Detecting Reconnection Events in Kinetic Vlasov Hybrid Simulations Using Clustering Techniques

Manuela Sisti1,2, Francesco Finelli2, Giorgio Pedrazzi3, Matteo Faganello1, Francesco Califano2, and Francesca Delli Ponti3

1 Aix-Marseille University, CNRS, PIIM UMR F-7345, Marseille, France; manuela.sisti@univ-amu.fr
2 Dipartimento di Fisica, Università di Pisa, Italy
3 HPC Department, Cinéca, Bologna, Italy

Received 2020 July 13; revised 2020 November 20; accepted 2020 December 8; published 2021 February 17

Abstract

Kinetic turbulence in magnetized space plasmas has been extensively studied via in situ observations, numerical simulations, and theoretical models. In this context, a key point concerns the formation of coherent current structures and their disruption through magnetic reconnection. We present automatic techniques aimed at detecting reconnection events in a large data set of numerical simulations. We make use of clustering techniques known as K-means and DBscan (usually referred to in literature as unsupervised machine-learning approaches), and other methods based on thresholds of standard reconnection proxies. All our techniques also use a threshold on the aspect ratio of the regions selected. We test the performance of our algorithms. We propose an optimal aspect ratio to be used in the automated machine-learning algorithm: AR = 18. The performance of the unsupervised approach results in it being strongly competitive with respect to those of other methods based on thresholds of standard reconnection proxies.

Unified Astronomy Thesaurus concepts: Solar magnetic reconnection (1504); Astronomy data analysis (1858); Astronomy data modeling (1859); Computational methods (1965); Clustering (1908)

1. Introduction

Magnetic reconnection is a ubiquitous phenomenon observed in the laboratory as well as in many space and astrophysical environments. It is the energy driver for solar flares and coronal mass ejections (Priest 1982). It occurs routinely at the boundary between the solar wind and the Earth’s magnetosphere (Dungey 1961; Priest & Forbes 2001), both at the dayside and at the magnetotail, or at the flanks induced by the presence of large-scale Kelvin–Helmholtz vortices (Faganello & Califano 2017). An important outcome is the injection of accelerated particles into the magnetosphere. In the context of turbulent plasmas, such as solar wind (Haynes et al. 2014; Osman et al. 2014) and magnetosheath plasma (Retinò et al. 2007), reconnection plays a central role representing a possible alternative path, as compared to the usual vortex–vortex interaction, for energy transfer toward smaller and smaller scales (Karimabadi et al. 2013; Cerri & Califano 2017; Camporeale et al. 2018). It is behind many of the risks associated with space weather such as, for example, disturbance of Global Navigation Satellite System signals and electronic damage to satellites or power grids (Cassak 2016).

Magnetic reconnection is probably the most known and important process in magnetized plasmas for which magnetohydrodynamics (MHD) accurately describe the evolution at large scales, where the magnetic field is frozen in the fluid motion of the plasma and its topology is preserved. In this context, magnetic reconnection is the only process that is able to rearrange the magnetic field topology on a global, MHD scale despite its occurrence in a very narrow region of typical size of the order of the ion kinetic scale length. The magnetic rearrangement, being on the large MHD scale, is accompanied by a strong energy release due to the conversion of magnetic energy into plasma flow, particle acceleration, heating, and wave generation. The classical picture of reconnection is obtained by considering a 2D domain where a large-scale magnetic field, the so-called equilibrium, presents an inversion line where it changes its direction. This is a 1D equilibrium configuration, since \( \partial_x = \partial_y = 0 \) while the inversion takes place along the \( y \)-direction, that corresponds to a kinetic equilibrium for the distribution functions known as the Harris sheet (Harris 1962). The 1D magnetic field can be represented as \( B = B(x) e_x \), directed along the \( x \)-axis and varying along the \( y \)-axis being zero along the neutral line at a given \( y = y_0 \). The corresponding out-of-plane current has a peak and the region around is dubbed the current sheet (hereafter CS) or current layer. In certain favorable conditions this layer becomes unstable to the 2D reconnection instability and the ideal MHD laws are locally violated (Furth et al. 1963; White 1986; Priest & Forbes 2001) leading to the breaking and reconnection of field lines and eventually to the formation of X-point-like structures characterized by an inflow/outflow ion and electron fluid velocity advecting the magnetic flux toward/away from the reconnecting regions. Often in the laboratory but also in space, a mean out-of-plane magnetic field, \( \mathbf{B}_0 = B_0 e_y \), is associated with the CS. In this case the dynamics is quasi-2D and reconnecting lines do not “cut&past” but actually slip across the CS changing their original connectivity, even if their projection still seems to break and reconnect. This configuration, the so-called guide-field regime, is often studied by means of 2D simulations.

In such a system, magnetic reconnection can be easily recognized. On the contrary, as soon as CSs have more complex shapes, ascertaining the presence of reconnection is far less simple, e.g., when CSs are dynamically generated by 2D (Henri et al. 2012; Daughton et al. 2014) or 3D (Faganello et al. 2014; Borgogno et al. 2015; Sisti et al. 2019) Kelvin–Helmholtz vortices or, even worse, by magnetized turbulence (Servidio et al. 2010; Zhdankin et al. 2013). For each of these cases a different ad hoc method has been developed, based on a deep knowledge of the peculiarity of each system.
To investigate the complex nonlinear dynamics of such systems, numerical simulations are widely used, providing important support for the understanding of satellite data. However, this approach generates an impressive amount of data to deal with, in particular when working with kinetic simulations spanning the entire phase space. Large data sets, as well as the difficulty of easily recognizing reconnecting structures in these sets, suggest a possible application of machine learning for reconnection analysis. Recently, in Dupuis et al. (2020), some machine-learning techniques based on signatures in the particle velocity distribution function in the phase space have been tested to identify reconnection regions. Moreover, in Hu et al. (2020) for the same purpose to automatically find magnetic reconnection sites, supervised machine-learning methods based on the Convolutional Neural Network are used. Hu et al. (2020) have shown that the supervised machine-learning approach performs very well and is therefore promising and that the algorithm seems to be capable of detecting reconnection even in cases in which human labeling has failed. However, the setup of a large labeled training data set is a long-term process that requires contributions from a number of experts in the field. On the other hand, an unsupervised algorithm does not require a training data set, thus being more flexible and quicker to implement (but not simpler). Here we analyze the performances of unsupervised algorithms in order to prove that such an approach can be an efficient method to automatically detect reconnection. For this purpose, we set up a hybrid 2D-3V (2D in space, 3D in velocity space) simulation with kinetic ions and fluid electrons, describing the turbulent dynamics of a collisionless plasma, where magnetic structures and CSs, continuously generated by plasma motions, interact together and eventually coalesce or disrupt due to magnetic reconnection. Beyond the physical interest of this simulation set on typical solar wind turbulence parameters, it represents a good test for a machine-learning approach since reconnection occurs in a variety of different configurations far from the idealized 1D Harris-like.

