Novel Application of Statistical Tools for Big Data Analyses of Solar Physics

Lars K. S. Daldorff [*][†] Siavoush Mohammadi [‡] Brett Isham [§]

Abstract
Space plasma simulations are known to generate vast amounts of numerical data. In recent times,with greater availability of computer power, the proportion of data generated has increased exponentially posing new challenges in its analysis. Simulations can be scaled up, but matching methods for analysis have not been developed at the same pace, except in the industry setting where companies utilize big-data techniques to build advanced analysis engines. Although many of these methods need developers working with low-level big-data programming, today there exist powerful out-of-the box solutions that can be easily employed. In this paper, using SAS Visual Analytics as one of these tools, we will demonstrate how existing statistical tools and analytical platforms can be used to analyze simulations of the Sun in a novel way. In our case, simulations generated roughly 660mb/time unit and hundreds of them were needed to provide an adequate analysis. Thus, application of our method allows instant analysis and prevents the need to guess "when" and "where" the interesting physical phenomena occurs, thus effectively getting rid of data which has little or no scientific value.

Key Words: Big Data, Data extraction , Explorative analytics, Heliophysics, Sun

1. Introduction

What could a space plasma physicist have in common with a data warehouse consultant? At first glance, not much, but is it really so? If we try to remove space physics terminology, ignore fluid dynamics and Maxwells equations, and instead focus on what these types of scientists actually do with their data, it might not be that foreign after all. The fact is that both numerical-physicists and experimentalists are completely dependent on their data for new insights. The data source is usually numerical simulations based on a model, or as in the case of an experimentalist, the instruments with which they use in their experiment. Both cases however can produce really Big Data, huge even. Next, the physicist would prepare the data for analysis in different ways. If you have a data warehouse background, you might have just thought quietly for yourself "ETL?" (Extract-Transform-Load)[Kimball and Ross (2013)] and you would be right to think so, in some sense it is an ETL-process. Even though this might be true, data warehousing [Kimball and Ross (2013)] is used very little in this type of academic research. The question that comes to mind is: Can Big Data be a connection point between academia and the business world, where the two help each other learn new and old methods so that both parties can obtain insights quickly? Yes, we believe that it could happen.

With these reoccurring thoughts we started working together and asked ourselves a simple question: What would happen if we were to take numerical plasma simulations of the Sun and structure them in such a way that they could be uploaded into a Big data in-memory environment that used out-of-the-box analytical methods? The challenge that we faced at NASA was not in producing big data volumes, but in analyzing it effectively.

[*] NASA, Goddard Space Flight Center, Maryland, USA

[†] University of Michigan, Atmospheric, Oceanic, and Space Science, Michigan, USA

[‡] Infotrek, Stockholm, Sweden

[§] Interamerican University of Puerto Rico, Department of Electrical and Computer Engineering, Bayamón, Puerto Rico, USA

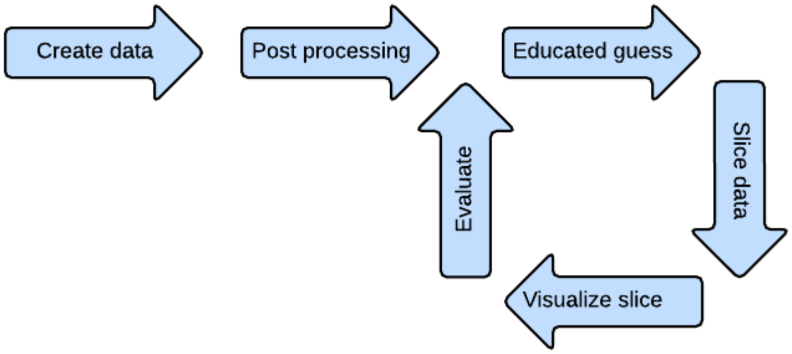


Figure 1: A simplified description of the existing process on how insights are drawn from scientific data.

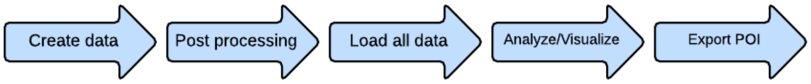


Figure 2: Simplified description of the new process that for creating/collecting data to insight , where we first load the entire data, automatically analyze and visualize all candidates of Point of Interest (POI) and then export the data for deeper analysis performed by the subject expert.

