British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

A comparison of data holdings at World Data Centres for Geomagnetism in Edinburgh and Kyoto

Ewan Dawson (ewan@bgs.ac.uk), Susan Macmillan, Thomas Humphries and Ciarán Beggan
British Geological Survey, West Mains Road, Edinburgh EH9 3LA, United Kingdom

1. Introduction

There are a number of World Data Centres (WDC) for Geomagnetism, each holding a collection of datasets gathered from a variety of geomagnetic observatories around the world. These WDCs are run by host institutes in Boulder, Edinburgh, Kyoto, Moscow and Mumbai, each having different data holdings, and offering different methods for data distribution.

In recent years, efforts have begun to harmonize geomagnetic data holdings across three of these WDCs, those in Boulder, Kyoto and Edinburgh. The data holdings of Boulder and Edinburgh are largely identical; however it is known that there are significant differences between the data holdings at Edinburgh and Kyoto.

Two years ago we compared the data holdings of Edinburgh and Kyoto in order to identify data held in only one or the other of these WDCs. Since then, data sharing between the WDCs has filled in many of these gaps in our respective datasets.

However, as well as differences in the temporal coverage of the data holdings, it is known that the data values held at Kyoto and Edinburgh are not always in agreement. Here we focus on the hourly-mean dataset, and present an analysis on the number of datasets in disagreement, and quantify the magnitude of the disagreement.

2. Comparing the data holdings

We compared the datasets held in common at both Edinburgh and Kyoto as follows:

- The entire hourly mean data holdings from 1883 to 2009 were downloaded from WDC Kyoto in IAGA-2002 format and converted into a series of CSV files containing the XYZ-component data.
- Similarly, the WDC Edinburgh data spanning the same years were downloaded, this time in WDC format, and converted into CSV files containing the XYZ-component data.
- Both datasets were loaded into the ‘R’ statistics application for analysis.
- Where the datasets were found to overlap, the values from the Kyoto data were subtracted from those of the Edinburgh data, to give a single dataset with ΔX, ΔY and ΔZ components. Data points not common to both WDCs were discarded.
- A threshold filter was applied to the differences, zeroing any with magnitude less than 0.5nT. This was done to remove any differences due to rounding; the IAGA-2002 format is capable of representing intensity values to a precision of 0.01 nT, while the WDC format is only precise to 1 nT.
- The data were reduced to a single component by computing ΔF from the ΔX, ΔY and ΔZ components; this gives a convenient metric for the difference between the datasets at each hour (note that this is not the same as the difference in the scalar field values at each hour).
- The resulting data were then reduced in time resolution from one sample per hour to one per year. The annual summary statistics computed were: 1) the number of non-zero ΔF hourly values for each year (*disagreement count*), and 2) the mean of these non-zero ΔF hourly values (*mean disagreement*).

The resulting dataset consisted of one time-series per observatory, each spanning the period of overlap between the data holdings at the two WDCs, and summarizing

3. Analysis of results

There are 286 observatories with hourly mean data held at both WDCs in Edinburgh and Kyoto. The overlap between the two data holdings spans 7,409 observatory-years of data.

Of this set of common data, we found a large proportion is in agreement:

- No disagreements were found in the hourly mean values for 91 of the 286 observatories.
- In terms of observatory-year datasets, 5,910 of the 7,409 datasets (almost 80%) are in agreement.

For most observatories, the number of annual hourly mean datasets which contain disagreements is a small proportion of the total number of annual datasets for that observatory. However, there are a small number of observatories for which the level of disagreement is very high; these are shown in Table 1 (right).

Note that because Dumont d’Urville is located close to the southern dip pole, the observed large disagreement may actually be a spurious result caused by the transformation from the original DHZ-component values into the XYZ-coordinate system. Also, the large number of disagreements at the observatories PSM, VLJ, CLF, DRV and CZT may be due to recent updates made to the data by the Institut de Physique du Globe de Paris which are not reflected in the Kyoto data.

Observatory	% Hourly mean values in disagreement	Mean size of disagreement (nT)
Parc St. Maur (PSM)	98%	12
Val Joyeux (VLJ)	96%	10
Arkhangelsk (ARK)	92%	38
South Georgia (SGE)	88%	12
Roburent (ROB)	75%	83
Chambon-la-Foret (CLF)	74%	9
Gornotayezhnaya (VLA)	70%	653
Hatizyo (HTY)	54%	23
Dumont d'Urville (DRV)	50%	1360
Port Alfred (CZT)	49%	1

Table 1: The ten observatories with the highest proportion of disagreeing values.

Figure 1 (below) illustrates how the total number of disagreements in an annual dataset relates to the mean disagreement.