Aiming at developing an automatic procedure for the detection of physical structures of basic interest, such as CSs and reconnecting structures, we have developed the following techniques. The first group relies on “standard” unsupervised machine-learning techniques, such as K-means (in particular, Lloyd’s algorithm; MacQueen 1967) and DBscan (Density Based Spatial Clustering of Applications with Noise; Ester et al. 1996). The second group instead is based on a number of standard proxies used in the literature, and the definition of the corresponding thresholds, as markers to detect and highlight the presence of a reconnection event. In particular we cite current density, electron vorticity, and decoupling of the electron dynamics from the magnetic field. Note that all methods use physical quantities particularly suited to be considered as a signature of reconnection. These quantities, extracted here from numerical simulations, are the ones usually measured by onboard instruments of in situ satellites. All methods are based on a final fundamental step for selecting reconnecting structures: the definition of a threshold on the aspect ratio of the CSs. This threshold is motivated by the physics of the reconnecting CSs whose shape (their typical length and width) is not random but imposed by the development of the reconnection instability.

There exist several models of reconnection that could be used to extrapolate a reasonable value for this threshold. The simplest one is the resistive Sweet–Parker model (Parker 1957; Sweet 1958) where the CS length \( L \) and width \( \delta \) depend on the local reconnection rate \( R \), i.e., on the outflow velocity (Cassak et al. 2017). Actually, the reconnection rate predicted from the Sweet–Parker model is too slow to account for observations and simulations in collisionless plasmas. Other models such as the so-called “fast reconnection” predict a reconnection rate of the order of \( \sim 0.1 \), quite in agreement with observations and simulation results recently found in the literature (Cassak et al. 2017). A different approach for the estimation of the CS aspect ratio relies on the role of the tearing mode as a sufficient condition for reconnection to occur. By looking at the wavenumber \( k_m \) of the most unstable mode in a CS of width \( \delta \), we get \( \delta / L \sim 0.08 \) (Karimabadi et al. 2005), not far from \( \sim 0.1 \). In other words, there is a direct link between the aspect ratio and the reconnection rate. Therefore, we make use of the aspect ratio to distinguish structures where most probably reconnection is on-going using it as a threshold.

The paper is organized as follows. In Section 2 the Hybrid Vlasov–Maxwell model (HVM) and the 2D-3V simulation data are introduced. In Section 3 we describe the methods developed to automatically identify CSs and reconnecting regions. In particular in Section 3.1 we discuss our main method, which uses the unsupervised machine-learning techniques. In Sections 3.2 and 3.3 we present the other two alternative methods that do not use machine learning but are still automatic algorithms to locate reconnecting regions using standard reconnection proxies. In Section 4 we discuss the performances of these different methods in finding reconnection sites. Finally, our results are summarized in Section 5.

2. Simulation

To identify a wide panorama of reconnecting sites emerging dynamically and not prepared ad hoc, we make use of a 2D-3V plasma turbulence simulation and use an hybrid model with kinetic ions and fluid electrons (with mass; Valentini et al. 2007). This simulation is run using the HVM code based on an Eulerian approach to solve the ion Vlasov equation (Mangeney et al. 2002) coupled to the Maxwell equations but neglecting the displacement current and adopting the quasi-neutral approximation, \( n_i = n_e = n \). We use ion quantities to normalize the equations. These are the ion mass \( m_i \), the ion cyclotron frequency \( \Omega_{ci} = eB / m_i c \), Alfvén velocity \( v_A = B / \sqrt{4\pi m_i n} \), and the ion skin depth \( d_i = \sqrt{m_i / \Omega_{ci}} \). The electron skin depth reads \( d_e = \sqrt{m_e / m_i} \), where \( m_e \) is the electron mass. Finally \( n, B, \Omega, f \) are the normalizing density, magnetic field, electric field, and distribution function, respectively. Then, the Vlasov equation for the ion distribution function \( f = f(x, y, v_x, v_y, v_z, t) \) reads:

\[
\frac{\partial f}{\partial t} + v \cdot \frac{\partial f}{\partial x} + (E + v \times B) \cdot \frac{\partial f}{\partial v} = 0.
\]

The Ohm’s equation for the electron response reads:

\[
E - d_e^2 \nabla^2 E = -u_e \times B - \frac{1}{n} \nabla p_e + d_i^2 \nabla \cdot [n(\nabla \Omega - u_i u_e)]
\]

where \( u = \frac{\sqrt{\mathbf{v}}}{\sqrt{\mathbf{f}} \, d\mathbf{r}} \) and \( u_e \) are the ion and the electron fluid velocities, respectively. Furthermore \( n \) is given by \( n = \int \mathbf{f} \, d\mathbf{r} \).
Finally the dimensionless Faraday and Ampere laws read:

\[ \nabla \times B = J; \quad \frac{\partial B}{\partial t} = -\nabla \times E. \]  

(3)

For the sake of simplicity, we take an isothermal closure: \( P_e = n T_0 e \). In this approach electron inertia (terms proportional to \( d_e^2 \) in Equation (2)) is a key ingredient since it allows for reconnection to occur by decoupling the electrons from the magnetic field.