In the academic numeric world, the fast paced technical development and access to large super computer centers has meant that production of data can easily be scaled up. However, much of the data produced is of little or no scientific interest, it is simply already known physics or noise of different sort.

Basically this is a needle in a haystack situation, the phenomena of interest are buried somewhere in the data, without a clue as to where or even when in the data it can be found. At the same time, the visualization and analysis methods typically employed are time consuming. As a consequence of this, the researcher in question (in this case, a physicist) needs to slice the data by making qualified guesses as to where and when in the data the needle lies. A simplified description of the process is depicted in Figure 1. The problem with this process is that even if one is lucky and just happens to find an interesting phenomenon on the first guess, one can't be sure that this is the only point of interest in the data. This problem means that the time between the gathering of data (from numerical simulations of the sun in our case) to drawing insights of data becomes very long. But what if one didn't have to visualize the data in slices? What if we could take out the guesswork from the process? What if it were possible to upload all of the data at once onto a platform which would instantly point out the location of the needle (or needles) by employing standardized methods? What if, after locating the needle(s), the full analysis is be done only on the data of interest by an easy and simple export?

The phenomena that we wanted to study were simulations of the sun, or more specifically the magnetic arches associated with the sun's atmosphere, Figure 3, not limited to solar spots, flares and coronal mass ejections, which contribute to a considerable increase in the X-ray and ultraviolet radiation from the outer solar atmosphere (and hence into the upper atmosphere of the earth) and how these arches arise. This phenomena has been observed for many years from the earth, as well as from satellites like the NASA mission

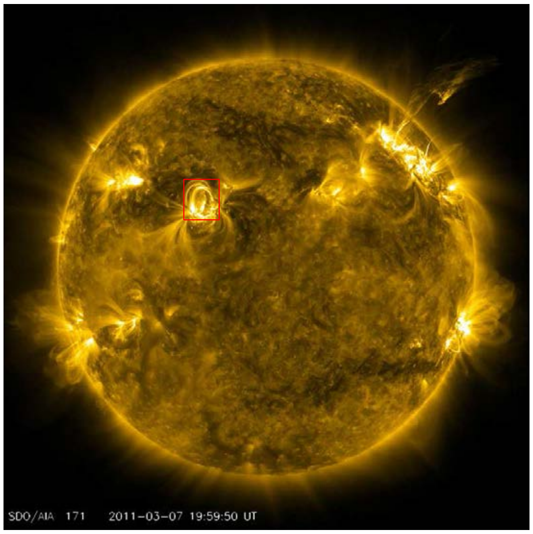


Figure 3: The background picture of the sun is taken by the SDO/AIA (Atmospheric Imaging Assembly) [Pesnell et al. (2012); Lemen et al. (2012)] and show the dynamics of the coronal plasma which has a temperature close to 0.6 million Kelvin. The figure is showing the plasma flows as it follows the magnetic fields structure on the Sun. In our models we are interested in the interaction of different field lines in a magnetic loop as marked by the red box. We want to see how the different magnetic field line bundles in the loop structure develop in time as shown in Figure 4. Most of the loops we have modeled in this paper are of a smaller size than the one marked by the red box.

Solar Dynamics Observatory (SDO) [Pesnell et al. (2012)].

Even in the present times, there are many open questions regarding the structures we see in Figure 3. When these powerful arches are created, there are situations wherein a phenomenon called magnetic reconnection occurs (Petschek, 1964; Sweet, 1958; Parker, 1957). Our understanding of this active research area has evolved substantially in recent years (e.g. Bhattacharjee, 2004; Zweibel and Yamada, 2009; Cassak and Drake, 2013). It is precisely this moment in the data simulation that one needs to identify, both spatially and in time, that is, both “where?” and “when?” the magnetic reconnection happens.

We loaded the entire data set into SAS Visual Analytics [SAS (2014)], in our own set up on Microsoft Azures cloud environment, and started looking for the phenomena of interest in the data. The aim was to automatically identify where and when the magnetic reconnection occurs, for all possible candidates. We wanted to replace the circular process described in Figure 1 with the linear process described in Figure 2. This could simplify and speed up how you actually get results and find insights, in this particular case regarding how the sun works, maybe in your case, how your customers work.

What we see in Figure 6 is how standardized methods which are widely used in the business world, suddenly find use for a completely different application. The methods for identifying Points of Interest, performing analysis, visualizations and creation of reports are the same, regardless of business or scientific data.

