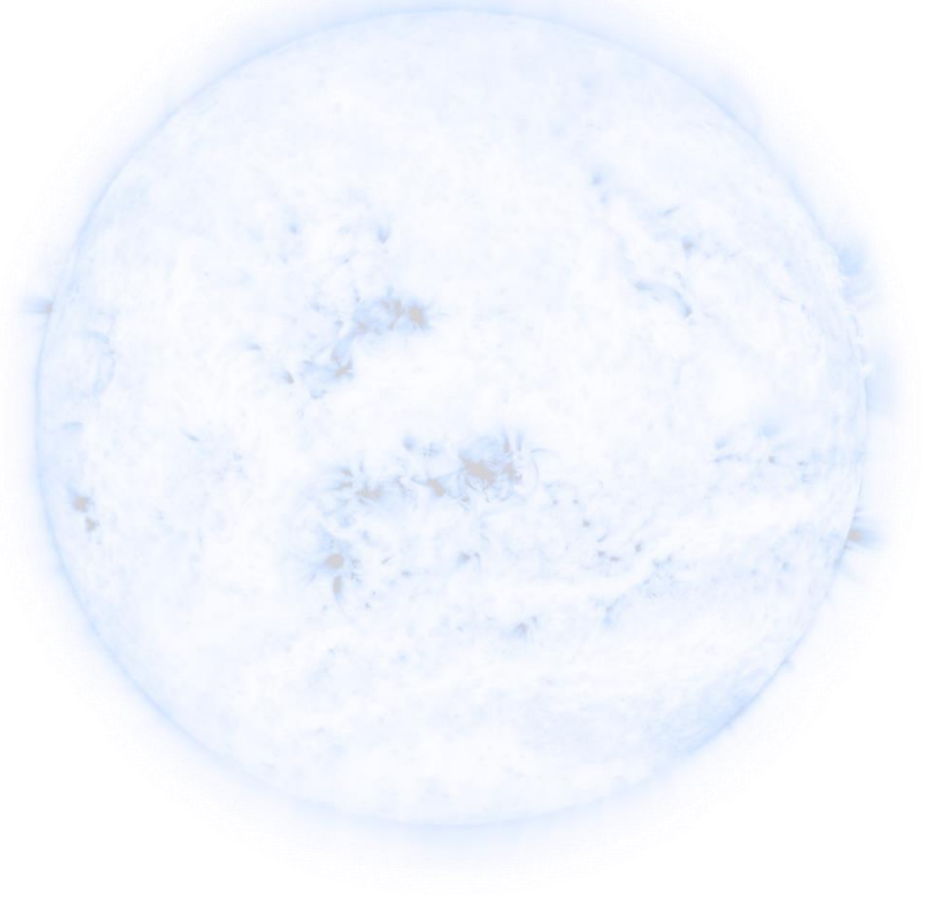
Cluster analysis of spectroscopic line profiles and EUV emission in RMHD simulations and observations of the solar atmosphere

Viacheslav Sadykov[1,2,3], Irina Kitiashvili[1,2], Alexander Kosovichev[2,3]

[1]BAERI, [2]NASA ARC, [3]NJIT



Spatially-resolved observations from the IRIS, SDO/AIA, and other space mission and ground-based telescopes, coupled with realistic 3D RMHD simulations, are a powerful tool for analysis of processes in the solar atmosphere. To better understand the dynamical and thermodynamic properties in the simulation data and their connection to observations, it is essential to determine similarities in the behaviors of the synthesized and observed emission. However, the complexity of observational data and physical processes makes comparison of observations and modeling results difficult. In this work, we show the initial results of application of K-Means clustering (unsupervised machine learning) algorithm to two different problems: 1) recognition of the typical spectroscopic line profiles observed by IRIS during solar flares and their typical dynamic behavior; 2) recognition of shocks and heating events in synthetic AIA emission data obtained from StellarBox quiet-Sun simulations. The average silhouette width technique for the K-Means algorithm is utilized in different ways to obtain optimal numbers of clusters. We discuss application of the emission clustering to visualizations of the computational volume, understanding its evolutionary trends and behavior patterns, and inversion (reconstruction) of physical properties of the solar atmosphere from synthesizes emission data.