We take a squared simulation box with \( L = L_x = L_y = 2\pi \times 50 \) using \( N_x = N_y = 3072 \) points. The corresponding physics goes from the MHD fluid-like behavior of the largest wavelength \( \lambda \sim L \) to the ion kinetic physics \( \lambda \sim d_i \) (=1 in dimensionless units) to the electron inertial scale, \( \lambda \sim d_e \). In terms of the corresponding frequencies we are across the ion cyclotron physics (=1 in dimensionless units). We set the magnitude of the initial guide field along the \( z \)-direction equal to one. The initial ion distribution function is given by a Maxwellian with corresponding uniform temperature. The electron temperature is set equal to that of the ions, \( T_{ei} = T_0 \). As a result, the plasma beta, the ratio between the fluid to the magnetic pressure, is equal to one (still in dimensionless units). We sample the velocity space using \( 51^3 \) uniformly distributed grid point spanning \([-5v_{th,i}, 5v_{th,i}]\) in each direction, where \( v_{th,i} = \sqrt{\beta_i/2} \) is the initial ion thermal velocity. The mass ratio is \( m_i/m_e = 100 \) allowing us to well separate the ion (\( d_i \)) and electron (\( d_e \)) skin depth. The simulation is initialized by adding an isotropic magnetic perturbation given by the sum of sinusoidal modes with random phase. The corresponding wavenumber interval is \( k \in [0.02, 0.12] \). The corresponding rms value of the magnetic field is \( \delta B_{\text{rms}} \approx 0.28 \). In the initial phase, the initial perturbation starts to produce a number of noninteracting CSs where, later, at about one eddy turnover time \( \sim 250 \), associated with the largest perturbed wavelength, reconnection becomes effective. After about two ion turnover times the system reaches a fully developed turbulent state. Two snapshots of the current density are shown in Figure 1 corresponding to the different stages: when CSs are well formed (left frame) and start to reconnect, and the fully developed turbulent state (right frame).

Figure 1. Snapshot of the contour plots of current density |\( J \)| for the two different times of our simulation: (a) \( t = 247 \) (\( 1/\Omega_{ci} \)) (current sheets formed), and (b) \( t = 494 \) (\( 1/\Omega_{ci} \)) (fully developed turbulence).

3. Methods

In order to speed up the identification of CSs and reconnecting structures, we develop three different techniques. The main one uses unsupervised machine learning, and will be addressed hereafter as “AML.” Two other alternative algorithms, which do not use machine learning, are developed in order to compare the performances; they will be addressed as “A1” and “A2.” A comparison between these three methods will be presented hereafter. The physical quantities we use as variables for detecting reconnection in AML are the current density magnitude |\( J \)|, the magnitude of the in-plane electron fluid velocity |\( u_{ei,\text{in-plane}} \)|, the magnitude of the in-plane magnetic field |\( B_{\text{in-plane}} \)|, the magnitude of the electron vorticity |\( \Omega_e = |\nabla \times u_e| \)|, the electron decoupling term |\( E_{\text{dec.e}} = |(E + u_e \times B)| \)| and the energy conversion term |\( J \cdot (E + u_e \times B) \)|. The first three variables are related to the geometry of the CSs. The electron decoupling term describes the de-freezing of magnetic field lines from the electron fluid motion. The last variable is a proxy that accounts for the energy dissipation at the reconnection sites (Zenitani et al. 2011). In A1 and A2 we use only some of them, in particular in A1 |\( J \)|, while in A2 \( J \), \( \Omega_e \), and \( E_{\text{dec.e}} \).

The aspect ratio (hereafter AR) of reconnecting CSs is not random but it is linked to reconnection development, thus all our algorithms have a common step: a threshold on the AR of the CSs selected. We define the AR as

\[ \text{AR} = \frac{\text{length}}{\text{width}} \]  

(4)

for each structure found. We give an estimation for the CSs’ width and length using the automated method explained in Califano et al. (2020) and references therein. Note that for methods A1 and A2 the CSs are defined as regions where the current density \( J \) is greater than a certain threshold, as defined in Zhdankin et al. (2013), while in AML, we make use of a different technique to define the CSs in the physical space (as explained in the following sections). Once a CS is defined as a collection of grid points in the 2D physical space, the automated procedure computes the Hessian matrix \( H \) of |\( J \)| at
the current peak (the local maximum of $J$). We look at the interpolated profile of $J$ along the direction given by the eigenvector associated to the largest eigenvalue of $H$ and define the CS’s width as the FWHM of $J$. A similar procedure for computing the length (i.e., interpolating $J$ along the direction of the eigenvector associated to the smallest eigenvalue) would be misleading because in a turbulent system CSs are seldom “straight” along that direction. We give an automated estimation for the length by computing the maximal distance between two points belonging to the same structure, similarly to what was done in Zhdankin et al. (2013), where the CSs are defined using a threshold on the current density.

It is worth noticing that all the variables we use are available in satellite data sets. In particular $\mathbf{u}_e$ and $|\mathbf{B}|$ are directly measured by onboard instruments, while $J = |\rho(\mathbf{u} - \mathbf{u}_e)|$, $E_{\text{dec,e}} = |(\mathbf{E} + \mathbf{u}_e \times \mathbf{B})|$, and $J \cdot (\mathbf{E} + \mathbf{u}_e \times \mathbf{B})$ are simple algebraic combinations of measured quantities. $\mathbf{u}_{\text{in-plane}}$ and $|\mathbf{B}_{\text{in-plane}}|$ can be computed by setting the out-of-plane direction aligned with the average (local) magnetic field. Finally, $\Omega_v$ and $AR$ can be calculated by measured quantities and the relative positions of spacecraft in the case of multisatellite missions, such as CLUSTER or MMS (Dunlop et al. 2002; Fadanelli et al. 2019).

3.1. AML

The core of the first method is constituted by two unsupervised machine-learning algorithms: $K$-means (Lloyd’s algorithm) and DBscan (Density Based Spatial Clustering of Applications with Noise). These are both clustering techniques aimed at learning a grouping structure in a data set. $K$-means algorithm requires us to set the parameter “$K$” corresponding to the number of clusters we expect to find in our data set in the variable space. To fix the parameter “$K$” we make a preprocessing step in which we tune the best value of “$K$” for our data. Our approach can then be summarized in the following steps:

1. Preprocessing step for the tuning of the “$K$” parameter for the $K$-means algorithm using a cross-validation-like approach with the Davis–Bouldin index applied to the predicted clusters as an internal cluster metric (Rhys 2020). The tuning is applied in the variable space defined at the beginning of Section 3. We choose to tune the “$K$” value at about one eddy-turnover time ($t \approx 247 \ [1/\Omega_v]$), corresponding to the phase when CSs are well formed but still not interacting with each other. The result of the tuning is shown in Figure 2. The best “$K$” value is the one for which the Davis–Bouldin index reaches its minimum, in our case it turns out to be $K = 11$, as shown in Figure 2.

