Software Manual - Section 3 (Reference)

Table of Contents:

General

The display has 2 parts: The lists, buttons and clickable areas to the left, and a grey **log panel** to the right. The log panel records commands, their results and the time it has taken to execute the command (in milliseconds), as well as errors and other messages. You can copy and paste data from the log panel.

Online vs. Offline Mode

As of version 1.0.5.0 the program can be run in either Online or Offline mode, which can be selected by the user. The *Online* checkbox in the *Manage Daily* tab show the current state. For details see the *Online checkbox* description in the *Manage Daily Tab* section below.

Start-Up

The program will carry out a number of activities and checks at start-up. The nature of these tasks varies depending on whether the program is running in **online** or **offline** mode:

Online

The Online startup procedures set up internal variables and the display screen. It fetches a list of all countries needed to populate the country combobox, thereby checking the status of the database. If specified by a parameter in the configuration file (see section on *supporting/configuration files* at the end of this document) the program will also start the MySQL database engine. The output in the log panel will look like this:

> **Initialising form and variables...**
> **MySQL started**
> **Form and variables initialised (10,535ms)**
>
> **All startup tasks completed.**

Any error messages here usually indicate that connection to the database or the Internet has not been established. The reasons for these are numerous, and you should contact the developer for advice.

> If the program is run for the first time in a new calendar year, it will attempt to load the station upload file from www.ClimateCHIP.org and update the *NOAAReadings* column in the *station table* accordingly (see coverage of the configuration file and station update file at the end of this document). This ensures accurate *data completeness* values to be shown in the station list later. If an update for a new calendar year has been detected, the program will prompt the user to update all their stations with the latest calendar year's data. Depending on the number of stations in the user's database this can take a while, especially in Online-mode. An estimate of the time it will possibly take is given, but this estimate will be rough and varies, among other things, with the available Internet bandwidth.

Offline

The Offline startup procedures set up internal variables and the display screen. It fetches a list of all countries needed to populate the country combobox, thereby checking the status of the database. If

specified by a parameter in the configuration file (see section on *supporting/configuration files* at the end of this document) the program will also start the MySQL database engine.

It also checks whether the NOAA index (more information about NOAA data below in the *Manage Daily Data Tab* description) is up to date. This should take only a few moments if your database is up to date. In this case the output looks like this:

> **Initialising form and variables...**
> **MySQL started**
> **Form and variables initialised (10,535ms)**
>
> **Checking NOAA Index...**
> **noaaindex exist (368ms)**
>
> **Reindexing NOAA files: . . . .**
> **0 NOAA entries indexed. (1,808ms)**
>
> **All startup tasks completed.**

Any error messages here usually indicate that connection to the database has not been established. The reasons for these are numerous, and you should contact the developer for advice. If you moved the location of the program or the NOAA data files, you will get an ***Invalid Data Path*** error message that explains that the program needs access to the data files in order to work properly. You will be given a choice to locate the data folder.

- If you click **No**, the program will stop trying to update the NOAA index and resume normal operation. It will not check if there is new data (e.g. a new calendar year) available. Also, you will not be able to import new stations.

- If you click **Yes**, the program will allow you to navigate to the NOAA data path and, if data files are found there, resume the re-indexing, concluding the startup tasks. If no data files were found the program skips the re-indexing task and resumes with the effects as described above (clicking No).

At any stage you can change the data path using the button near the bottom of the display (Manage Daily Data tab), which will trigger re-indexing the NOAA files.

If re-indexing NOAA files found new data the software will prompt the user to re-import all stations (see description in the Maintenance tab below) and give an estimated time this will take (calculated from the number of stations currently in the user's database). ("All stations" here means the stations the user has already in their individual database.) In most cases this is recommended as new NOAA data detected needs to be imported into update the user's database. In this case the import function will skip over existing records and be generally quite fast.

Manage Daily Data Tab

Station List

The table will show the FIPS country code, station number, station name as used by the NOAA GSOD data and level of data completeness of the stations found. Click on the table headings to sort. Click on a heading again to reverse the sorting order. Stations that are already in your database are highlighted with a yellow background.

The **data completeness** for each station is calculated as: *Total Number of NOAA readings* / (365 * *No of Years*). The *Total Number of NOAA readings* is the number of rows of data found in the GSOD data set over all years included in the data folder (*No of Years*). (There is one data row for each day where GSOD has at least one reading, e.g. Tmax and/or Tmean etc.

This completeness value should only be taken as guide: One can make no assumption to which parameter, temperature, wind speed, rainfall etc. the displayed completeness value applies, or indeed, if it applies to *any* parameter. E.g. a station may have Tmax readings for half of the time, and Td (dew-point temperature) for the other half. The completeness would be shown as 100% as there are data rows for the whole period, but Tmax and Td would really only have 50% data completeness each. In reality though, if there is a data row in the GSOD data, for most stations it is likely to contain Tmax, Tmean, Tmin, with the other readings tending to be less reliable.)

Completeness values exceeding 100% points to missing data files in the gsod folder. In a "normal" scenario data files starting with gsod_1980.tar to the last complete year should be present. E.g. in 2014 the last file would be gsod_2013.tar, a total of 34 files (years) would be in the folder. If all data files are present and the completeness values still appear incorrect, carry out a manual re-index of the NOAA data (see description of the Maintenance Tab below).

Buttons

Use the buttons described below to locate available stations. **Find Station**: To find a station by name, type any part of a station text string into the box and click the **Find Station** button (or press ENTER).

- Clicking **Find Station** (or pressing ENTER) with <u>*nothing*</u> in the text box: This shows **all** available stations, whether or not they are in the database. (This may take a moment or two.)

- **All in Database**: All stations for which data exists in our database will be listed.

