Cross-over analysis of hydrographic variables: XOVER v1.0

Matthew P. Humphreys (m.p.humphreys@soton.ac.uk)

Ocean and Earth Science, University of Southampton, UK

Friday 10th July 2015

Abstract

The program XOVER v1.0 has been written to enable cross-over analysis of hydrographic data in the MATLAB® (MathWorks®) computing environment. Fundamentally, 'cross-over analysis' is the comparison of measurements of a variable that have been carried out at different times and/or using different techniques, but in the same geographical location. Although it may be usable within other frameworks, XOVER is primarily designed for data sets that are organised like those typically produced by hydrographic research cruises, as follows. Each set of measurements is collected during a single research cruise, and should be quality-controlled such that its results are internally consistent. During each cruise, measurements are carried out at 'stations', each of which samples a range of depths in the water column at the same latitude-longitude point. The user provides XOVER with a 'test' data set of measurements and a 'master' data set to compare it with; XOVER then finds pairs of stations in the test and master data sets that are within a user-specified horizontal distance of each other, interpolates variables of interest in the master data set to match the test data set, and calculates the residuals for each input data point, station and cruise. The accompanying script XOVER_plots provides some tools to begin visualising and interpreting the results.

Both scripts can be downloaded from the MATLAB File Exchange:

- http://uk.mathworks.com/matlabcentral/fileexchange/52063-xover

Contents

1. Compatibility

XOVER and XOVER_plots were written using MATLAB 8.5 (R2015a) and are compatible with MATLAB 8.2 (R2013b) and newer. The main incompatibility with earlier versions of MATLAB arises from the use of the relatively new 'table' data type[1].

2. Methods

2.1. Conversion of latitude and longitude co-ordinates

Input latitude and longitude co-ordinates for each station are converted into a three-dimensional Cartesian co-ordinate system by the subfunction XOVER_ll2s, which is included in the XOVER program. The 3D co-ordinate system has its origin at the centre of the Earth, and assumes that the surface ocean at every station is at a constant radius of 6371 km.

2.2. Matching station pairs

The converted co-ordinates are used to find each master data station falling within the input hzlimit distance of each test data station, to form a list of 'station pairs'. The test and master station in each pair are considered to be in the same geographical location as each other. Each test station can be matched to multiple master stations, and vice versa.

2.3. Master data interpolations

At each master data station that contains at least 2 test variable results at different index variable values, XOVER generates a piecewise cubic Hermite interpolating polynomial (PCHIP) fit (Fritsch and Carlson, 1980; Kahaner et al., 1988) between the index variable and each test variable. Each PCHIP fit is then used to interpolate values of the test variables to the values of the index variable observed at all test data stations paired with the master station (Figure 1).

2.4. Initial calculation of residuals

The residuals are first calculated for each variable at each test data point and master data station by subtracting the interpolated master data from the corresponding measured values in the test data set (Figure 1). Residuals are calculated only for the test data falling within the range of the index variable that was used to generate each interpolation; no extrapolation is permitted. These residuals are stored in the output t_ms.

2.5. Station pair, station, and cruise mean residuals

Finally, the individual data point residuals are combined in various ways. Firstly, the mean and standard deviation (SD) of all residuals is calculated for each station pair, and stored in the outputs ts_ms and sp. Secondly, the mean and SD of residuals between each master station and all test stations is stored in ms. Thirdly, the mean of all individual data point residuals for each master cruise is calculated, and stored in mc. Consequently, each match between a test data point and a master data interpolation is given equal weight in each of these calculated means.