2. $K$-means (Lloyd’s algorithm) is applied to the chosen variables, which we list here again: the current density magnitude $J = |J|$, the magnitude of the in-plane electron fluid velocity $|V_{e,\text{in-plane}}|$, the magnitude of the in-plane magnetic field $|B_{\text{in-plane}}|$, the magnitude of the electron vorticity $\Omega_v = |\nabla \times V_e|$, the electron decoupling term (i.e., $E_{\text{dec,e}} = |(\mathbf{E} + \nabla \times \mathbf{B})|$), and the energy conversion term $J \cdot (\mathbf{E} + \mathbf{u}_e \times \mathbf{B})$. These variables are normalized between 0 and 1. This is a common requirement for many machine-learning estimators, since in the non-normalized data set the presence of outliers could alter the result. The $K$-means algorithm returns $K = 11$ clusters in the variable space, as we show in Table 1 for time $t \approx 247 \ [1/\Omega_v]$.

In Table 1 we report for each cluster the mean value of our variables (in dimensionless units) and the number of grid points, which belong to it. Among these clusters we choose to analyze the one with the highest mean value of the current density (since a necessary condition for reconnection to occur is the presence of a peak in the current density value), which is cluster “1.” This particular cluster in the variable space is made up of different structures in the physical space of our box. In Figure 3, top panel, we draw in the simulation $(x, y)$ space domain the shaded isocontours of the 11 clusters calculated by $K$-means (in the variable space). Cluster 1 is represented by the red color, the others are represented by the different blue variations. In the same figure, bottom panel, we draw the shaded isocontours of cluster 1 regions (red) superimposed to the contour shaded plots of the current density $|J|$, suggesting that cluster 1 roughly corresponds to the ensemble of CSs.

It is worth noting that other unsupervised methods are suitable for the same purpose such as, for instance, the $K$-medoids technique. In particular, by applying $K$-medoids at $t = 247$ we did not observed any significant improvement in the results. Therefore, $K$-means turns out to be a very good approach in this context, a baseline for this kind of analysis.

3. DBscan. From Figure 3 we see that cluster 1 in the variable space is composed by different structures in physical space. Thus we need another algorithm able to distinguish different structures using their location in physical space. A technique that is suited for this scope is DBscan, which takes as input the $x$- and $y$-coordinates of the points belonging to cluster 1. We take as searching radius for DBscan $\epsilon = 50$ grid points that correspond to about $5d_i$, and $\text{Min}_\text{pts} = 100$ points, as the minimum number of points for a single structure, where $\epsilon$ and $\text{Min}_\text{pts}$ are intrinsic variables of the DBscan algorithm. The results obtained by using DBscan are shown in...
Figure 4 where the different colors represent the different structures that are identified through the DBscan algorithm applied to cluster “1” of Table 1.

Table 1

| Cluster | $\mathcal{J}$ | $\mathcal{V}$ | $\mathcal{W}$ | $\mathcal{J}_{\text{AW}}$, $10^{-2}$ | $\mathcal{J}_{\text{AW, plane}}$ | $\mathcal{J} - (\mathcal{L} + \mathcal{V} + \mathcal{B})$, $10^{-3}$ | Grid Point Number |
|---------|-------------|-------------|-------------|----------------|----------------|-----------------|-----------------|
| 1       | 0.369       | 0.253       | 1.239       | 0.031          | 0.138          | 0.930           | 37,776          |
| 2       | 0.059       | 0.125       | 0.162       | 0.021          | 0.402          | 0.170           | 160,928         |
| 3       | 0.036       | 0.239       | 0.049       | 0.022          | 0.109          | 0.079           | 593,803         |
| 4       | 0.033       | 0.173       | 0.055       | 0.021          | 0.242          | 0.060           | 660,889         |
| 5       | 0.027       | 0.078       | 0.045       | 0.021          | 0.286          | 0.053           | 685,840         |
| 6       | 0.031       | 0.145       | 0.040       | 0.021          | 0.049          | 0.062           | 813,751         |
| 7       | 0.025       | 0.153       | 0.031       | 0.021          | 0.168          | 0.041           | 1,165,723       |
| 8       | 0.022       | 0.063       | 0.023       | 0.021          | 0.072          | 0.039           | 1,243,427       |
| 9       | 0.022       | 0.069       | 0.029       | 0.021          | 0.207          | 0.039           | 1,278,303       |
| 10      | 0.022       | 0.070       | 0.024       | 0.021          | 0.144          | 0.037           | 1,325,111       |
| 11      | 0.021       | 0.128       | 0.022       | 0.021          | 0.108          | 0.034           | 1,471,633       |

Note. For each cluster, identified by an index, we report the mean value of our variables and the number of grid points that belong to it. In bold the “interesting” cluster with the highest value of mean current density.

Figure 3. Top panel: the regions that constitute all 11 clusters in the variable space; cluster “1” in red, and all other clusters in various shades of blue (please note that each of the colors refers to a physical variable). Bottom panel: the regions that constitute cluster “1” (the interesting one) in the variable space, in red, superimposed to the contour plot of current density $|\mathcal{J}|$.

Figure 4. Contour plot of current density $|\mathcal{J}|$. Displayed in different colors are the different structures identified through the DBscan algorithm applied to cluster “1” of Table 1.

3.2. A1

A1 is the simplest algorithm to identify structures. It is based on two steps:

1. A threshold on the current density $|\mathcal{J}|$ defined as $\mathcal{J}_{\text{thr}} = \sqrt{\langle \mathcal{J}^2 \rangle} + 3\sigma$, where $\sigma = \sqrt{\langle \mathcal{J}^4 \rangle - \langle \mathcal{J}^2 \rangle^2}$, see

3. Threshold on the AR of structures found. For each structure found at step 3 we compute the AR (see Section 3, Equation (4)). We consider different possible thresholds for the AR to achieve better performance. In particular we have tried the following values: 10, 12.5, 20, 30, 50, and 70.
Table 2  
Summary of the Steps for Each Method

| Method | Steps |
|--------|-------|
| AML    | 0) selection of physical variables we want to use  
1) preprocessing: tuning of “K” value for the K-means  
2) K-means on variable space  
3) DBscan on physical space  
4) threshold on aspect ratio |
| A1     | 1) threshold on the current density | |
|        | } in the physical space  
2) threshold on the aspect ratio |
| A2     | 1) threshold on the current density | |
|        | } in the physical space  
2) request of points that overcome thresholds on | |
|        | } and | |
|        | } where the selected/excluded structures are all those possible reconnection regions whose ARs overcome/do not overcome the AR threshold as schematized in the flowchart in Figure 5. |

Zhdankin et al. (2013). More precisely, we select all regions in the physical space where the current density overcomes the threshold \( J_{thr} \). However, since a region of enhanced current is a necessary condition but not a sufficient one for reconnection, we add a second step.