Description of data processing and clustering algorithms

- **Statistical moments of line profiles.** The zeroth moment will represent the maximum or the integrated line intensity. The first (Doppler shift), second (line width), third (line asymmetry), and higher statistical moments can be computed as:

$$S_k = \sqrt[k]{\int (\lambda - \langle\lambda\rangle)^k I(\lambda)d\lambda \Big/ \int I(\lambda)d\lambda}, \qquad \langle\lambda\rangle = \int \lambda I(\lambda)d\lambda \Big/ \int I(\lambda)d\lambda$$

- **K-Means clustering.** The K-Means takes the number of clusters as an input parameter, and initially seeds the cluster centers randomly among data points. After this, K-Means assigns the points to belong to the nearest cluster center (i.e. labels them), recomputes the cluster centers as the means among the points of the same labels, and repeats the procedure until there are no changes in labels for points.
- **Average silhouette width.** The silhouette is defined for data *i* point as $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, where *a(i)* is the average distance from the point *i* to points of the same cluster, *b(i)* is the average distance from the point *i* to points of another **closest** cluster. The average *s(i)* across the points indicates how well the points lye within their clusters.
  - The optimal number of clusters can be estimated by maximization of the *s(i)*.
  - When *s(i) < 0*, the points no longer "belong" to their clusters.

Identifying typical response of the upper chromosphere and lower transition region for the flare heating from IRIS observations

- The Interface Region Imaging Spectrograph (IRIS, De Pontieu et al. 2014) has observed hundreds of flares of ≥ C1.0. However, statistical studies of atmospheric response to the flare heating by IRIS are hard to perform because of the complexity of imaging spectroscopy data: their high dimensionality, large data volumes, optically-thick nature of the lines.
- Finding compact illustrative representation of the atmospheric response to the flare heating using unsupervised machine learning (clustering) techniques can simplify the analysis of large observational data sets and increase their understanding.
- An example of clustering of C II 1334.5 Å line profiles for the M1.0 flare of June 12, 2014, is presented in Figure 1. The maps of the line profile representatives indicate that most of the southern part of the flare region exhibits redshifts of the C II line profiles, while the northern part does not show any strong Doppler shift with respect to unperturbed gray line profile. Line clustering was previously used by Panos et al. (2018) and Sainz Dalda et al. (2019).
- Figure 2 illustrates the typical evolution of Mg II k 2796 Å line profiles during the M1.8 class solar flare of February 13, 2014. The K-Means clustering was performed simultaneously for the line intensity, Doppler shift, and line width evolution, with equal contribution from each considered statistical moment. The red, blue, and black clusters are of special interest: while red cluster behaves as typically expected during "explosive" chromospheric evaporation, blue and black clusters revealed slight redshift followed by a strong blueshift of the spectral lines.