Something the academic world is generally very good at, is to experiment with their data, dare to play with it, explore it with the mindset ”I don’t really know what I will find, but I hope it’s something interesting!”, or to quote a famous physicist

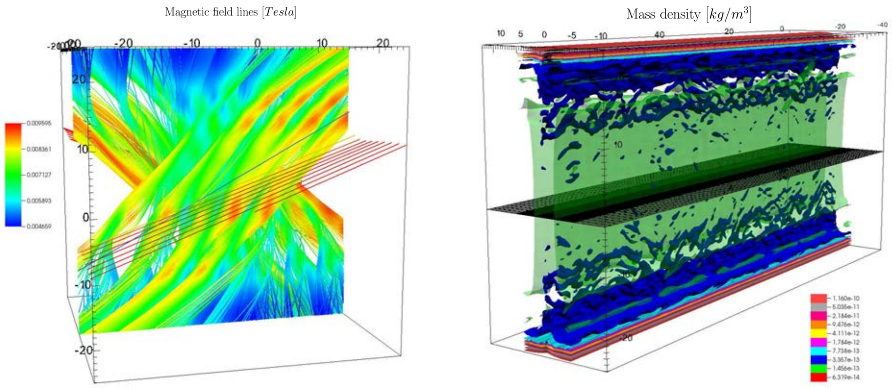


Figure 4: The figures shows two results from two different simulations where we have stretched out the magnetic loop structure as given in Figure 5. To the left we have a resistive magneto hydrodynamic (rMHD) simulation run, showing the resulted recombined magnetic field (stream lines) colored by the fields magnitude $[Tesla]$. Right panel show the mass density profile for a simulation which also contain radiation loss, electron heat conduction, and bulk heating. We clearly see the large densities $[Kg/m^3]$ of the chromosphere taken in the lower and upper section. The grid plane near the middle shows where we extracted for data used in Figure 7. The different resolutions of the grid can also be seen. The wave like structure in the density comes as a quniqence of the heating associated with magnetic reconnection.

"*Experiment is the only means of knowledge at our disposal. Everything else is poetry, imagination.*"- Max Planck

In section 2 we will describe the simulations done for generating the data used in this study. Section 3 will outline how data from the simulations was organized and stored, while section 4 describes the process of identifying the position in time and space of the magnetic reconnection from the physical variables in the simulation.We conclude with a discussion in section 5.

2. simutalion

To simulate the Solar atmosphere we are using the Space Weather Modeling Framework (SWMF) with the Block Adaptive Tree Solar-wind Roe Upwind Scheme (BATS-R-US) fluid solver [Tóth et al. (2005, 2012); Powell et al. (1999)] from the University of Michigan. We will assume that all the gass on the Sun will be fully ionized so we can use magneto hydrodynamic description of the fluid, plasma, picture to describe its behavior. We are solving the resistive magneto hydrodynamic (rMHD) combined with *Spitzer* heat conduction $Q_{Spitzer}$, optical thin radiation loss function $Q_{radiation}$ [Klimchuk et al. (2008)] and an artificial volumetric heating rate $Q_{heating}$ based on the steady state model of coronal loops [Rosner et al. (1978); Klimchuk et al. (2008)]:

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) &= 0 \\
\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot \left(\rho \mathbf{u}\mathbf{u} + \left(p + \frac{|\mathbf{B}^2|}{2\mu_0} \right) I - \frac{\mathbf{B}\mathbf{B}}{\mu_0} \right) &= -\mathbf{g}\rho \\
\frac{\partial \mathbf{B}}{\partial t} - \nabla \times (\mathbf{u} \times \mathbf{B}) &= \frac{\eta}{\mu_0} \nabla^2 \mathbf{B}
\end{aligned}$$

$$\begin{aligned}\frac{\partial e}{\partial t} + \nabla \cdot \left(\left(e + p + \frac{|\mathbf{B}^2|}{2\mu_0} \right) \mathbf{u} - \frac{(\mathbf{u} \cdot \mathbf{B})\, \mathbf{B}}{\mu_0} \right) & \\ + Q_{radiation} - Q_{Spitzer} - Q_{heating} &= -\mathbf{g} \cdot \mathbf{u}\rho + \frac{\eta}{\mu_0^2} |\nabla \times \mathbf{B}|^2\end{aligned}$$

$$\begin{aligned}Q_{radiation} &= N_e N_i \Lambda(T_e) \\ Q_{Spitzer} &= \nabla \cdot \left(\kappa_e T_e^{5/2} \mathbf{b}\mathbf{b} \cdot \nabla T_e \right)\end{aligned}$$

Where we have the mass density ρ, bulk velocity vector $\mathbf{u}$, pressure p, magnetic field vector $\mathbf{B}$, gravity g, resistivity η, total energy e, number density for electron and ions $N_{e,i}$, electron temperature T_e and the optical thin radiative loss function Λ.