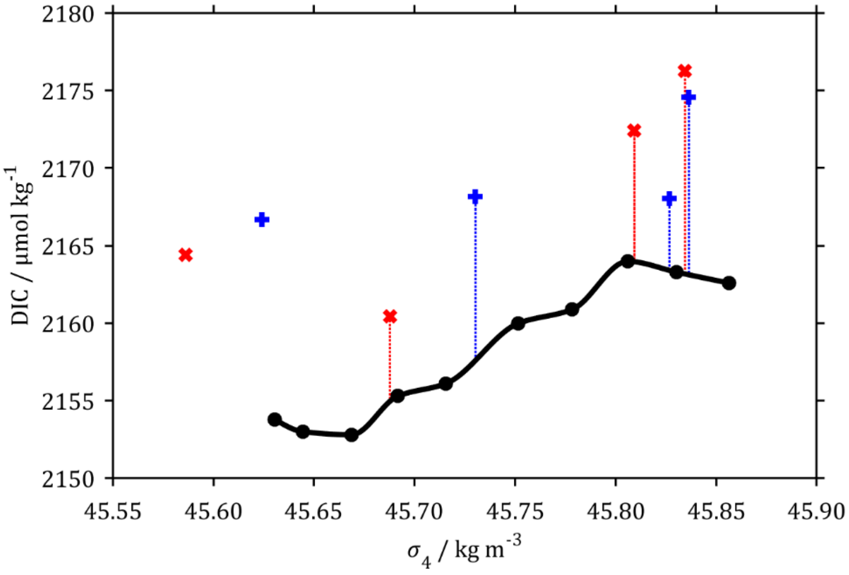

Figure 1. Filled black circles show all data points at an example master data station, and the thick black line joining them is the PCHIP fit. In this example, the index variable is σ_4 (potential density anomaly at 4000 dbar), and the test variable is DIC (dissolved inorganic carbon). The red crosses and blue plusses are the data from two test data stations in the same geographical location as the master data station. The red and blue dotted lines indicate the residuals, which are all positive in this example. Note that two data points have lower σ_4 values than the minimum σ_4 at the master data station, so residuals are not calculated for these measurements. Figures like this can be generated using XOVER_plots; see 5.1.

[1] http://uk.mathworks.com/help/matlab/tables.html

3. XOVER inputs

The 5 inputs required by XOVER are t, m, ivar, tvars, and hzlimit.

3.1. Test data set (t)

The first input t contains the 'test data set' as a table. It must contain at least the following columns, none of which should have any missing values:

t.cruise An identifier that is unique to each test data set cruise; can be a cell[2] of strings[3] or a double[4].

t.station An identifier that is unique to each station within each test data set cruise; can be a cell of strings or a double. Values can be replicated between different cruises.

t.lat The decimal latitude of each data point, between -90 and +90 in °N of the equator, as a double.

t.lon The decimal longitude of each data point, between -180 and +180 in °E of the equator, as a double.

t.ndate The sampling date and time of each data point, in the output format of MATLAB's datenum function, as a double.

Typically, the test data set consists of a new set of measurements that you wish to compare with previously reported results. The data for each 'cruise' should be a set of measurements that are internally consistent. The data for each 'station' should be at roughly the same latitude, longitude, date and time at each other, covering a range of depths. Depth is not required as an input, unless it is selected as the index variable ivar.

In order to compare a subset of the test data set with the master data set – for example to consider test data only from below a certain depth – it is necessary to provide only that subset of the test data set as the input t; XOVER will perform the cross-over analysis using all of the data provided to it. An example of how to do this is included in XOVER_plots.

3.2. Master data set (m)

The second input m contains the 'master data set' as a table. It must contain the same columns as t, and they must be named in the same way as in t. Example data sets that could be used for m include the GLODAP (Key et al., 2004), CARINA (Key et al., 2010) and PACIFICA (Suzuki et al., 2013) quality-controlled syntheses, or the GEOTRACES Intermediate Data Product 2014 (http://www.bodc.ac.uk/geotraces/data/idp2014/); MATLAB scripts to load some of these data sets into a format suitable for input into XOVER are available (e.g. Humphreys, 2015), and more are planned.

Like for the input t (3.1), in order to compare the test data set with a subset of the master data set – for example to consider master data only from below a certain depth – it is necessary to only provide that subset of the master data set as the input m; XOVER will perform the cross-over analysis using all of the data provided to it.

3.3. Index variable (ivar)

The third input ivar is the column name of the index variable in t and m that will be used to interpolate the master data m to match the test data t, as a string. Typical choices may include potential density anomaly, neutral density, or depth, but any quantitative variable can be used. NaN values[5] are permitted but will be excluded from the cross-over analysis.

Example

```
ivar = 's4';
```

3.4. Test variables (tvars)

The fourth input tvars contains the column names of the 'test variables' that are to be compared between t and m, as a cell of strings. If only one variable is desired, it must still be a string in a cell, not just a string.