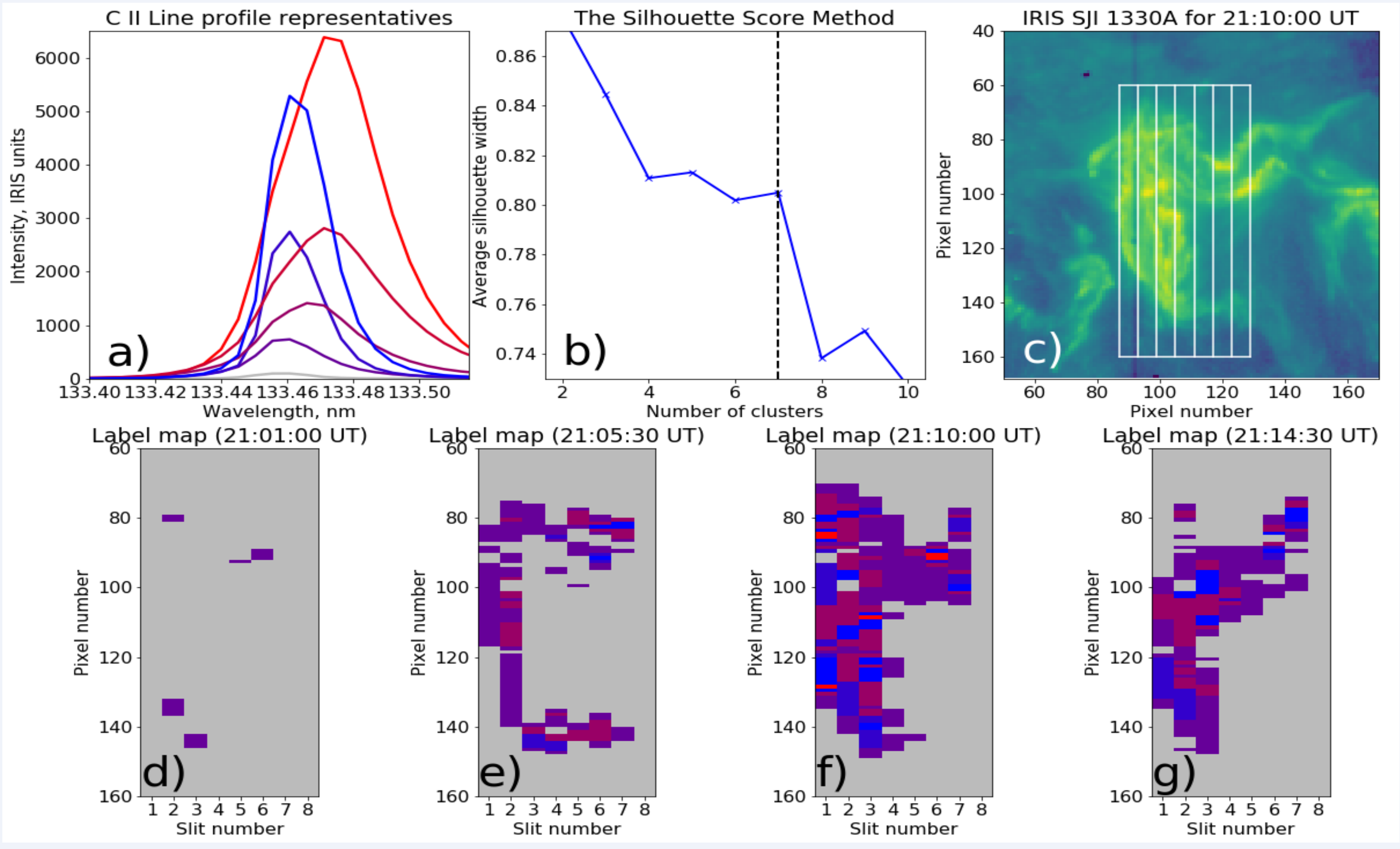


Figure 1. a) C II 1334.5 Å line profile representatives (cluster centers) for the M1.0 class solar flare observed on June 12, 2014; b) the average silhouette width as a function of number of cluster used for K-Means algorithm. Dashed black vertical line points out the optimal selected number of clusters (7); c) IRIS SJI 1330 Å image for the peak time of the flare. White lines point out slit positions for the panels d-g); d-g) maps of line profile representatives for different times of the flare. The color code is in accordance with the panel a).

