Format Description for Hypocenter Data File (HDF) files as produced by MLOC v9.9.6

June 18, 2014

Eric Bergman
Global Seismological Services
bergman@seismo.com

The multiple event relocation code **mloc** (which implements the Hypocentroidal Decomposition method of Jordan and Sverdrup, 1981) creates one or more output files that summarize the relocated hypocenters of all events in a cluster in a "one line per event" format. The reason for multiple cases of the file is explained below. These files are known generically as **hdf** files and they all contain "hdf" as part of the file name suffix. The format of these files was for a long time based closely on the format of the same name used by Bob Engdahl for output of his single-event relocation studies (e.g., the EHB catalog). There were always some differences in the **mloc** version of **hdf** files, however, and a number of fields did not translate well from a single event relocation process to a cluster-based process. In recent years the format has been modified more extensively to better serve the research needs that are commonly addressed by **mloc**.

The **hdf** format as defined in **mloc** version 9.9.6 (release date November 29, 2013) is documented here for the benefit of persons who may be extracting information from **hdf** files for further analysis. This document does not apply to "hdf" files from any other source, in particular those produced by earlier versions of **mloc** or those created by Bob Engdahl's single event relocation code.

Different Flavors of HDF files in mloc

The need for several flavors of **hdf** file in **mloc** arises from the different scenarios that are possible:

- No calibration. Only a file with suffix ".hdf" is created. Uncertainties are for relative location only (cluster vectors).
- Direct calibration by locating the hypocentroid with local readings. Only a file with suffix ".hdf_dcal" is created. Uncertainties are for absolute location (cluster vector plus hypocentroid).
- Indirect calibration by shifting the cluster to match *a priori* calibration data. Two files are created, one with the suffix ".hdf" (with uncertainties for relative location) and one with the suffix ".hdf_cal" (with uncertainties for absolute location).
- Direct calibration, followed by indirect calibration in the same run. Two files are created, one with the suffix ".hdf_dcal" and one with the suffix ".hdf_cal". Indirect calibration takes precedence for other output files.

All **hdf** files, regardless of flavor, may be read with the same code. The interpretation of some fields varies slightly, depending on flavor, as documented below.

Version of an HDF-formatted file

The version and release date of **mloc** that was used in any given relocation is carried in the first line of the **summary** file (filename suffix ".summary"). HDF files do not carry any version information, but they can be associated with the corresponding version of **mloc**.

HDF v9.9.6 Format

Columns	Variable (format)
1:4	Origin year (i4)
6:7	Origin month (i2)
9:10	Origin day (i2)
12:13	Origin hour (i2)
15:16	Origin minute (i2)
18:22	Origin seconds (f6.2)
24:32	Geographic latitude (f9.5)
34:43	Geographic longitude (f10.5)
45:50	Focal depth, km (f6.2)
52:52	How starting depth was set (a1)
53:53	Free depth flag (a1)
54:59	Depth from input file (f6.2)
61:63	Magnitude (f3.1)
64:65	Magnitude scale (a2)
67:76	Event ID (a10)
78:81	Number of observations contributed to hypocentroid estimation (i4)
83:86	Number of observations used for cluster vector (i4)
88:91	Number of observations flagged as outliers, fcode = 'x' (i4)
93:98	Normalized sample variance for the cluster vector (f6.2)
100:104	Uncertainty in origin time, sec (f5.2)
106:109	+ uncertainty in depth (deeper), in km (f4.1)
111:114	- uncertainty in depth (shallower), in km (f4.1)
116:120	Epicentral distance of nearest station for cluster vector (f5.1)
122:126	Epicentral distance of farthest station for cluster vector (f5.1)
128:132	Largest open azimuth for cluster vector (f5.1)
134:136	Semi-axis azimuth (i3)
138:142	Semi-axis length, km (f5.2)
144:146	Semi-axis azimuth (i3)
148:152	Semi-axis length, km (f5.2)
154:159	Area of confidence ellipse, km^2 (f6.1)
161:164	Calibration code (a4)
166:185	Annotation (a20)