Examples

```
tvars = {'no3'}; % one test variable
tvars = {'dic' 'ta' 'o2'}; % several test variables
```

3.5. Horizontal distance limit (hzlimit)

The fifth and final input hzlimit is a horizontal distance in kilometres, as a double. Only master data stations that fall within the hzlimit distance of each test data station will be used in the cross-over analysis.

Example

```
hzlimit = 50;
```

[2] http://uk.mathworks.com/help/matlab/cell-arrays.html
[3] http://uk.mathworks.com/help/matlab/characters-and-strings.html
[4] http://uk.mathworks.com/help/matlab/numeric-types.html
[5] http://uk.mathworks.com/help/matlab/ref/nan.html

4. XOVER outputs

The seven outputs produced by XOVER are sp, ts, ms, t_ms, ts_ms, mc and t.

4.1. Station pairs (sp)

The first output sp is a table of all the pairs of stations falling within the hzlimit distance of each other. The columns sp.t_cruise, sp.t_station, sp.t_lat, sp.t_lon and sp.t_ndate, with the prefix t_, provide the cruise, station, and mean latitude, longitude, and date and time of a station in t. The cruise and station fields are cells of strings, while the others are doubles. The columns sp.m_cruise, sp.m_station, sp.m_lat, sp.m_lon and sp.m_ndate, with the prefix m_, provide the cruise, station, and mean latitude, longitude, and date and time of a station in m that is within the hzlimit distance of the station in t that is in the same row of sp. The cruise and station fields are cells of strings, while the others are doubles. The horizontal distance between the two stations on the row in kilometres, as a double, is given by sp.hzdist. The columns sp.*_resid, sp.*_std, and sp.*_nobs are replicated for each variable in the input tvars, with the * replaced by the test variable name in each case. The sp.*_resid field provides the mean residual between all measurements in t at the station sp.t_station and the interpolated values from sp.m_station in the same row of sp, while the sp.*_std and sp.*_nobs fields provide the standard deviation and number of these measurements respectively.

4.2. Test stations (ts)

The second output ts is a table in which each row represents a unique station in the test data set t. Its columns ts.cruise, ts.station, ts.lat, ts.lon and ts.ndate provide the cruise, station, and mean latitude, longitude and date and time of each station in the test data set t. The cruise and station fields are cells of strings, while the others are doubles. The three-dimensional Cartesian co-ordinates for each station (2.1) are given by ts.sxyz0.

4.3. Master stations (ms)

The third output ms is a table in which each row represents a unique station in the master data set m. It contains the same fields as the output ts (4.2), and additionally the columns ms.*, ms.*_interp, ms.*_resid, ms.*_std, and ms.*_nobs, which are replicated for each variable in the input tvars, with the * replaced by the test variable name in each case. The field ms.* contains the data from the master data set that was used to generate the interpolation for that station; each row is a cell containing a matrix (double) which has the index variable values in the first column and the test variable in the second. The ms.*_interp contains the output of the MATLAB function pchip that is used to generate each interpolation, which can be evaluated using the MATLAB function ppval. The ms.*_resid field provides the mean residual between all measurements of each variable in t at all of the test stations matching each master data station, while the ms.*_std and ms.*_nobs fields provide the standard deviation and number of these measurements respectively.

4.4. All test data points vs master stations (t_ms)

The fourth output is a structure containing matrices (doubles) of the individual interpolated data points and their residuals. The rows in each matrix correspond to rows of the output t, while the columns correspond to rows of ms. The structure contains fields t_ms.* and t_ms.*_resid for each test variable *; the former is the original interpolated values of the test variable, while the latter is the residuals (i.e. interpolations subtracted from the test data set values). There are NaN values where either the master station was not within the hzlimit distance of the corresponding test station, or the test station index variable was outside the interpolated range of the same variable at the corresponding master station (i.e. the value would have been an extrapolation).

4.5. All test stations vs master stations (ts_ms)

The fifth output is a structure containing matrices (doubles) of the residuals for each station pair. The rows in each matrix correspond to rows of ts, while the columns correspond to rows of ms. The double matrix ts_ms.dist gives the horizontal distance between each potential pair of stations, while the logical matrix ts_ms.match is true where the values in ts_ms.dist are less than or equal to hzlimit, and false elsewhere. The structure also contains fields ts_ms.*_resid, ts_ms.*_std and ts_ms.*_nobs for each test variable; these provide the mean, standard deviation and number of individual data point residuals that correspond to each station pair.