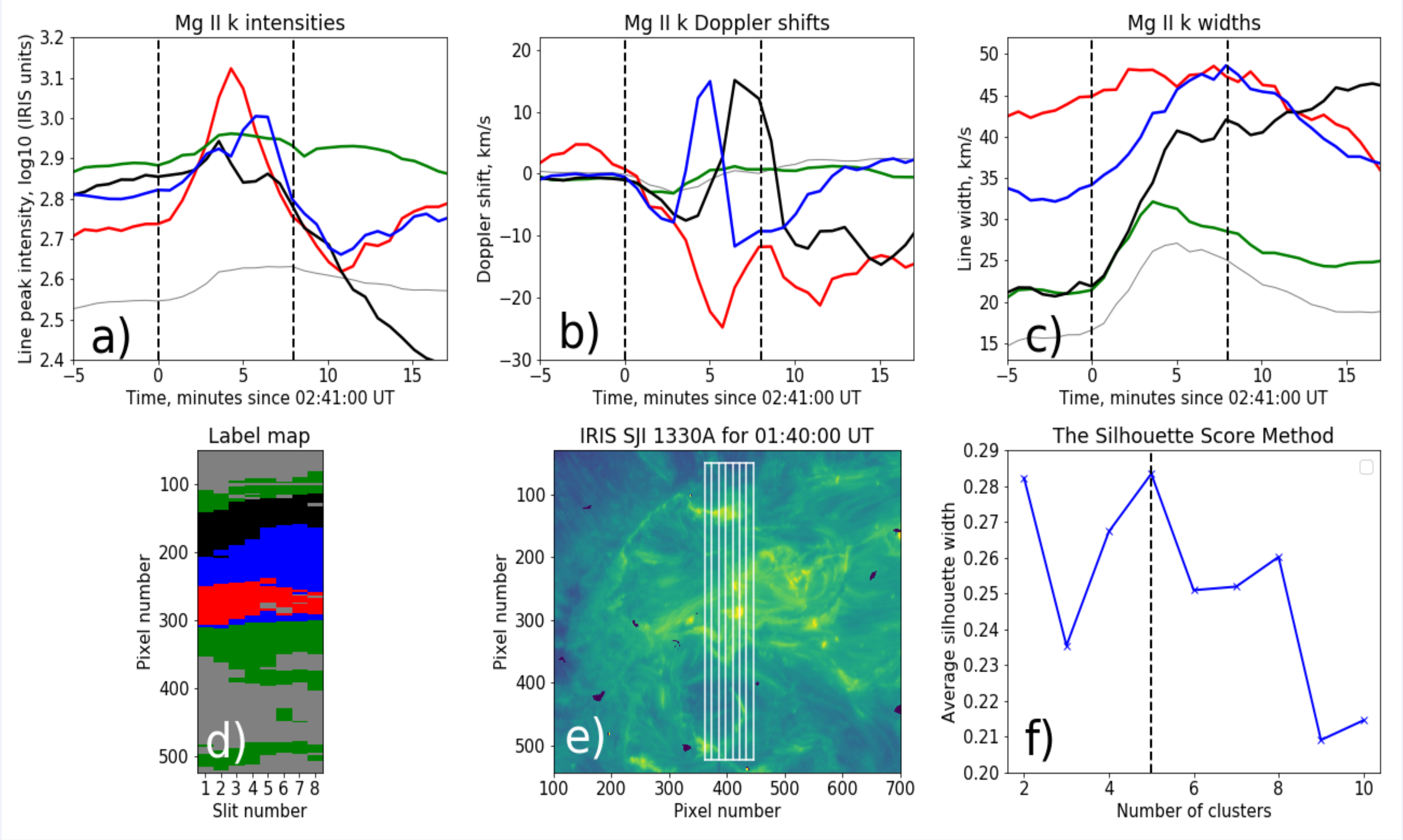


Figure 2. Typical behavior (cluster centers) of the Mg II k 2796 Å line intensity (a), Doppler shift (b) and line width (c) during M1.8 class flare of February 13, 2014. Dashed black vertical lines mark the flare start and flare peak times. (d) Map of behavior representatives for the flare. Color coding is kept in accordance with panels (a-c). (e) IRIS SJI 1330 Å image for the flare peak time. White lines point out slit positions for the panels (d). (f) the average silhouette width as a function of number of cluster used for K-Means algorithm. Dashed black vertical line points out the optimal selected number of clusters (5).