The cluster of points in the bottom-right shows that there is a large number of datasets where there is almost no agreement between the data held at the two WDCs, but the mean disagreement is small (~1 nT). Of greater concern is the cluster of points along the right edge; these show a significant number of datasets where the data differ wildly between the two data centres. It is datasets like these which contribute to the large disagreements seen in the data for Gomotayezhnaya and Dumont d’Urville observatories. In cases like these the two WDCs will have to work closely with the institutes responsible for the data to try to resolve these discrepancies.

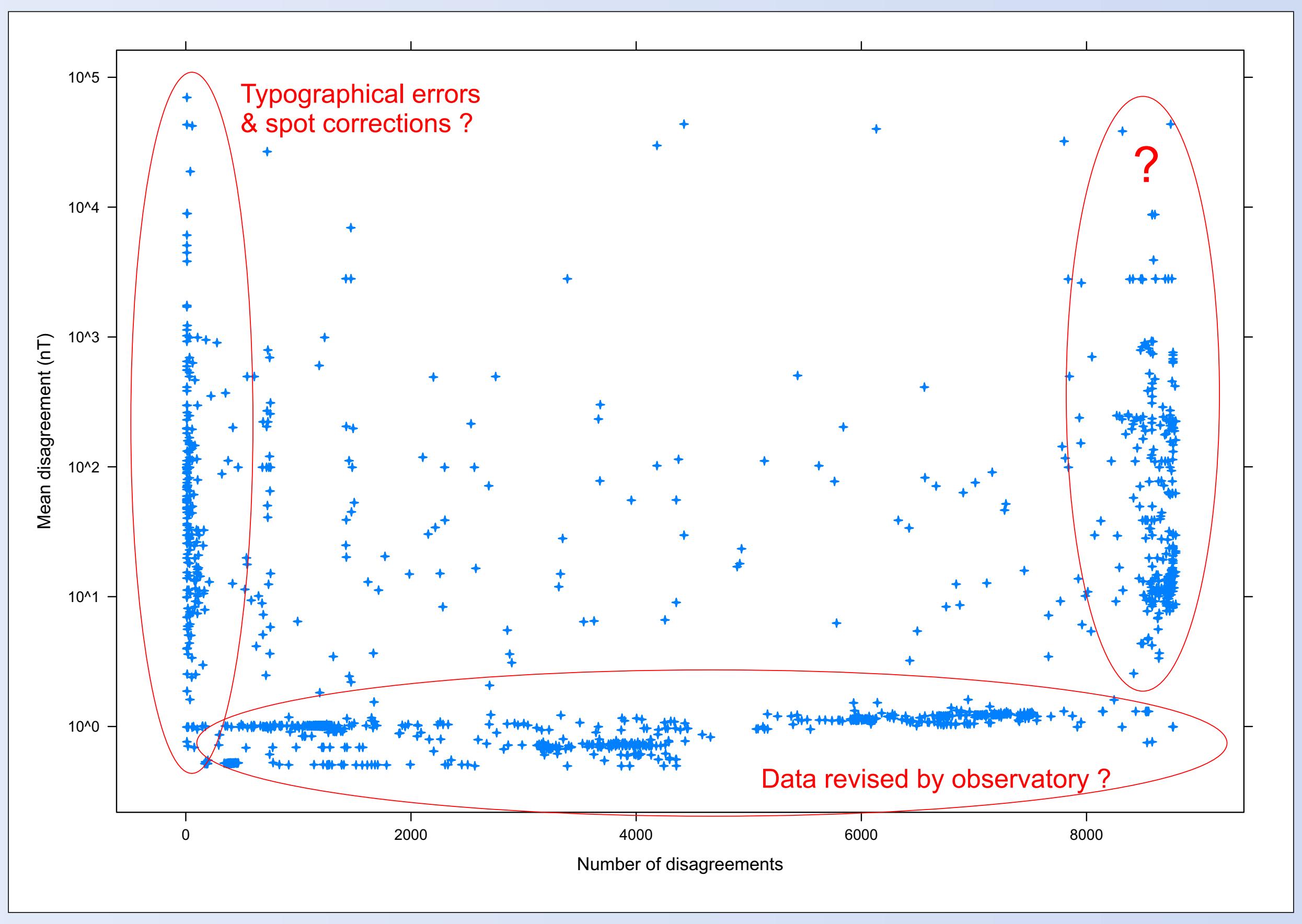

Figure 1: The relationship between the number of disagreeing samples and the mean size of the disagreements, for each observatory-year of data. Both axes are scaled logarithmically.

The points in the left half of the figure should be much easier to deal with, as they represent datasets with only a small number of discrepancies between the WDCs. In particular, the points along the left edge of the plot most likely represent isolated typographical errors, and it should therefore be straightforward to determine which is the correct value.

4. A closer look - Alibag, August 2003

The hourly mean dataset for Alibag 2003 has a high number of discrepancies (7,951), with a fairly low mean disagreement (6.14nT). A comparison of the month of August are shown in Figure 3 below, with the differences in each component on the left, and the actual component values on the right.

