**How to prepare the reference database for OW**

*Annie Wong, October 2008*

1). Make sure the data are clean. Remove duplicates. Only use casts that are deeper than 900 dbar. The 900 dbar criterion is for excluding coastal stations and stations shallower than the T-min layers. Please refer to the Argo QC Manual for a detailed list of reference data selection criteria.

2). Separate data into 10°×10° WMO boxes. BOTTLE data are stored under filename bot_****.mat in /data/climatology/historical_bot. CTD data are stored under filename ctd_****.mat in /data/climatology/historical_ctd. Argo data are stored under filename argo_****.mat in /data/climatology/argo_profiles. **** is the relevant WMO number. For each 10°×10° WMO box, store the following variables:

| | |
|---|---|
| dates | (1 × n, yyyymmddhhmmss) |
| lat | (1 × n, north is +, south is –) |
| long | (1 × n, 0° to 360°, e.g. 60°W = 300°) |
| pres | (m × n, in dbar) |
| sal | (m × n, salinity in PSS-78) |
| temp | (m × n, in-situ temperature in ITS-90) |
| ptmp | (m × n, potential temperature relative to 0 dbar) |
| source | (1 × n, unique character strings stored in a cell array) |

*Note (a)*. “dates” here takes a different format from “DATES” in float_source.

*Note (b)*. Make sure “long” goes from 0° to 360°.

*Note (c)*. With regard to “source”, for CTD and BOTTLE data, you can choose whatever unique identifier you like, as long as you store them as character strings in a cell array. For Argo data, you must use the file names from the single profile Argo netcdf files. For example, source = [ cellstr(‘19019_021’), cellstr(‘19010_022’), cellstr(‘291999_078’) ]. This is because OW uses “source” to exclude Argo data that come from the same float being analysed from the historical data selection. It is optional whether you want to include the prefix of ‘R’ or ‘D’, and the suffix ‘.nc’, in the character strings.

*Note (d)*. The four matrices “pres”, “sal”, “temp” and “ptmp” do not have to line up along a fixed y-axis. OW takes reference data in any vertical resolution and any lengths. There is no need to do any vertical interpolation. So “m”, the number of rows in a matrix, is not fixed. For example, if you have 1-dbar bin CTD data that go to 6000-dbar, then m=6000. If you have BOTTLE casts that have 38 bottle samples, then m=38. If you have Argo data, then m = however many levels the floats report. So “m” will be different in every WMO box. Unfortunately, within each WMO box, the nature of matrices dictates that “m” has to be the length of the longest vector in the WMO box. So use NaNs to fill in the empty spaces.

*Note (e).* The reference database should contain full-depth profiles. This is because even though Argo now only samples to 2000-dbar, future technology will improve to a point where Argo can sample abyssal depths, and we will need full-depth reference data for calibration. Moreover, for people who continue to use WJO, they need the full depths for vertical interpolation, because WJO interpolation starts from the bottom.

3). Edit /data/constants/wmo_boxes.mat. The 1st column is the list of all WMO numbers. The 2nd column denotes existence of CTD data. The 3rd column denotes existence of bottle data. The 4th column denotes existence of Argo data. 0 = no data, or do not use. 1 = data exist, and use them. It is important that you edit this file, because OW uses this file to check which WMO box has what data available. If, for example, you do not want to use bottle data, simply set the 3rd column to all 0s. The order of the WMO boxes in the 1st column is fixed, so please do not change anything in the 1st column, or delete any rows.

4). Please email awong@ocean.washington.edu with further queries.