Notes

Source of Focal Depth Constraint

Most **mloc** analyses are conducted with focal depth fixed (not necessarily at the same value) for all events. Therefore it is necessary to keep track of the type of information that has been used to set the focal depth. Even in the case of a free-depth solution, it is necessary to specify a starting depth and a record is needed of the type of information used for that purpose. This is done with a single character flag, for which a set of standard definitions are used in **mloc**:

- c = cluster default depth
- d = depth phases
- e = engineered (e.g., man-made explosions)
- f = fault model (e.g., InSAR, GPS)
- i = input data file
- l = local-distance readings (more than 2-3 focal depths)
- m = mloc free depth solution
- n = near-source station readings
- r = relocation (outside mloc) with free depth
- u = unknown (just a guess)
- w = waveform analysis

The difference between the flags “l” and “n” is subtle. Readings at very close distances (flag “n”), less than about two focal depths, provide a strong constraint on focal depth because of raypath geometry, i.e., changes in depth change the raypath length and therefore the travel time. In the context of calibrated multiple event location, however, it is possible for direct crustal phases observed at greater distance (flag “l”) to also provide constraint on focal depth, but through the origin time, not the raypath geometry. This is possible because in most cases there are many more regional and teleseismic arrivals than local distance ones. The arrival times of these phases become referenced (through the assumed velocity model) to the average depth of events in the cluster. If a new event is introduced to the cluster with a significantly biased focal depth, it will not be possible to simultaneously fit the local distance (Pg and Sg) readings and the regional Pn and teleseismic readings to the pattern of arrival times established by the other events. In that case the origin time will be set to best match the more numerous Pn and teleseismic data, leaving the local-distance readings with distinct residuals. By adjusting depth until these residuals are small, one can make the depth of the new event consistent with the rest of the cluster. The key, obviously is that some subset of the cluster events must have more direct constraint on depth, such as near-source readings, and these must be propagated through the rest of the cluster reasonably accurately.

Free Depth Flag

The column immediately after the focal depth constraint flag is used for a special flag “f” which indicates that the focal depth has been a free parameter in the relocation.

Depth from Input File

This variable holds whatever focal depth was specified for an event in the input data file. Depending on what other commands are issued in the command file or interactively this value may or may not be used as the starting depth for relocation. It is sometimes useful to have this for comparison with the depth actually set for (or determined in) relocation. Large discrepancies, more than, say, 10 or 15 km, may warrant investigation.

Magnitude and Magnitude Scale

A single representative (or preferred) magnitude is carried through the **mloc** analysis to assist in interpretation. It has no bearing on the relocation. MNF input files can carry multiple estimates of magnitude, but one will be specified (or selected by default) as the preferred magnitude. If no magnitude estimate is available, these two fields will be blank. In some cases a magnitude may be available but not a scale, in which case the magnitude scale field will be blank.

Event ID

This field is provided to carry an EVID that was assigned by some relational database, but as far as mloc is concerned it is simply a character variable of 10 characters. If the field has an entry it will have been read from an MNF input file. If an integer EVID is stored in the field it should be right-justified.

Number of Observations (phase readings)

The **hdf** format carries three different counts of number of readings:

- Number of observations contributed to hypocentroid estimation
- Number of observations used for the cluster vector
- Number of observations flagged as outliers

It is wise to review carefully the reliability of the relative location (cluster vector) for an event that contributes many readings to the hypocentroid in a direct calibration study. Conversely, events with relatively small number of readings contributing to the cluster vector should be reviewed carefully, because the relative location of such events is likely to be poorly determined..

The third variable in this set carries the number of flagged readings that are due to being judged to have a large residual (so-called “cluster residuals” in which the residuals are compared on the basis of mutual consistency rather than absolute value). These are the readings for which fcode = ‘x’. It is not uncommon for the number of outliers to be a significant fraction of the number of readings used for the cluster vector, but when the fraction becomes notably large (say, more than 25%) investigation is warranted.