Recognition of shocks and heating events from synthetic AIA emission

- **"StellarBox" code** solves the fully compressible MHD equations with radiative transfer solved by ray-tracing and opacity binning techniques, and large-eddy simulation (LES) treatment of subgrid turbulent transport (Wray et al. 2018). The current version of the code supports option to extend of the computational domain to corona and deeper convective layers and in horizontal directions.
- **The computational domain** of 12.8 x 12.8 x 15.2 Mm includes a 10-Mm layer from the photosphere to the low corona. The grid-size is 25km in the horizontal directions; a variable grid-spacing of similar size is used in the vertical direction. The lateral boundary conditions are periodic. For initial conditions of the chromosphere and corona, the model by Vernazza et al. (1981) has been used. The 176 simulation snapshots delivered with 2s temporal cadence are analyzed.
- **Synthetic top-view AIA emission** is computed for each snapshot for each column separately, using SDO/AIA temperature response functions available from SSW IDL. Strong impacts ("shocks" hereafter) are observed in AIA running difference images.
- **K-Means** clustering is performed for sparse selection of columns and snapshots for all AIA channels together. The contribution of each channels was normalized. Seven clusters are used (the reason is explained below).
- **Preliminary result:** one of the clusters (cyan) correlates well with the shock signatures. The corresponding differential emission measure profile (DEM, cyan) shows the peak at ~1MK and contribution from ~400kK plasma.