To simplify the calculation we will take the Corona loop and straighten it into a rectangular box, see Figure 5. The two ends, which are also the footprints on the Suns surface, are now on the top and bottom of the box. As the hydrostatic scale length are large compared to the perpendicular dimensions of the loop We have only included gravity along the loops length.

For the initial solution of the loops we define the temperature of the Corona and Chromosphere and the length of the loop together with a theoretical equilibrium model [Rosner et al. (1978); Klimchuk et al. (2008)] for the loop, we can set up the initial solution. This solution will not necessarily be a perfect stable solution for our numerical solver so we will run it to the plasma in the box has stabilized. Typical values are $T_{Corona} = 10^6$K, $T_{chromosphere} = 10^4$K and $L_{loop} = 5.0 \cdot 10^7$ m.

When solving numerical the physics of a Coronal loop, as shown in Figure 5, we need to set the proper boundary condition. At the foot points of the loop, in the vertical direction, we enforce no flows out of the domain and hydrostatic equilibrium solution for mass density and pressure. For the other variables we assume no change at the boundary. In the other direction we use periodic boundaries.

2.1 Shearing boundary

To mimic the photospheric flows on the Suns surface which can generate condition where the magnetic field can recombine Parker (1983), we set up a velocity shearing profile at the loops footpoints pendicular to it length, see figure 5. This will generate a magnetic field components out of the plain which will recombine when it has grown large enough.

3. The data

A single simulation will produce about 25 trillion data points in both time and space containing the state variable describing the plasma as well each grid cells position. Each of them are internally represented by a 64 bit floating point number, but can be stored in any wanted precision to reduce storage needs. The three dimensional grid data is organized in discrete time steps which are stored at regular intervals. The grid is composed of a set of blocks each containing a homogeneous cartesian grid. A cut plane through the box showing the grid in the left panel in Figure 4, where we can see that the resolution is increasing as we get closer to the center of the box where we have vertical sheet like structure. Resolution change between neighboring blocks can only change by a factor of two. The possibility of having different area with different resolution is used to resolve the physics properly where its necessary and use less computational resources where we have smoother solution, less

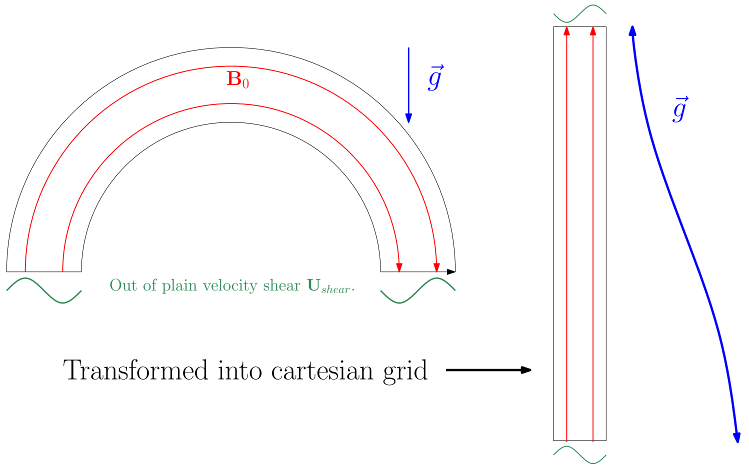

$\mathbf{B}_0$
$\vec{g}$
Out of plain velocity shear $\mathbf{U}_{shear}$.
Transformed into cartesian grid
$\vec{g}$

Figure 5: In Figure 4 we have marked a loop structure on the solar surface. In this work we want so simulate similar but smaller loop structure on the Sun. To do this we will simplify the geometry as shown in this figure. To the left we see a 2D cut of the loop similar to in the image in Figure 3, we have a magnetic field generating the loop, guide field, in red. And solar surface flow pattern out of the plane as illustrated by the green function at the footpoints of the loop. For the simulation we will straighten the loop out to a box where the foot points will end up at the top and bottom. To get the stratification of the solar atmosphere we will change what is a constant gravity field, blue, on the loop, left, to a function of height as shown to the right. We are ignoring gravitational effects across the magnetic field lines, red.

interested in the local physics or close to a boundary to make it less costly to push them fare from the interior of the simulation. In this study our data is describe as table of cells given their position in space, time and state variable.