4.6. Master cruises (mc)

The sixth output mc is a table in which each row represents a unique cruise in the master data set m. The column m.cruise is a list of the cruises, as a cell of strings. The columns m.ndate and m.ndate_range provide the mean, minimum and maximum date and time for the data used in the cross-over analysis for each cruise, in MATLAB's datenum format. The fields mc.*_resid, mc.*_std and mc.*_nobs contain the mean and SD of all residuals from each cruise, and the number of residuals, for each test variable *. These are calculated from all of the individual data point residuals for each cruise; they are not derived from the test or master station means.

4.7. Updated test data set (t)

The final output t is an updated version of the input data set t, including only the data points used in the cross-over analysis. The rows of this output correspond to the rows of the variables in t_ms.

5. Visualising and interpreting the results with XOVER_plots

The accompanying script XOVER_plots provides some tools to begin visualising and interpreting the outputs generated by XOVER. Brief descriptions are presented here. You can use whatever master and test data you like for this script, but to load the example data used as default by XOVER_plots, you need to use the function loadGLODAPv1 (Humphreys, 2015), which can be downloaded from http://uk.mathworks.com/matlabcentral/fileexchange/51961-loading-glodap-into-matlab. An example of how to implement loadGLODAPv1 to work with XOVER and XOVER_plots is included in the latter script. If loadGLODAPv1 and its associated data files (Humphreys, 2015), XOVER and XOVER_plots are all saved on the MATLAB search path[6], then XOVER_plots is set up such that it will perform an example cross-over analysis of one cruise from GLODAP (Key et al., 2004), using the rest of the GLODAP data as the master data set, and generate the figures in this documentation. XOVER_plots can then be easily modified to assess different variables or cruises, and to use any user-generated test data set.

5.1. Individual interpolations

The first figure generated by XOVER_plots illustrates the individual interpolations used to perform the cross-over analysis, similar to Figure 1.

Which test variable is shown on the vertical axis can be modified by changing the variable f1tvar in the script. Which master station's interpolation is shown can be adjusted through the variable f1ms_station; this input dictates which master station to show in terms of its row number in ms. Initially, the variable is set as follows:

```
f1ms_station = 1; + f1ms_station;
```

Running the code section[7] for the figure will therefore generate a plot for the master station in the first row of ms. If the first semicolon is then deleted:

```
f1ms_station = 1 + f1ms_station;
```

then the code section can be repeatedly evaluated (Control and Enter), and the figure will update each time to show the master station in the next row of ms. Once all rows have been shown, i.e. the value of f1ms_station becomes greater than the number of rows in ms, evaluating the code section will return an error. To return to the start, replace the first semicolon in the above line of script.

5.2. Map of cross-overs

The second figure generated by XOVER_plots is a map of the test and master stations (Figure 2), with the matching stations pairs indicated.

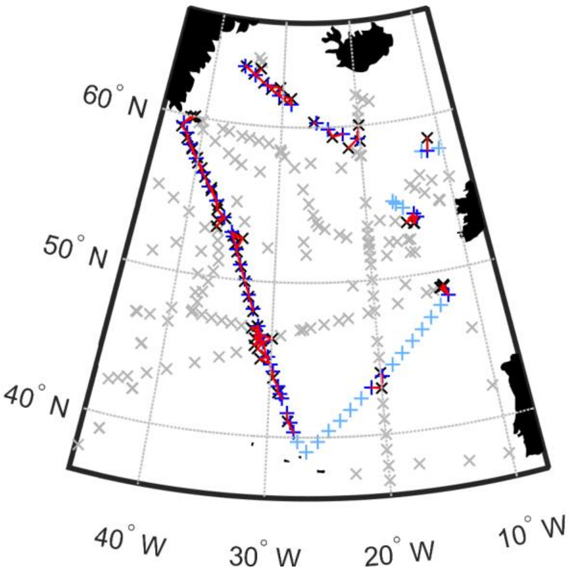

Figure 2. Example map of test and master stations generated by XOVER_plots. Grey × = all master stations; black × = master stations within hzlimit distance of a test station; light blue + = all test stations; dark blue + = test stations with matching master stations; red lines = test-master station pairs.