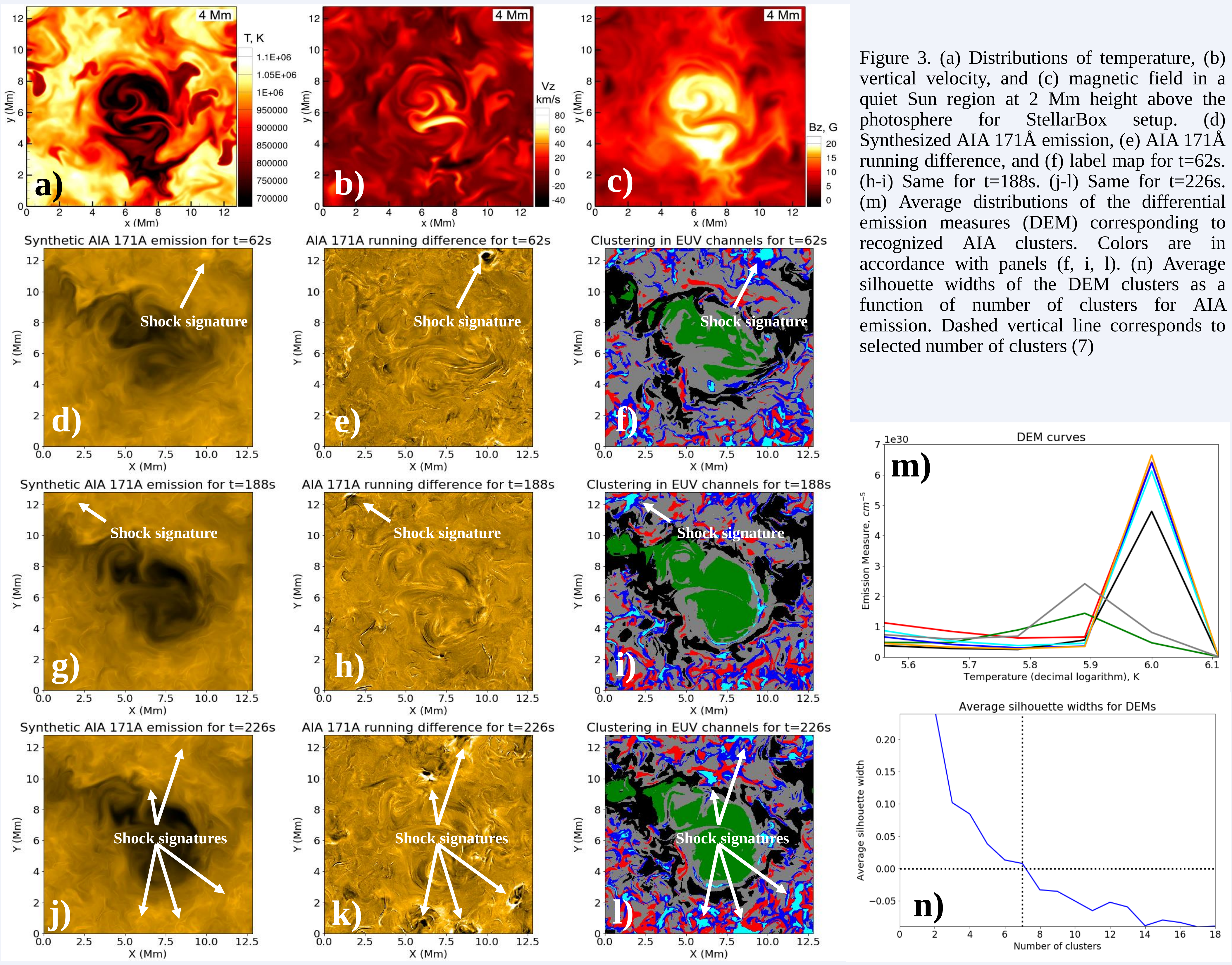
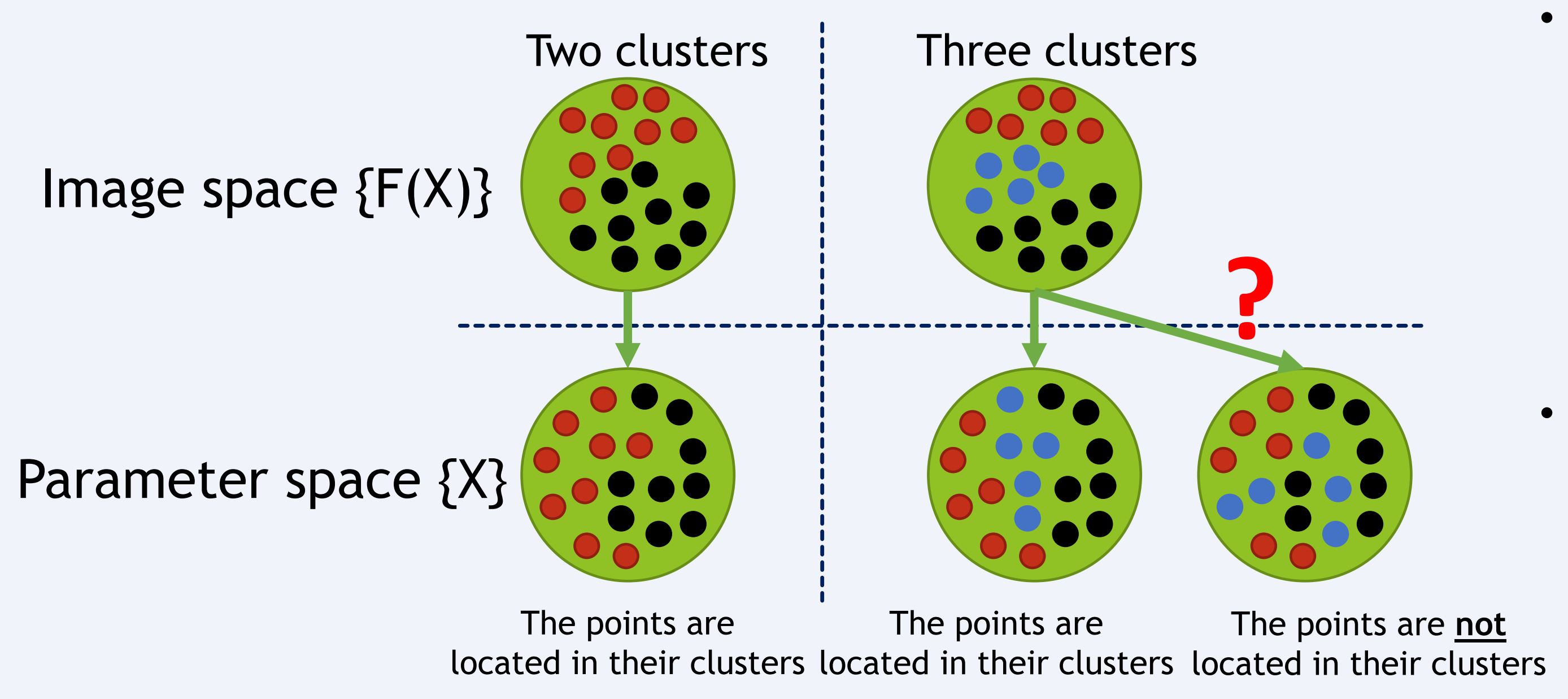