We use a couple of data sets from different simulations to represent the span of parameters dependent on the plasma environment. The data sets varied in size from about 1 GB to 15 GB size and with different spatial and temporal resolution to test different algorithms for detection of the magnetic reconnection (previously mentioned ”Point of Interest”). Figure 6 shows an example from a low resolution run, taking the whole domain, while the data for figure 8 was taken around zenit of the loop configuration, as illustrated by the horizontal grid shown in Figure 4.

4. The analysis

We loaded one data set at a time with sizes up to about 15GB into SAS Visual Analytics to look for different statistical identifiers in order to identify for magnetic reconnection which is the previously mentioned proverbial “needle in the haystack”. The search was split into two groups, spatial and time evolution.

What becomes clearly visible even for the lowest resolution runs of rMHD shown in Figure 6 is the large expansion in parameter space occurring at the moment the magnetic reconnection starts. The upper left Figure 6 shows the heat map of the span of pressure values in time. The sudden increase with the sudden release of magnetic energy when the magnetic reconnection starts giving rise to pressure waves propagating in the system. In the upper right panel in Figure 6 we see the corresponding values for the generated magnetic field. We see the same bursty nature of the initial state of the reconnection but opposite to

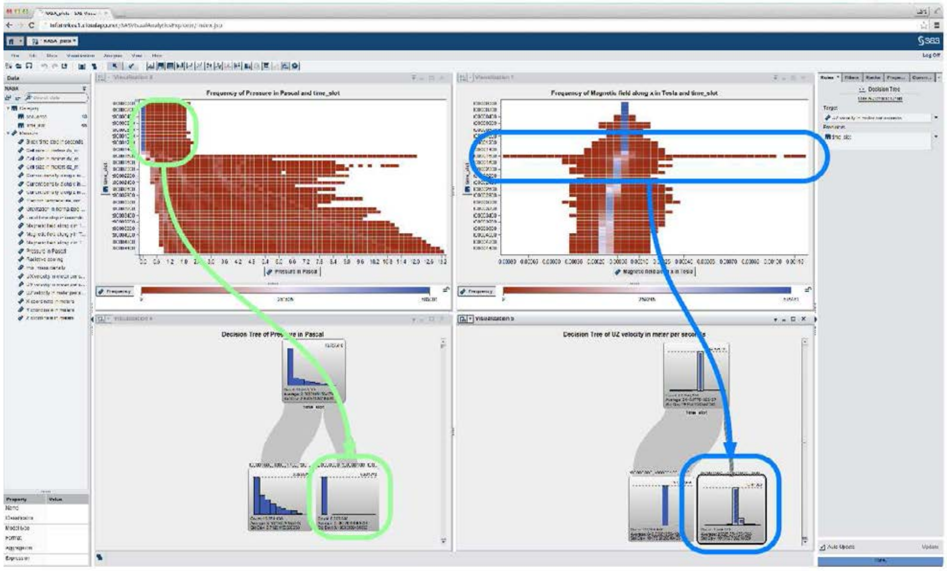

Frequency of Pressure in Pascal and time_slot
Frequency of Magnetic field along x in Tesla and time_slot
Decision Tree of Pressure in Pascal
Decision Tree of UZ velocity in meter per second
Decision Tree

Figure 6: Shows simulated data for one of the many arches that are formed at the surface of the sun and how we used SAS Visual Analytics [SAS (2014)] to identify the crucial moment, the proverbial needle in the haystack, with the help of heat-maps and decision trees.

the pressure we an slow build up and after initial start of reconnection we see a relaxation of the solution. This observation can be an artifact of the low resolution of the run. In the lower part of the Figure 6 we see the same clear difference in the histogram for the decision tree where the histogram for pressure, lower right panel in Figure 6, are separated in before and after the initial reconnection event while for the generated magnetic field , lower left panel in Figure 6, the decision tree only takes out the time step where the magnetic reconnection starts.