[6] http://uk.mathworks.com/help/matlab/matlab_env/what-is-the-matlab-search-path.html
[7] http://uk.mathworks.com/help/matlab/matlab_prog/run-sections-of-programs.html

5.3. Histograms of residuals

Histograms, and the mean and standard deviation of all residuals, provide a useful way to evaluate the cross-over analysis for each test variable. However, there are several different ways in which all of the residuals can be combined to achieve this goal, and it is left to the user to decide which is the most appropriate in each context.

The third figure generated by `XOVER_plots` illustrates three different ways that this could be done. The first subplot (Figure 3, left) histogram and statistics are derived as follows: at each test data point, the mean residual across all master stations matched to that data point is calculated, from the output `t_ms`. Consequently, each test data point is weighted equally in the histogram and its statistics, regardless of how many master data stations that test data point was matched to. In the second subplot (Figure 3, centre), a similar process was carried out, but calculating the mean residual at each test data station (from the output `ts_ms`), rather than at each test data point. In this case, each test data *station* is weighted equally, regardless of (i) how many master data stations it was matched to, and (ii) how many test data points there are at that station. In the third example (Figure 3, right), the mean of all test data point residuals at each master station is calculated, from `t_ms`. This time, each master station is weighted equally in the histogram and statistics, regardless of how many test data point residuals were calculated for each master station; and, the test data stations *are* weighted by the number of test data points that each one contains. The subplots also show the *p*-value of a one-sample *t*-test for the null hypothesis that the mean of the data in the histogram is from a normal distribution with a mean equal to 0.

Which test variable is shown in the histograms can be modified by changing the variable `f3var` in the script. The set of residuals that are used to generate each histogram and associated statistics can be modified through the variables `f3data1`, `f3data2` and `f3data3`.

There are many more possible ways to calculate the 'overall' residual, each with a different weighting of master and test stations and data points, that can be carried out using the information contained in outputs `t_ms`, `ts_ms` and `sp`. Ideally, if the test cruise data is internally consistent, and the master data set is internally consistent, and if there are many station pairs, then there should be little difference between these different ways of calculating an 'overall' residual.

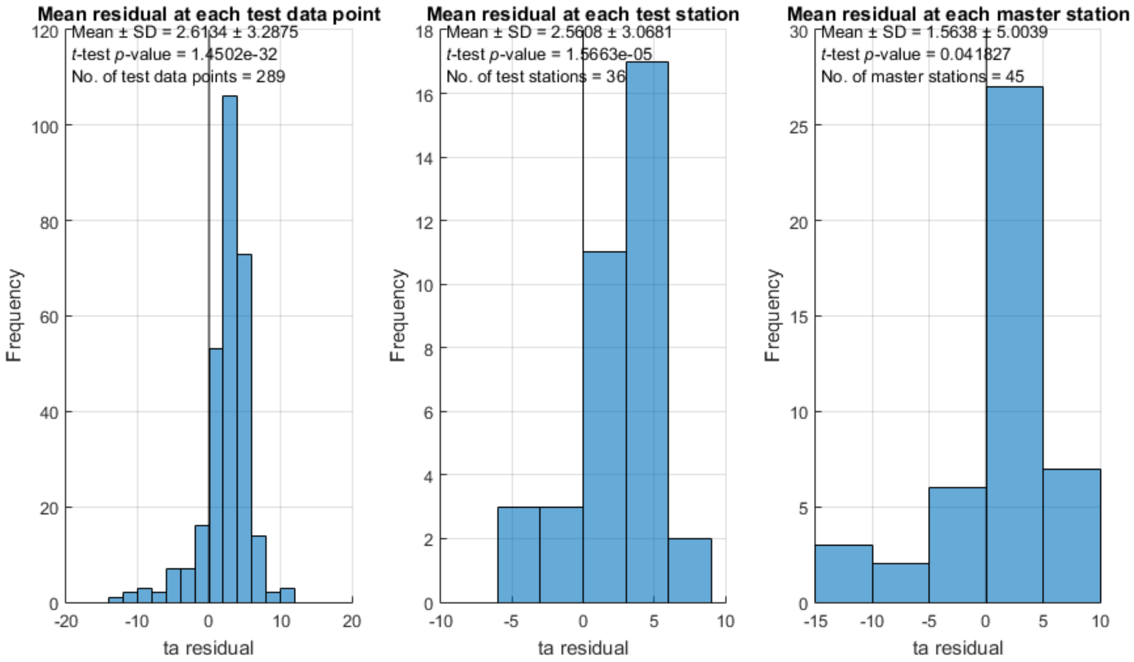

Figure 3. Example histograms of residuals, and associated statistics.