Through consultation with the WDC for Geomagnetism in Mumbai, we discovered that the file held at WDC Kyoto contains data from an Izmiran analogue magnetometer, while the WDC Edinburgh data was produced by a DMI digital fluxgate magnetometer. The two instruments were running side-by-side at Alibag observatory throughout 2003. Analysis has shown that the data produced by the analogue system were of higher quality that year. Therefore in this case the Kyoto data should be considered definitive, and replace the data held at Edinburgh.

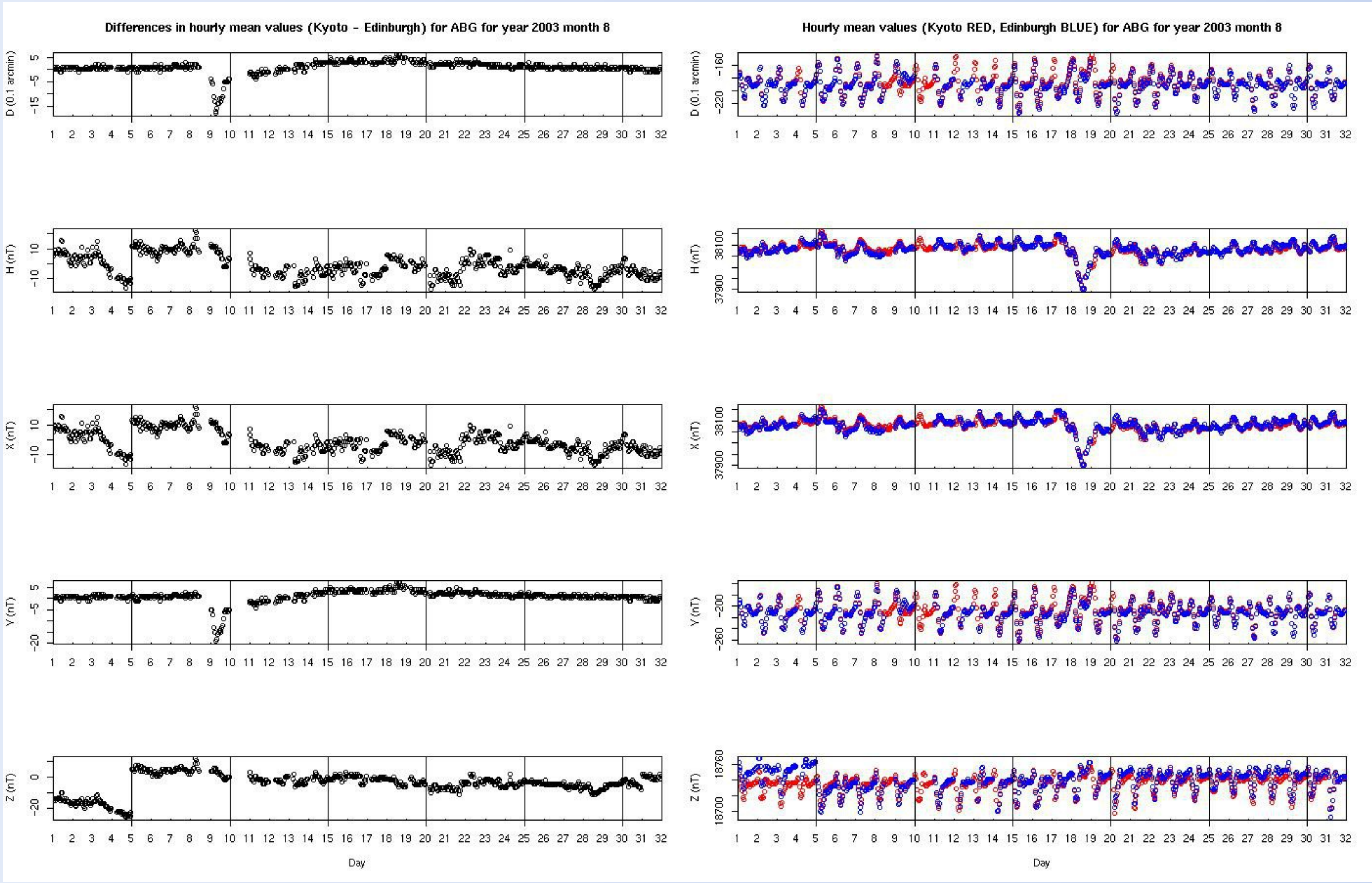

Figure 2: Detailed comparison of hourly mean values at Alibag, August 2003, held at WDCs for Geomagnetism in Edinburgh and Kyoto.

4. Next steps

Using this detailed information on the disagreements between the data holdings, we will work together with our colleagues in Kyoto, and the institutes responsible for the data, to harmonize our data holdings and produce a single definitive set of geomagnetic hourly mean values.

Reference

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Association: IAGA
Symposium: A131
Poster: 5710

IUGG Melbourne, Australia
28th June - 7th July 2011