The same physical model of rMHD was used for Figure 8 but with a higher resolution. Figure 8 shows box plots of five time steps while magnetic reconnection is proceeding spaced by 100 seconds around the central sheet shown where the magnetic fields are intertwining in the left in figure 4. In SAS Visual Analytics the box of the box-plot [SAS (2014)] is between first and third quartiles, the black line is the median value and the line spans the whole range of values. In the lower left of Figure 8 the “S” shaped curve is the driving magnetic fields which is increased by $\sim 30\%$ during the time span of the figure. We also see the guided, vertical, component field at around 0.005 Tesla and the generated magnetic field across the vertical sheet structure seen in the center of both pictures in figure 4. All components shows clear signs of wave activity but not as clearly as the three components of the velocity vector shown the upper panels and lower right panels in Figure 8 or in Figure 7. The main conclusion to take away form this is that there are no clear trends in the fluctuations of the velocities and the size of the plotted boxes are larger than the max range of values which are limited by physics of the local plasma.

When we include more of the physical processes than what was used for Figure 8 , for the Solar Corona as electron heat conduction, radiation loss in a optical thin medium and plasma bulk heating to match the radiative loss we get a similar picture as illustrated in figure 7. Here we see three time steps, separated by 100 seconds showing the vertical velocity components associated with the magnetic reconnection. We see a slow build up phase in the first two time steps. Where the box-plot box is not much smaller than the full

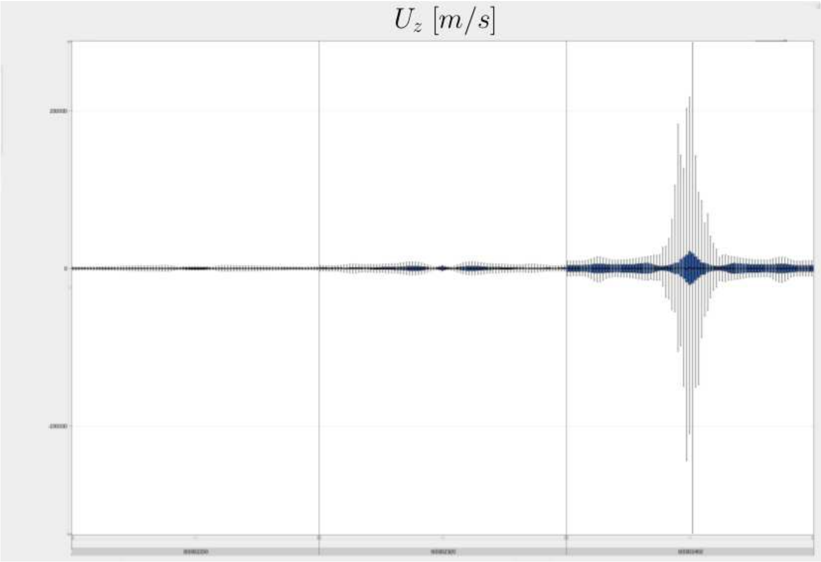
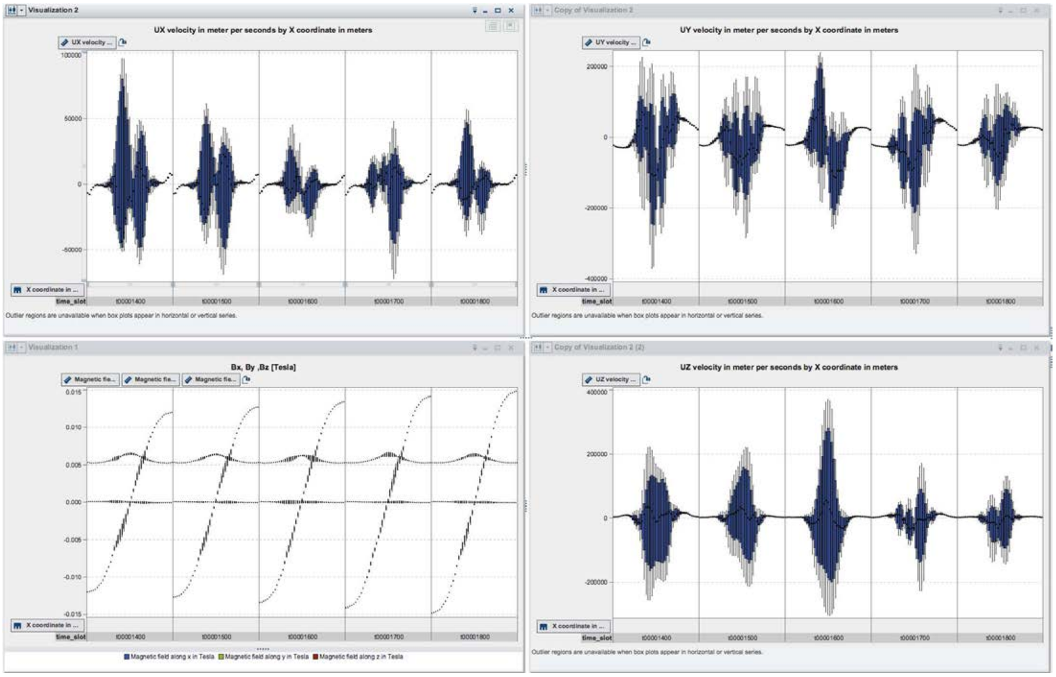