(2) A threshold on the AR of the structures in order to select the most likely reconnecting structures. As done in AML, we consider different possible threshold values on AR: 10, 12.5, 20, 30, 50, and 70.

3.3. A2

A2 is a refinement of A1. Another step is added in order to increase the precision in finding reconnection events. The method is summarized as follows.

(1) As in A1 we fix a threshold on the current density and select the regions that overcome this threshold.

(2) We look at each structure and we select only the ones that include some points (at least two) that overcome both a threshold over electron vorticity \( \Omega_e \) and over \( E_{dec,e} \). In particular these two thresholds are defined as \( \Omega_{e,thr} = \sqrt{\langle \Omega_e^2 \rangle + 3 \sigma_\Omega} \), where \( \sigma_\Omega = \sqrt{\langle \Omega_e^4 \rangle - (\langle \Omega_e^2 \rangle)^2} \), and \( E_{dec,e,thr} = \sqrt{\langle E_{dec,e}^2 \rangle + 3 \sigma_{E_{dec}}} \), where \( \sigma_{E_{dec}} = \sqrt{\langle E_{dec,e}^4 \rangle - (\langle E_{dec,e} \rangle)^2} \).

(3) Finally, as in A1, we set a threshold on the AR of the selected structures. Also in this case, we consider different possible values for the AR: 10, 12.5, 20, 30, 50, and 70.

In Table 2 we give a summary of the steps of each method, while in Figure 5 we show a flowchart to sketch the selection procedure to detect possible reconnecting structures.

4. Results

Once three sets of candidate reconnecting structures have been selected using the three different methods (see the flowchart in Figure 5), we estimate the accuracy of these techniques by looking at candidate sites one by one. In particular, we check if reconnection is going on in each single site by looking to see if typical signatures of reconnection are present: inversion of the in-plane magnetic field, X-point configuration of the magnetic flux \( \Psi \), converging electron inflows toward the X-point and diverging outflows, magnetic fluctuation along the guide-field direction, peaked electron decoupling and energy conversion terms.

Now, to quantitatively compare the performances of the three algorithms AML, A1, and A2 for different values of the AR threshold. We define the following quality parameters:

\[
\text{Precision} = \frac{\# \text{ reconnecting structures among selected structures}}{\# \text{ selected structures}}
\]

\[
\text{nMR-precision} = \frac{\# \text{ nonreconnecting structures among excluded structures}}{\# \text{ excluded structures}},
\]

where the selected/excluded structures are all those possible reconnection regions whose ARs overcome/do not overcome the AR threshold as schematized in the flowchart in Figure 5.

The best algorithm performance is obtained when both precision and nMR-precision (which stands for non-magnetic-reconnection precision) are \( \sim 1 \). Indeed, in this case we would have, at the same time, that the algorithm has selected only reconnection sites and has not excluded reconnection sites.

In Tables 3–5 we show the values of these quality parameters for AML, A1, and A2, respectively, and for different AR thresholds. The results are shown at five time instants of our simulation: \( t_1 \sim 20 \ [1/\Omega_{ci}] \) (initial phase, no evidence of CS structures), \( t_2 \sim 230 \ [1/\Omega_{ci}] \), \( t_3 \sim 247 \ [1/\Omega_{ci}] \), \( t_4 \sim 282 \ [1/\Omega_{ci}] \) (CS developed regime), and \( t_5 \sim 494 \ [1/\Omega_{ci}] \) (fully developed turbulence). We will refer to these three distinct phases as phases I, II, and III. In the last two columns we list the mean value of our quality parameters during phase II, \( t = t_2, t_3, t_4 \) and during phase II and III, \( t = t_2, t_3, t_4, t_5 \). Looking at Table 3 corresponding to AML we observe (a) precision increases when AR threshold increases, (b) nMR-precision decreases when AR threshold increases, meaning that we are loosing “good” (reconnecting) sites, and (c) precision worsens when we include our analysis phase III (turbulent regime). The same conclusions are drawn by looking at Tables 4 and 5, corresponding to methods A1 and A2. In order to ease the comparison for the different performances, we plot in Figure 6 precision (solid red, dashed orange, and dotted brown lines, AML, A1, and A2, respectively) and nMR-precision (solid blue, dashed cyan and dotted violet lines, AML, A1, and A2, respectively) averaged over phase II (left panel) and over phase II and III (right panel) as a function of the AR threshold. By comparing these plots, left panel, we see that the unsupervised machine-learning algorithm AML turns out to be strongly competitive with respect to the methods based on thresholds on standard proxies. In particular precision performs better for AML rather than for A1. For what concerns A2, we see that the precision appears to be less influenced by the AR threshold variations. In particular, the precision of A2 (brown dotted line) turns out to be better at small AR thresholds than for AML; on the other hand, the nMR-precision (dotted violet line) worsens, so that lots of “good” reconnection sites are excluded. The same holds when including in the average the values at \( t = t_5 \) (fully developed turbulent phase), see the right panel of Figure 6. However, as a quite general result, precision and nMR-precision worsen a bit when considering the full turbulent phase. This worsening is the consequence of the dynamics driven by the turbulence, which advects, shrinks, breaks, and deforms the CSs. This dynamics together with the merging of nearby structures typical of a 2D geometry makes much more difficult a correct estimation of the typical width and length. Moreover, turbulence also creates
Figure 5. Flowchart for the three methods we are considering.