5.4. Cruise residuals and secular trends

The fourth figure generated by XOVER_plots can be used to identify secular trends in the residuals. For some variables in some parts of the ocean – for example, dissolved inorganic carbon at high latitudes – these could be real, and thus might not indicate a problem with the test data set. This figure (Figure 4) plots the mean residuals for the master cruises and their standard deviations against the cruise data, and performs an ordinary least-squares (OLS) regression to check for a trend, using the data in the output mc. The rate of change of the residual, the r^2 value for the regression, the regression-predicted value of the residual at the date of the test cruise (the 'test cruise intercept'), and the number of cruises (n) are shown. Small r^2 values indicate less chance of a secular trend in the residual; a test cruise intercept close to 0 is suggestive of a 'good' cross-over analysis result.

Which test variable is shown in the plot can be modified by changing the variable f4var in the script.

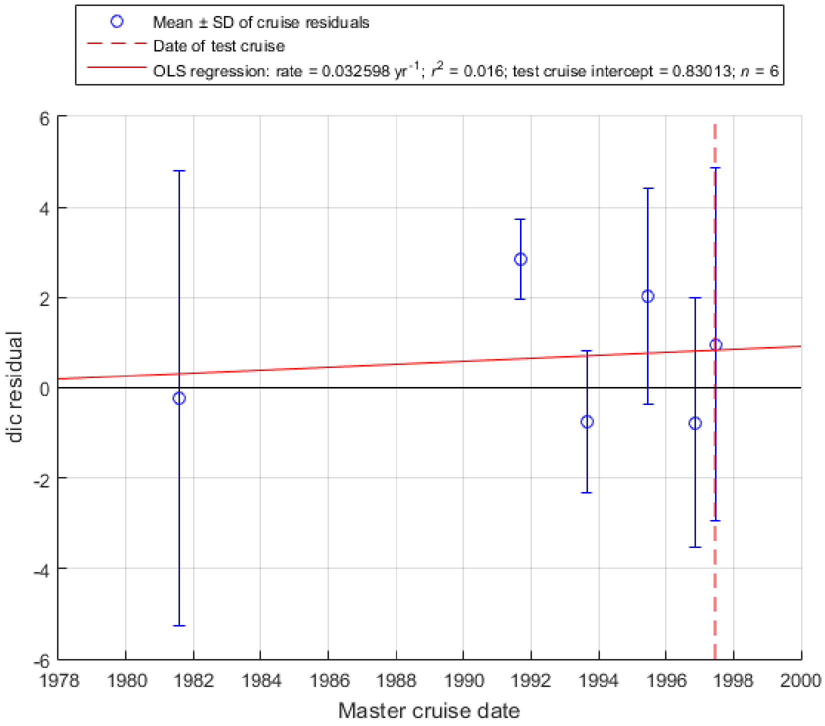

Figure 4. Example of plot to check for secular trends in residuals.

5.5. Relationships between residuals and other variables

In the event of poor cross-over analysis results, it may be instructive to evaluate the relationship between the residuals and other variables (Figure 5), for example using the fifth figure created by XOVER_plots. For example, this may highlight that there are elevated residuals for a test variable at shallower depths, and thus suggest using a deeper subset of the master and cruise data as inputs to XOVER for the cross-over analysis. It may also be possible to identify if any specific geographical region of the cruise has particularly anomalous residuals.

Which test variable residual is shown on the vertical axes can be modified by changing the variable f5var in the script. The variables on the horizontal axes can be similarly adjusted through the variable f5others.