Figure 7: The figure shows a box plot of the horizontal velocity component along the sheet structure sampled on the horizontal grid shown in right picture in figure 4 for tree time steps from left to right 100 seconds apart. For the to first snapshots we only see small variations in the system up to the point where reconnection is starting, right most figure. The clearest indicator is the large range for outliers. The dataset was taken from a simulation using resistive magneto hydrodynamic with additional terms for electron heat conduction, radiation loss in thin medium and had-hoc buck plasma heating.

range of velocity values. This change is dramatic when the magnetic reconnection sets in. In both figure 7 and 8 we see that the variations outside the reconnection area are much lower and the box plots proposing is similar to what you expect for a normal distribution. As we get closer to the center the two figures start to diverge, as the rMHD runs of figure 8 have a mostly a steady increase in the variation while the run with more physics show an decrease before the we get into the reconnection region. In figure 7 we have also a max variable range much larger than the area covered by first to third quartiles.

5. Conclusion

Many of the observed differences between the simulation runs can be because we only dump data for post processing at preset intervals. As we change the physics we will not be able to get the data at the same time in regards to the onset of magnetic reconnection. Therefore making it harder to know what is the influential physical aspect of the onset mechanism for magnetic reconnection. One way of solving this is to store data to disk much more often, but this will waste storage space as most of the data is not of interest. Instead we can find the variability of the velocity variable to check for there changes and then only store it if it passes a threshold value. When we have reached steady reconnection levels we can go back to regular storage intervals. We have seen that using what is to the business and statistical world, familiar methods, faster insights can be achieved by shortening the process from what is described in Figure 1 to Figure 2. Explorative analysis can be a powerful tool, regardless if your trying to understand the sun, studying effects of new drugs, or trying to understand your customers better.



Outlier regions are unavailable when box plots appear in horizontal or vertical series.

Outlier regions are unavailable when box plots appear in horizontal or vertical series.

Outlier regions are unavailable when box plots appear in horizontal or vertical series.

Figure 8: The data is taken form a resistive magnetohydrodynamic simulation, same as shown in figure 4 to the left but narrowed in on the central sheet where we have magnetic reconnection. The figure shows the variations in velocity components, to upper and lower right, and lower left magnetic as they evolve in time with each snapshot separated by 100 seconds. The lower right is equivalent to Figure 7 which is taken from the same simulation as shown in the right pane of Figure 4. We can clearly see the enhanced variability of the velocity in the reconnection region, with a sharp decrease as soon as we are outside the reconnecting sheet structure. In the lower left picture the magnetic fields we see, the "S" shaped driving magnetic field generated by the flows on the boundary is increasing during the time period shown. The vertical magnetic field has been compressed with a resulting gaussian shape and the generated magnetic field by the magnetic reconnection is near horizontal line. We can see the some increase in variation in all the magnetic field components but not as clearly as in the velocity components.

6. Acknowledgments

This work has been made possible by a very good collaboration between Lars K. S. Daldorff (contracted researcher at NASA) and Infotrek (a data warehouse consultancy company in Sweden) with contributions by Saiam Mufti, Lars Tynelius and support from SAS institute Sweden. Also thanks to Laura Fernandes and Sara Baghchehsaraee for helping with this text. The authors also want to thanks James A. Klimchuk for his support. B. Isham was supported in part by ARO grants W911NF-11-0217 and W911NF-10-1-0002. L. k. S. Daldorff was supported from grants from NASA. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Advanced Supercomputing (NAS) Division at Ames Research Center and the NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center and Infotrek.