Figure 3. (a) Distributions of temperature, (b) vertical velocity, and (c) magnetic field in a quiet Sun region at 2 Mm height above the photosphere for StellarBox setup. (d) Synthesized AIA 171Å emission, (e) AIA 171Å running difference, and (f) label map for t=62s. (h-i) Same for t=188s. (j-l) Same for t=226s. (m) Average distributions of the differential emission measures (DEM) corresponding to recognized AIA clusters. Colors are in accordance with panels (f, i, l). (n) Average silhouette widths of the DEM clusters as a function of number of clusters for AIA emission. Dashed vertical line corresponds to selected number of clusters (7)

How many clusters to select? Inverse problem POV.

- The number of clusters depends on the problem type: some problems (e.g. recognition of tiny features in line profiles) may require selection of more clusters than dictated by the maximization of the average silhouette width.
- In the example above, synthetic AIA emission is a function of the Differential Emission Measure (DEM, Cheung et al. 2015) of the computational domain. Reconstruction of the DEM from AIA emission is an example of ill-posed inverse problem.



Two clusters

Three clusters

Image space {F(X)}

?

Parameter space {X}

The points are located in their clusters

The points are located in their clusters

The points are **not** located in their clusters

Figure 4. Illustration of how the clustering in image space can be reflected in parameter space. The lower right circle represents the case when identification of new cluster in image space does not help to locate the generating parameters in parameter space.

- Figure 3n illustrates that, if more than 7 clusters are selected in AIA emission space, the corresponding DEMs of these clusters strongly mix with each other. The average silhouette computed in DEM (generating parameter) space becomes negative, i.e. the points no longer correspond to their clusters in that space.
- In general, the combination of unsupervised clustering in image space and measure of how well the data is clustered in generating parameter space can increase understanding and provide diagnostics of any inverse problem solely based on the known forward modeling results (see Figure 4)

Future plans and ideas.

- Recognition of typical line profiles and dynamical responses of the atmosphere to flare heating from IRIS data based on large statistics of flare events.
- Correlation of appearance of certain line profile shapes and dynamical behavior with properties of hard X-rays (from RHESSI and Konus-WIND) and soft X-rays (from GOES).
- Computation of IRIS line profiles (Mg II, C II, Si IV) for the considered StellarBox run using RH radiative transfer code. Testing the clustering algorithms on the synthesized emission reduced to IRIS and SDO/AIA instrumental resolutions.
- Development of diagnostics tool for recognition of shocks, strong flows and heating events from SDO/AIA data and IRIS data based on cluster analysis discoveries.

References.

1. Cheung, M.C.M., Boerner, P., Schrijver, C. J. et al. 2019, ApJ, 807, 143
2. De Pontieu, B., Title, A.M., Lemen, J.R. et al. 2014, Solar Physics, 289, 2733
3. Panos, B., Kleint, L., Huwyler, C. et al. 2018, ApJ, 861, 62.
4. Sainz Dalda, A., de la Cruz Rodríguez, J., De Pontieu, B., and Gošić, M. 2019, ApJL, 875 L18.
5. Vernazza, J.E., Avrett, E.H., Loeser, R. 1981, ApJSS, 45, 635
6. Wray A.A., Bensassi K., Kitiashvili I.N. et al. 2018, In Book: "Variability of the Sun and Sun-like Stars: from Asteroseismology to Space Weather". Eds: J.-P. Rozelot, E.S. Babaev, EDP Sciences, p.39-62.