Table 3

| Tempo $[1/\Omega_c]$ | 20  | 230 | 247 | 282 | 494 | Mean (3t) | Mean (4t) |
|----------------------|-----|-----|-----|-----|-----|-----------|-----------|
| N. structures        | 35  | 29  | 19  | 24  | 32  |           |           |
| N. structures AR > 10| 0   | 17  | 17  | 21  | 28  | 0.71      | 0.64      |
| AR > 12.5            | 0   | 16  | 17  | 20  | 25  | 0.74      | 0.64      |
| AR > 20              | 0   | 14  | 15  | 14  | 19  | 0.79      | 0.7       |
| AR > 30              | 0   | 10  | 10  | 13  | 12  | 0.82      | 0.72      |
| AR > 50              | 0   | 6   | 9   | 12  | 7   | 0.92      | 0.79      |
| AR > 70              | 0   | 3   | 7   | 6   | 4   | 0.94      | 0.89      |
| Precision AR > 10    | …   | 0.65| 0.82| 0.67| 0.43| 0.71      | 0.64      |
| AR > 12.5            | …   | 0.69| 0.82| 0.7 | 0.36| 0.74      | 0.64      |
| AR > 20              | …   | 0.79| 0.80| 0.79| 0.42| 0.79      | 0.7       |
| AR > 30              | …   | 0.8 | 1   | 0.77| 0.33| 0.82      | 0.72      |
| AR > 50              | …   | 1   | 1   | 0.75| 0.43| 0.92      | 0.79      |
| AR > 70              | …   | 1   | 1   | 0.83| 0.75| 0.94      | 0.89      |
| nMR-precision AR < 10| 1   | 0.92| 1   | 1   | 1   | 0.97      | 0.98      |
| AR < 12.5            | 1   | 0.92| 1   | 1   | 0.57| 0.97      | 0.87      |
| AR < 20              | 1   | 0.93| 0.5 | 0.7 | 0.69| 0.71      | 0.7       |
| AR < 30              | 1   | 0.79| 0.55| 0.64| 0.6 | 0.66      | 0.64      |
| AR < 50              | 1   | 0.74| 0.5 | 0.58| 0.64| 0.61      | 0.61      |
| AR < 70              | 1   | 0.65| 0.42| 0.5 | 0.68| 0.52      | 0.56      |

Note. The results are shown for five different times of our simulation: $t \sim 20\, [1/\Omega_c]$, $t \sim 230\, [1/\Omega_c]$, $t \sim 247\, [1/\Omega_c]$, $t \sim 282\, [1/\Omega_c]$, and $t \sim 494\, [1/\Omega_c]$. In the last two columns we report the mean value of our quality parameters for the three central times (230, 247, and 282) and for four times (230, 247, 282, and 494). In the second row we list the number of structures found at the end of the third step; in the third to eighth rows we give the number of detected structures overcoming the specified AR threshold. In the ninth to fourteenth rows we give the values of precision for different AR threshold computed among the structures enumerated in the third to eighth rows; finally, in the fifteenth to twentieth rows we give values for nMR-precision.
regions with sharp variations in current density where, however, reconnection is not occurring, thus “confusing” the various methods and decreasing their accuracy. Except for this worsening in precision at the end of the simulations, all methods perform well both during the initial phase I when, correctly, no reconnection sites are detected, as well as during the central phase II when the CSs are well detected. According to us, it is possible to define an optimal AR threshold for AML methods.

### Table 4

| Tempo [1/Ωci] | 20 | 230 | 247 | 282 | 494 | Mean (3r) | Mean (4r) |
|---------------|----|-----|-----|-----|-----|----------|----------|
| N. structures | 11 | 45  | 47  | 52  | 454 |
| AR > 10       | 0  | 22  | 23  | 18  | 88  |
| AR > 12.5     | 0  | 20  | 20  | 17  | 70  |
| AR > 20       | 0  | 15  | 15  | 12  | 43  |
| AR > 30       | 0  | 13  | 12  | 10  | 26  |
| AR > 50       | 0  | 11  | 9   | 7   | 11  |
| AR > 70       | 0  | 6   | 6   | 5   | 4   |
| precision     | AR > 10 | ... | 0.68 | 0.61 | 0.61 | 0.26 | 0.63 | 0.54 |
|               | AR > 12.5 | ... | 0.75 | 0.7  | 0.59 | 0.27 | 0.68 | 0.58 |
|               | AR > 20 | ... | 0.8  | 0.8  | 0.75 | 0.33 | 0.78 | 0.67 |
|               | AR > 30 | ... | 0.77 | 0.83 | 0.8  | 0.35 | 0.8  | 0.69 |
|               | AR > 50 | ... | 0.82 | 1    | 0.86 | 0.36 | 0.89 | 0.76 |
|               | AR > 70 | ... | 1    | 1    | 0.8  | 0.25 | 0.93 | 0.76 |
| nMR-precision | AR < 10 | 1   | 0.74 | 0.71 | 0.65 | 0.90 | 0.7  | 0.75 |
|               | AR < 12.5 | 1   | 0.76 | 0.74 | 0.63 | 0.90 | 0.71 | 0.76 |
|               | AR < 20 | 1   | 0.7  | 0.72 | 0.65 | 0.89 | 0.69 | 0.74 |
|               | AR < 30 | 1   | 0.66 | 0.68 | 0.64 | 0.88 | 0.66 | 0.71 |
|               | AR < 50 | 1   | 0.65 | 0.68 | 0.62 | 0.88 | 0.65 | 0.71 |
|               | AR < 70 | 1   | 0.61 | 0.63 | 0.60 | 0.87 | 0.61 | 0.68 |

**Note.** The results are shown for five different times of our simulation: \( t \sim 20 \, [1/\Omega ci], \ t \sim 230 \, [1/\Omega ci], \ t \sim 247 \, [1/\Omega ci], \ t \sim 282 \, [1/\Omega ci], \) and \( t \sim 494 \, [1/\Omega ci]. \) In the last two columns we report the mean value of our quality parameters for the three central times (230, 247, and 282) and for four times (230, 247, 282, and 494). In the second row we list the number of structures found at the end of the third step; in the third to eighth rows we give the number of detected structures overcoming the specified AR threshold. In the ninth to fourteenth rows we give the values of precision for different AR threshold computed among the structures enumerated in the third to eighth rows; finally, in the fifteenth to twentieth rows we give values for nMR-precision.