References

Bhattacharjee, A. (2004), “IMPULSIVE MAGNETIC RECONNECTION IN THE EARTH’S MAGNETOTAIL AND THE SOLAR CORONA,” *Annual Review of Astronomy and Astrophysics*, 42, 365–384.

Cassak, P. A. and Drake, J. F. (2013), “On phase diagrams of magnetic reconnection,” *Physics of Plasmas*, 20.

Kimball, R. and Ross, M. (2013), *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, Wiely.

Klimchuk, J. A., Patsourakos, S., and Cargill, P. J. (2008), “Highly Efficient Modeling of Dynamic Coronal Loops,” *The Astrophysical Journal*, 682, 1351–1362.

Lemen, J. R., Title, A. M., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., Duncan, D. W., Edwards, C. G., Friedlaender, F. M., Heyman, G. F., Hurlburt, N. E., Katz, N. L., Kushner, G. D., Levay, M., Lindgren, R. W., Mathur, D. P., McFeaters, E. L., Mitchell, S., Rehse, R. A., Schrijver, C. J., Springer, L. A., Stern, R. A., Tarbell, T. D., Wuelser, J.-P., Wolfson, C. J., Yanari, C., Bookbinder, J. A., Cheimets, P. N., Caldwell, D., Deluca, E. E., Gates, R., Golub, L., Park, S., Podgorski, W. A., Bush, R. I., Scherrer, P. H., Gummin, M. A., Smith, P., Auker, G., Jerram, P., Pool, P., Soufli, R., Windt, D. L., Beardsley, S., Clapp, M., Lang, J., and Waltham, N. (2012), “The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO),” *Solar Phys.*, 275, 17–40.

Parker, E. N. (1983), “Magnetic Neutral Sheets in Evolving Fields - Part Two - Formation of the Solar Corona,” *ApJ*, 264, 642.

Parker, N. E. (1957), “Sweet’s mechanism for merging magnetic fields in conducting fluids,” *JGR*, 62, 509.

Pesnell, W. D., Thompson, B. J., and Chamberlin, P. C. (2012), “The Solar Dynamics Observatory (SDO),” *Solar Phys.*, 275, 3–15.

Petschek, H. E. (1964), “Magnetic Field Annihilation,” *NASA Special Publication*, 50, 425.

Powell, K. G., Roe, P. L., Linde, T. J., Gombosi, T. I., and Zeeuw, D. L. D. (1999), “A Solution-Adaptive Upwind Scheme for Ideal Magnetohydrodynamics,” *Journal of Computational Physics*, 154, 284 – 309.

Rosner, R., Tucker, W. H., and Vaiana, G. S. (1978), “Dynamics of the quiescent solar corona,” *Astrophys. J.*, 220, 643–645.

SAS (2014), *SAS Visual Analytics 7.1: User’s Guide*, Cary, NC: SAS Institute Inc.

Sweet, P. A. (1958), “The Neutral Point Theory of Solar Flares,” in *Electromagnetic Phenomena in Cosmical Physics*, ed. Lehnert, B., vol. 6 of *IAU Symposium*, p. 123.

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., Hansen, K. C., Kane, K. J., Manchester, W. B., Oehmke, R. C., Powell, K. G., Ridley, A. J., Roussev, I. I., Stout, Q. F., Volberg, O., Wolf, R. A., Sazykin, S., Chan, A., Yu, B., and Kta, J. (2005), “Space Weather Modeling Framework: A new tool for the space science community,” *Journal of Geophysical Research: Space Physics*, 110, n/a–n/a, a12226.

Tóth, G., van der Holst, B., Sokolov, I. V., Zeeuw, D. L. D., Gombosi, T. I., Fang, F., Manchester, W. B., Meng, X., Najib, D., Powell, K. G., Stout, Q. F., Glocer, A., Ma, Y.-J., and Opher, M. (2012), "Adaptive numerical algorithms in space weather modeling," *Journal of Computational Physics*, 231, 870 – 903, special Issue: Computational Plasma PhysicsSpecial Issue: Computational Plasma Physics.

Zweibel, E. G. and Yamada, M. (2009), "Magnetic Reconnection in Astrophysical and Laboratory Plasmas," *Annual Review of Astronomy and Astrophysics*, 47, 291–332.