There are other reasons why a reading may be flagged to prevent its being used in the relocation but these are not tabulated in the **hdf** file. Examples include duplicate readings, unknown phase types or phase types that are not used for relocation, and stations for which no coordinates are available.

Normalized Sample Variance

This field carries a measure of the statistical self-consistency of the error budget related to the cluster vector, the normalized sample variance. If the data followed our statistical model perfectly (i.e., data drawn randomly from a normal distribution with spread equal to our estimated empirical reading error) the expected value would be 1.0. Values both larger and smaller are actually observed, naturally, and the range of variation declines as the analysis proceeds and outlier readings are identified and removed by flagging.

A Bayesian term in the calculation of the normalized cluster sample variance represents our *a priori* state of knowledge about its expected variability and prevents it from going unrealistically small for events with few data points. This expectation on the spread of values for the normalized sample variance (about 0.35) provides a check on the internal consistency of the statistical model for each event. Values larger than about 2.0 (~3σ) may reveal the presence of readings that violate the statistical model (e.g., outliers). By the end of a relocation analysis there should be few such cases.

Uncertainty of Origin Time

In seconds. For uncalibrated clusters and direct calibration, it is taken from the covariance matrix of the relocation, including uncertainty of hypocentroid and cluster vector. For indirect calibration it is based on the uncertainty of the calibration shift plus the uncertainty of relative origin time from the cluster vector.

Uncertainty in Focal Depth

Uncertainty in focal depth can be read from the input file, from the command file, or from the relocation. Uncertainty in focal depth is carried in two fields because it is not uncommon for the uncertainty to be asymmetric. This can arise when inferring depths from teleseismic depth phases and there is uncertainty about the correct identification of the phase (pP, sP, or pwP). It can also arise in waveform analyses where the error vs depth curve is not symmetric. If focal depth has been a free parameter in the relocation it will be symmetric, and it will over-ride any specification in the input file and command file. For uncalibrated clusters and direct calibration with a free depth solution, it is taken from the covariance matrix of the relocation, including uncertainty of hypocentroid and cluster vector. For indirect calibration it is based on the uncertainty of the calibration shift plus the uncertainty of relative depth from the cluster vector. If no estimate of uncertainty in focal depth is available, the fields are blank.

Epicentral Distance Range

The epicentral distance, in degrees, of the nearest and farthest reading used for the cluster vector of an event is carried in these fields. This is useful for judging how an event is connected to the cluster, i.e., through local readings or teleseismic readings, or both.

Open Azimuth

The largest open azimuth for the readings used to estimate the cluster vector is carried in this field. The reliability of the relative location of an event should be questioned if this value is much greater than 180°.

Confidence Ellipse

Confidence ellipses (90% confidence level) for the epicenter are given in four columns which give the azimuth and length of each semi-axis (half-length) for the ellipse. The shorter semi-axis is given first. Azimuth is in integer degrees, clockwise from North. Semi-axis lengths are given in decimal km.

A fifth column carries the area of the 90% confidence ellipse, in km^2. This is a convenient metric to monitor when searching for events with problems. A circle of 5 km radius (the canonical GT5 location) has an area slightly greater than 75 km^2.

As discussed above, the interpretation of the confidence ellipse (relative vs absolute location) depends on whether the cluster has been calibrated or not, which is indicated by the file name suffix.

Calibration Code

I have recently developed a code (GTCNU) to carry information about the calibration status of an event to replace the widely-abused "GTX" formulation introduced by Bondar et al. (2004). This subject is documented elsewhere. **mloc** calculates this code based on the nature of the relocation and calibration data available. The calibration code includes a scale parameter (location accuracy level).

Annotation

It is possible to declare a comment or annotation for an event in the **mloc** command file, using the "anno" command. An annotation can also be read from an MNF input file (in the event record), which will take precedence over an annotation given in the command file. Any annotation will be appended to the end of the line in all **hdf** files. The comment is limited to 20 characters.