W., et al. 2014, GeoRL, 41, 5389
Shuster, J., Chen, L.-J., Hesse, M., et al. 2015, GeoRL, 42, 2586
Sitnov, M., Buzulukova, N., Swisdak, M., Merkin, V., & Moore, T. 2013, GeoRL, 40, 22
Souza, V. M., Medeiros, C., Koga, D., et al. 2018, Machine Learning Techniques for Space Weather (Amsterdam: Elsevier), 329
Summers, D., & Thorne, R. M. 1991, PhFlB, 3, 1835
Swisdak, M. 2016, GeoRL, 43, 43
Titov, V. S. 2007, ApJ, 660, 863
Vasyliunas, V. M. 1968, JGR, 73, 2839
Vasyliunas, V. M. 1975, RvGeo, 13, 303
Wilson, L. B., III, Chen, L.-J., Wang, S., et al. 2019, arXiv:1902.01476
Wu, P., Shay, M., Phan, T., Oieroset, M., & Oka, M. 2011, PhPl, 18, 111204
Yamada, M., Yoo, J., Jara-Almonte, J., et al. 2014, NatCo, 5, 4774
Zenitani, S., Hesse, M., Klimas, A., & Kuznetsova, M. 2011, PhRvL, 106, 195003