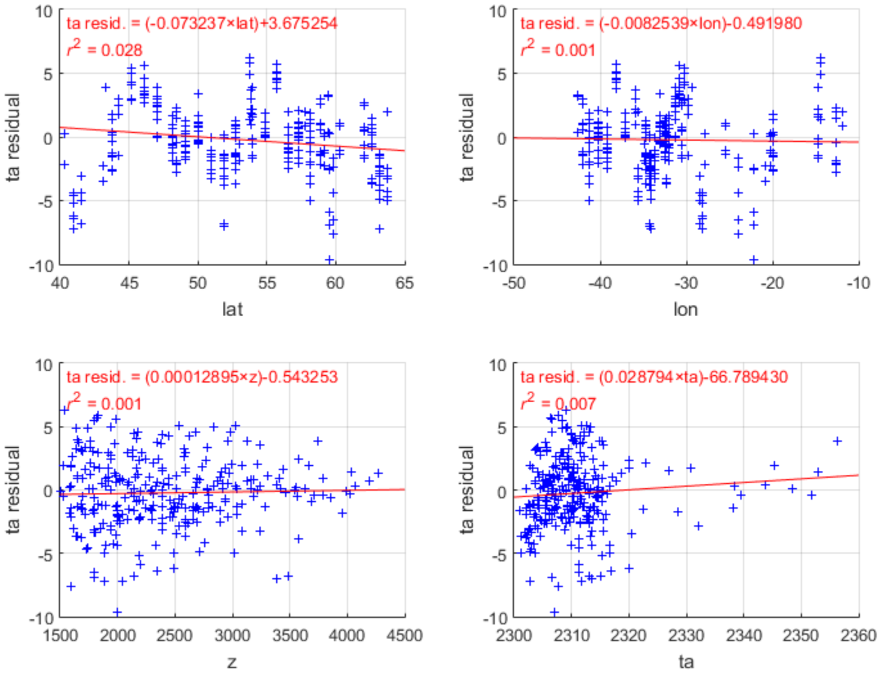

Figure 5. Example relationships between total alkalinity (ta) residuals and latitude (lat), longitude(lon), depth (z) and total alkalinity. In each panel, the blue + show the test data points, and the red line and text show the results of an OLS regression through the plotted data.

Acknowledgements

Matthew P. Humphreys is a postdoctoral research assistant for the UK Shelf Seas Biogeochemistry research programme, funded by the Natural Environment Research Council (UK). The XOVER program was inspired by Toste Tanhua's quality control toolbox for MATLAB (Tanhua, 2010). A developmental version of XOVER was used to perform a cross-over analysis by Humphreys et al. (2015).

References

Fritsch, F., Carlson, R., 1980. Monotone Piecewise Cubic Interpolation. SIAM J. Numer. Anal. 17, 238–246. doi:10.1137/0717021

Humphreys, M.P., 2015. Loading GLODAP into MATLAB. Ocean and Earth Science, University of Southampton, UK. pp 3. doi:10.13140/RG.2.1.2681.9687

Humphreys, M.P., Achterberg, E.P., Griffiths, A.M., McDonald, A., Boyce, A.J., 2015. Measurements of the stable carbon isotope composition of dissolved inorganic carbon in the northeastern Atlantic and Nordic Seas during summer 2012. Earth Syst. Sci. Data 7, 127–135. doi:10.5194/essd-7-127-2015

Kahaner, D., Moler, C., Nash, S., 1988. Numerical Methods and Software. Prentice Hall, NJ, USA.

Key, R.M., Kozyr, A., Sabine, C.L., Lee, K., Wanninkhof, R., Bullister, J.L., Feely, R.A., Millero, F.J., Mordy, C., Peng, T.-H., 2004. A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). Global Biogeochem. Cy. 18, GB4031. doi:10.1029/2004GB002247

Key, R.M., Tanhua, T., Olsen, A., Hoppema, M., Jutterström, S., Schirnick, C., van Heuven, S., Kozyr, A., Lin, X., Velo, A., Wallace, D.W.R., Mintrop, L., 2010. The CARINA data synthesis project: introduction and overview. Earth Syst. Sci. Data 2, 105–121. doi:10.5194/essd-2-105-2010

Suzuki, T., Ishii, M., Aoyama, M., Christian, J.R., Enyo, K., Kawano, T., Key, R.M., Kosugi, N., Kozyr, A., Miller, L.A., Murata, A., Nakano, T., Ono, T., Saino, T., Sasaki, K., Sasano, D., Takatani, Y., Wakita, M., Sabine, C., 2013. PACIFICA Data Synthesis Project. ORNL/CDIAC-159, NDP-092. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, USA. doi:10.3334/CDIAC/OTG.PACIFICA_NDP092

Tanhua, T., 2010. MATLAB Toolbox to perform secondary quality control on hydrographic data. ORNL/CDIAC-158. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, USA. doi:10.3334/CDIAC/otg.CDIAC_158