### Table 5

| Tempo [1/Ωci] | 20 | 230 | 247 | 282 | 494 | Mean (3r) | Mean (4r) |
|---------------|----|-----|-----|-----|-----|----------|----------|
| N. structures | 2  | 30  | 27  | 26  | 126 |
| AR > 10       | 0  | 21  | 18  | 12  | 52  |
| AR > 12.5     | 0  | 19  | 17  | 11  | 43  |
| AR > 20       | 0  | 16  | 14  | 10  | 29  |
| AR > 30       | 0  | 13  | 11  | 9   | 20  |
| AR > 50       | 0  | 11  | 9   | 7   | 8   |
| AR > 70       | 0  | 6   | 6   | 5   | 3   |
| precision     | AR > 10 | ... | 0.76 | 0.78 | 0.92 | 0.33 | 0.82 | 0.70 |
|               | AR > 12.5 | ... | 0.84 | 0.82 | 0.91 | 0.33 | 0.86 | 0.72 |
|               | AR > 20 | ... | 0.81 | 0.86 | 0.9  | 0.38 | 0.86 | 0.74 |
|               | AR > 30 | ... | 0.77 | 0.91 | 0.89 | 0.4  | 0.86 | 0.74 |
|               | AR > 50 | ... | 0.82 | 1    | 0.86 | 0.37 | 0.89 | 0.76 |
|               | AR > 70 | ... | 1    | 1    | 0.8  | 0.33 | 0.93 | 0.78 |
| nMR-precision | AR < 10 | 1   | 0.89 | 0.44 | 0.64 | 0.81 | 0.66 | 0.69 |
|               | AR < 12.5 | 1   | 0.91 | 0.5  | 0.6  | 0.79 | 0.67 | 0.7  |
|               | AR < 20 | 1   | 0.71 | 0.46 | 0.56 | 0.79 | 0.58 | 0.63 |
|               | AR < 30 | 1   | 0.59 | 0.44 | 0.53 | 0.78 | 0.52 | 0.58 |
|               | AR < 50 | 1   | 0.58 | 0.44 | 0.47 | 0.77 | 0.5  | 0.56 |
|               | AR < 70 | 1   | 0.54 | 0.38 | 0.43 | 0.76 | 0.45 | 0.53 |

**Note.** The results are shown for five different times of our simulation: \( t \sim 20 \, [1/\Omega ci], \ t \sim 230 \, [1/\Omega ci], \ t \sim 247 \, [1/\Omega ci], \ t \sim 282 \, [1/\Omega ci], \) and \( t \sim 494 \, [1/\Omega ci]. \) In the last two columns we report the mean value of our quality parameters for the three central times (230, 247, and 282) and for four times (230, 247, 282, and 494). In the second row we list the number of structures found at the end of the third step; in the third to eighth rows we give the number of detected structures overcoming the specified AR threshold. In the ninth to fourteenth rows we give the values of precision for different AR threshold computed among the structures enumerated in the third to eighth rows; finally, in the fifteenth to twentieth rows we give values for nMR-precision.
and A1, which corresponds to a compromise between having a good precision and not losing lots of good reconnection sites. For AML, this optimal AR threshold corresponds to the one at the intersection between the plot of precision (left panel of Figure 6, red solid line) and the plot of nMR-precision (left panel of Figure 6, blue solid line). In this case, we get AR \( \approx 18 \), which gives a precision and nMR-precision accuracy of the order of 78%. For A1, the same is obtained for AR \( \approx 14 \) with a score of \( \approx 71\% \). In both cases AR values are in agreement with theoretical estimations. For A2, it is not possible to make the same estimation of the optimal AR since the two plots do not intersect; this is a limitation of A2. Finally, in our simulation the performance of the AML method improves a bit (reaching 80%) if we use, in the variable set, \( J \cdot (E + u_e \times B + \nabla \phi / ne) \) instead of \( J \cdot (E + u_e \times B) \). This can be due to the fact that in our model we consider a scalar pressure; thus, in the absence of off-diagonal terms in the pressure tensor, this term does not contribute to the effective ohmic heating (Birn & Hesse 2010).

### 5. Conclusions

In summary, we developed three different methods aimed at automatically detecting reconnecting CSs in a 2D simulation of turbulence. AML uses unsupervised machine-learning techniques for the scope, while methods A1 and A2 use only use thresholds on the standard reconnection proxies. All methods are based on a threshold on the structures’ AR since the AR is directly linked to the reconnection rate and is a physical adimensional parameter independent from all simulation setups. We have defined two different quality parameters in order to quantitatively estimate the performance of our algorithms: precision, which measures the capability of the method to select good reconnection sites, and nMR-precision, which gives us information about excluded regions, in particular how many of them are not really reconnecting sites. This parameter provides information about good sites that have been excluded. AML is the one that performs the best. Moreover, although A2 has better precision at the smallest AR threshold, for AML, nMR-precision is better, and it is possible to define an optimal AR for it (as the intersection between the plot of precision and nMR-precision), which cannot be done for A2. The impossibility to define an optimal AR for A2, together with the bad values of nMR-precision for this method (lots of good sites lost), brings us to the conclusion that A2 is not the best method to be applied. Considering only time periods where CSs are formed but still not interacting or significantly affected by the turbulent dynamics, we found for AML an optimal AR of about 18, which gives both a precision and an nMR-precision of \( \sim 78\% \). These average performances worsen a bit if we consider also times at developed turbulence. This worsening is expected since at the fully developed turbulent phase the dynamics advects, shrinks, breaks, and deforms the CSs. Moreover, CS merging could lead to problems in separating different peaks and developed turbulence creates regions with sharp variations in current density where, however, reconnection is not occurring, thus “confusing” the automatic methods and decreasing their accuracy. Despite this worsening during the fully developed turbulent phase, the automatic methods (and in particular the AML one) are still valid tools for speeding up reconnection site identification. In principle they could be applied to any plasma simulation including multiple potential reconnection sites, efficiently providing a list of “candidate” sites for a detailed human-based reconnection analysis.

These automatic methods are not based on the analysis of images as, for instance, 2D charts of simulation data, but use physical, measured variables as relevant signatures of reconnection. These quantities, here taken from numerical simulations, are also available in satellite data sets. We are presently working on adapting these methods to the analysis of data collected on 1D trajectories, more precisely cuts in the

---

**Figure 6.** Plots of the values of precision (red, orange, and brown lines) and nMR-precision (blue, cyan, and violet lines) averaged over the three central times (left panel) and over four times (right panel), as a function of the AR threshold, for the three different methods we are analyzing: solid line, AML; dashed line, A1; dotted line, A2.
simulation or temporal series in the case of a satellite surveying plasma/magnetic structures. Such an upgrade would greatly improve the analysis of magnetic reconnection in satellite data, quickly highlighting interesting temporal intervals with active reconnection randomly scattered in the huge amount of data produced by past and present missions.

