Background Discussion with Steve Worley on use of UIDs – the ICOADS perspective
23 Aug 2011

The Databank needs to consider whether a UID should be incorporated into the Databank data provenance architecture. To begin assessing whether the use of UIDs is necessary, Steve Worley provided background information on the planned use for UIDs in ICOADS.

A UID is a unique identifier that is attached to every observation. All observations made at a single point and place in time are given the same UID. While a ship or buoy will have an identifier – the UID is different in that it is applied to every observation uniquely.

The UID is useful in that any user working with the dataset and finding incorrect values can focus on the specific UID in informing the ICOADS developers – the user doesn't need to know the release number, the deck number, etc.

The goals in using UIDs are the following:

- Track records locally
- Track records through reanalyses, easily make links back, gain metadata from feedback
- Easy to draw a focus on answer questions from user
- Easy to trace recommended fixes from users
- Allow for bias corrections to be received and added, from the supporting (international) community.

Examples of Benefits to use of UID
Use of ocean obs in reanalysis – identification of systematic and random errors. If a UID is carried along with every single observation in ICOADS it is easier to identify the source of the problem.

Before final release of new version – will find for example 30% of identical duplicates. 5-8% are near duplicates but don't know which to choose – the approximate duplicate will retain the UID even if it is excluded from the new release. Then when the next version is developed the observation can be revisited. Additional information at that time may help resolve the duplicate issues and associated uncertainty.

Inclusion of Adjusted data values – As ICOADS Value Added Data (IVAD) is developed, data providers will have the opportunity to provide bias corrections. An attachment will be developed that contains the Adjusted data. The same UID will be used for the Adjusted and the Raw data.

What does a UID look like
Objective was to make the UID simple; don't attempt to make the ID more than a single purpose number. Any more elaborate scheme will probably not endure over time. Experience from

NODC WOD.

UID is a base36 number

- Allows us to go from a 9-10 digit integer to a 5 digit alphanumeric string
- It is total unique, looks a little strange (0-9,A-Z)
- For example, 546342564 is 1JXYT0 when converted to a base36 number
- 546342564333 is M3N30M3
- Base36 data converters and descriptions are available online for better understanding
- Currently have about 300M records, need to allow for 1-9B
- Adding roughly, 2M/month of near real-time preliminary
- It is easily sortable in Unix
- Saves storage space, order GB's
- Easily handled in a DB

Reasons to avoid the use of UIDs in the land surface databank

The unique nature of ocean observations creates a greater need for the use of UIDs than there is for land surface observations.

- Ocean obs – the platform is typically moving. Thus it is more difficult to track and manage the observations.
- Land observations are taken at a fixed point in space, and although stations can move, that is undesirable from the standpoint of data continuity.
- When a land surface station moves, the station will have new metadata on station location, environment, etc, and may (often?) be given a new station identifier – in effect becoming a new station with the previous station closing.

For a single element observed from a land surface station – e.g., daily mean temperature – a UID does not seem to be necessary – because it is easy to communicate with users about particular observations simply by using the station identifier and time (year, month, day) the observation was taken.

If multiple elements are recorded by a land surface station (max temp, min temp, precipitation, solar radiation) all of those elements for the same place and point in time should be given the same UID.

- Land surface observation elements are often managed in different datasets, unlike ICOADS, that retains and manages all observations together in a single dataset for every observation point and time.
- This could potentially be a maintenance nightmare for the land surface databank as it expands into additional elements. Attempting to identify and assign the same UID to

various elements taken by the same station at particular points in time could be resource intensive.

Key Question: Are there benefits to the use of UIDs for the land surface databank that outweigh the additional processing complexities involved in assigning and managing UIDs.