TMAP DATA FORMATS
a brief overview

TMAP's data formats are simple, self-checking layouts of data designed to facilitate

a) the handling of very large data sets through i) allowing portions of the data set to be off-line; and ii) allowing arbitrary four dimensional rectangular sub-regions to exist as separate (reduced size) data sets with access identical to the parent data set

b) high speed random access to arbitrary four dimensional rectangular sub-regions of the data that is on-line

c) internal self-checking to ensure data integrity

d) "intelligence" in the sense that all required scaling, units, variable names, coordinate locations, etc. are part of the data set.

To achieve these goals a TMAP data set is broken down into three parts:

A descriptor file - which contains the intelligence of the data set (as described above) and points to a grid file and various data files.

A grid file - which contains definitions of coordinate axes and assembles various combinations of axes into grids of one to four dimensions.

Data files - random access by record, unformatted (binary) and redundantly self-checking for correct positioning of records

In considering performance the orientation of the records with respect to the underlying axes of the grid is vital. For example, if each record contains a time series of n points at a single point in space then accessing an m x p grid in space will require m*p record reads and will process n times more data that is actually required. Similarly, if each record contains a line of n points lying in an east-west orientation then accessing a time series will be inefficient.

There is no general solution to this problem that optimizes efficiency for all cases. The TMAP formats address the problem by offering a variety of options with respect to record ordering and record orientation. The user must pick the option that provides the best overall compromise given the intended mix of data access styles.

For accesses that will be predominantly in spatial (X,Y,Z) grids the user will probably want to use "GT" (grids at time steps) format. This format lays down full spatial grids (one to three dimensions) for each variable, then runs through the collection of

variables - each on a possibly different grid - and finally repeats the entire sequence for each time step. Time steps may be located in multiple files allowing some of the data to be off-line at any given time. (The GT descriptor file parameter "d_ordering" controls the ordering of the X, Y and Z axes.)

For access styles that will be predominantly time series the user will probably want to use "TS" (time series) format. This format lays down complete time series at each point (with a record length selected by the user via the TS descriptor file parameter "d_ndataperrec"). Selecting a shorter record length will make access to spatial grids more efficient at a slight penalty in time series performance. Successive time series will march through the grid points of the file - always in the order X then Y then Z. And successive variables follow in turn - possibly in multiple files to permit some of the data to be off-line at any given time.

"GT" and "TS" descriptor files have distinct, although very similar, formats.

TMAP grid files allow a user to define axes which are a sequence of coordinates, not necessarily regular and not necessarily ordered (although only ordered axes are supported as of 5/90). Time axes may have date to timestep conversion information embedded. Each axis is given a name.

Grids are composed of a collection of one to four axes with some axes associated as "outer products" (producing, for example, Nx*Ny points from the X and Y axes together) and some axes associated as "inner products" (producing, for example, Nx points from the X and Y axes together, where Nx must equal Ny). (Only outer products are supported as of 5/90).