We also stress the importance of setting a threshold for the CS’s AR for better performance of our methods. Indeed the aspect ratio of a reconnecting current sheet is a physical, adimensional, simulation-independent parameter directly linked to the reconnection process. The increasing of the precision with the AR threshold value is linked to the fact that once the characteristic length of the reconnecting structure is fixed (by the boundaries of adjacent flux ropes), reconnection is able to effectively shrink its width, leading to increasing values of aspect ratio. However, increasing the AR threshold too much above the theoretical estimation would lead to incorrect results since excessive thinning corresponds to a strongly nonlinear phase eventually leading to the site breaking into smaller magnetic islands. In fact nMR-precision worsens with too large AR threshold as reconnection itself limits the AR of the CS. As a consequence, we propose an optimal AR threshold for which both precision and nMR-precision perform well. We conclude by saying that the AML method, based on the machine-learning approach, turns out to be the most performing one.

In the context of the EU AIDA project, we are presently working to set up a tool, namely Unsupervised ML Reconnection, in Python language for the AML method. This tool will become available, free-to-use, in the online AIDA-repository at https://gitlab.com/aidaspace/aidapy aimed at being of utility for the (space) plasma physics community. Today, some preliminary scripts are available at https://gitlab.com/aidaspace/aidapy/-/tree/unsupervised_reconnection/aidapy// WIP-unsupervised_reconnection, and also on Zenodo at https://doi.org/10.5281/zenodo.4282289. Finally, the simulation data set (TURB 2D) is available at Cineca AIDA-DB. In order to access the meta-information and the link to “TURB 2D” simulation data see the tutorial at http://aida-space.eu/AIDAdb-iRODS.

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 776262 (AIDA, www.aida-space.eu). Numerical simulations discussed here have been performed on Marconi at Cineca (Italy) under the IS CRA initiative. F.C. thanks Dr. M. Guarrasi (Cineca, Italy) for his essential contribution to the implementation of the kinetic Vlasov code on Marconi supercomputer at Cineca (Italy).

References

Birn, J., & Hesse, M. 2010, PhPl, 17, 012109
Borogogn, D., Califano, F., Faganello, M., & Pegoraro, F. 2015, PhPl, 22, 032301
Califano, F., Cerri, S. S., Faganello, M., & et al. 2020, PhP, 8, 317
Camporeale, E., Sorriso-Valvo, L., Califano, F., & Retino, A. 2018, PhRvL, 120, 125101
Cassak, P. A. 2016, SpWea, 14, 186
Cassak, P. A., Liu, Y.-H., & Shay, M. 2017, JPP, 83, 71583051
Cerri, S. S., & Califano, F. 2017, NJPh, 19, 025007
Daughton, W., Nakamura, T. K. M., Karimabadi, H., Roytershteyn, V., & Loring, B. 2014, PhPl, 21, 052307
Dunlop, M. W., Balogh, A., Glassmeier, K. H., & Robert, P. 2002, JGRA, 107, 1384
Dupuis, R., Goldman, M. V., Newman, D. L., Amaya, J., & Lapenta, G. 2020, ApI, 889, 22
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. Second Int. Conf. on Knowledge Discovery and Data Mining, ed. E. Simoudis, J. Han, & U. Fayyad (Menlo Park, CA: AAAI Press), 226
Fadanelli, S., Lavraud, B., Califano, F., et al. 2019, JGRA, 124, 6850
Faganello, M., & Califano, F. 2017, JIPhPh, 83, 535830601
Faganello, M., Califano, F., Pegoraro, F., & Retino, A. 2014, EL, 107, 19001
Furth, H. P., Killeen, J., & Rosenbluth, M. N. 1963, PhPl, 6, 459
Harris, E. G. 1962, NCim, 23, 115
Haynes, C. T., Burgess, D., & Camporeale, E. 2014, ApI, 783, 38
Henri, P., Califano, F., Faganello, M., & Pegoraro, F. 2012, PhPl, 19, 072908
Hu, A., Sisti, M., Finelli, et al. 2020, ApI, 900, 86
Karimabadi, H., Daughton, W., & Quest, K. 2005, JGRA, 110, A03214
Karimabadi, H., Roytershteyn, V., Wan, M., et al. 2013, PhPl, 20, 012303
MacQueen, J. B. 1967, in Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability, ed. L. M. Le Cam & J. Neyman (Berkeley, CA: Univ. California Press), 281
Mangeney, A., Califano, F., Cavazzoni, C., & Travnicek, P. 2002, JCoPh, 179, 495
Osman, K., Matthaeus, W., Gosling, J., et al. 2014, PhRvL, 112, 215002
Parker, E. N. 1957, JGr, 62, 509
Priest, E. R. 1982, Solar Magnetohydrodynamics (Dordrecht: Springer)
Priest, E. R., & Forbes, T. G. 2001, JIPhPh, 66, 363
Retino, A., Sundkvist, D., Vaiavds, A., et al. 2007, NatPh, 3, 235
Rhys, H. I. 2020, Machine Learning with R, the tidyverse, and mlr (Shelter Island, NY: Manning Publications)
Servidio, S., Matthaeus, W., Shay, M., et al. 2010, PhPl, 17, 032315
Sisti, M., Faganello, M., Califano, F., & Lavraud, B. 2019, GeoRS, 46, 11597
Sweet, P. A. 1958, in IAU Symp. 6, Electromagnetic Phenomena in Cosmical Physics, ed. B. Lehnhert (Cambridge: Cambridge Univ. Press), 123
Valentini, F., Califano, F., Hellinger, P., & Mangeney, A. 2007, JCoPh, 225, 753
White, R. 1986, RvMP, 58, 183
Zenitani, S., Hesse, M., Klimas, A., & Kuznetsova, M. 2011, PhRvL, 106, 195003
Zhdankin, V., Uzdensky, D., Perez, J., & Boldyrev, S. 2013, ApI, 771, 124