Once a station in the list is selected/highlighted its *Longitude & Latitude* is displayed in the text boxes. You can override the longitude/latitude by typing in a geo location for use with the following function:

- **Find Nearby**: All stations in the vicinity are listed, using the specified distance from the displayed longitude/latitude:
  - o Enter a *number* to specify the distance to look for nearby stations. Choose the unit: *km, miles or degrees*.

o The search area may be a *circle* or a *square* around the currently displayed longitude/latitude. Example: If **1** is entered in the number field and **deg** as the unit and **Sq-Dist** is selected, the search area includes a field 1 degree north, east, south and west of the displayed longitude/latitude, effectively a 2 degree grid-square.

• **Change Data Path**: This button is only enabled (and relevant) in *offline-mode*. It tells the program where to find the NOAA data. (The location will be saved, along with other customised settings, in an XML file called HothapsSoftConfig.xml, which lives in the same folder as the executable file; see section on *supporting/configuration files* at the end of this document). All NOAA data are expected to be stored in a single folder, in the form of gsod_YYYY.tar files, with YYYY the year of the recording. As new annual data files become available place them into this folder. At start of the program, as well as after changing the NOAA data path using this button, the program will scan the files and re-index any entries not already in the database.

• **Online (checkbox)**: As of version 1.0.5.0 the program can be run in either Online or Offline mode, which can be selected by the user. The online mode is always available (subject to a working Internet connection), but offline mode should only be selected if the required data files are present on the user's computer:

o **Offline mode** (normal mode up to v 1.0.4.x) requires all supporting data files to be present on the user's storage medium. These files include all gsod_YYY.tar files up to the previous full calendar year, and the CRU database. These files require around 4 GB of disk space. In this mode no Internet connection is required to run Hothaps-Soft.

o **Online mode** (since v 1.0.5.0) needs access to an Internet connection to run. It does not require the GSOD data files or the CRU database to be present on the user's local storage. It will access GSOD data from the NOAA ftp site, obtain CRU from a service hosted by ClimateCHIP.org and will access a station-update file once a year from ClimateCHIP.org. In this mode the data footprint of the entire application is around 200 MB. This is the preferred mode for the downloadable version.

Context Menu

Select one or more stations in the list (use Ctrl or Shift-Click for multiple selections). You access the *context menu* by right-clicking anywhere on the station list. Some items are disabled, because they are for development and maintenance purposes only. The explanations of the disabled items are intentionally greyed out:

• **Analyse**: Analyses the database as well the raw NOAA files for the selected station(s), producing two separate results sets. Each result set shows the number of years with any data in them, the number of records found and the corresponding percentage of data density for those years e.g.:
Analysing NOAA for DALLAS HENSLEY FIELD NAS (722585-93901):
19 year(s), 6425 records (92.5%)
(Here 19 years containing any data were found in the NOAA data. 19 years have approximately 6,940 days. The 6,425 records found amount to 92.5% data density in those years.)

This is followed by a breakdown of the years containing any data, showing the data density for each year in brackets.

The analyses may take a while, depending on how many stations have been selected. Each analysis (NOAA and database) is a separate process, with its own progress window. This better utilises multi-core computers and leaves the software responsive, i.e. you can perform other tasks with the program without waiting for the analyses to complete. The results will be shown in the log panel upon completion. You may have to scroll the display in the log panel to view analyses completed earlier.

The NOAA analysis takes usually longer in online-mode than in offline-mode because the raw GSOD data file has to be downloaded from the NOAA website for each station analysed. The performance in online-mode greatly depends on the available Internet bandwidth and the responsiveness of the NOAA data server.

This menu item is disabled if the NOAA scan/indexing is in progress, or was unsuccessful (see discussion of error messages at start-up above).

- **Import NOAA and Calc Chart Avgs:** Combines two functions described below: Import NOAA data, and calculate monthly averages. This is the most common function for making weather station data available for this program.

  This menu item is disabled if the NOAA scan/indexing is in progress, or was unsuccessful (see discussion of error messages at start-up above).

  This function shows progress-bars in 2 stages (i.e. 2 times 0-100%): for (1) importing NOAA data, and (2) calculating averages. You can import several stations at the same time (in fact this will save time on a PC with a multi-core processor) but you must wait for all the progress bars to have finished before shutting down the program.

- **Export Daily Data:** Creates one comma-delimited text file containing daily records for each station selected, between the years displayed in **Start Year** and **End Year** boxes at the bottom. After selecting this menu item the user is prompted for the desired values (temperatures, dew point, wind speed etc. including any of the heat indexes). You can select any number of values (Ctrl or Shift-Click).

  Next the user is prompted to select the location (folder) for the generated files. The filenames contain the station, years and the selected values, followed by the csv extension, e.g.:

  CAIRO AIRPORT 1980-1990 , Temp Min, Dew Point, Wind Speed.csv

  The text files themselves contain a header, consisting of the station name and numbers, and a line containing the comma-delimited column headers, which will produce suitable column headers when imported into an Excel spreadsheet, e.g.:

  CAIRO AIRPORT (623660-99999)

  Rec Date,Temp Min,Dew Point,Wind Speed
  1/01/1980,9,2.05556,3.24275
  2/01/1980,11,2.11111,2.47067
  ...

  Missing data readings are indicated by a comma without a data value, e.g.:

  26/03/1980,,19.2778,2.50927

If there are no daily readings for a day, the record is not included in the exported text file, e.g.:

```
7/09/1982,14.5,13.8889,3.93762
8/09/1982,18,17.1111,2.04602
10/09/1982,19,17.1111,2.58648
12/09/1982,15.1111,13.7222,4.43948
```

Notice the missing records for the 9/9/1982 and the 11/9/1982. *(Depending on requirements it is envisaged to amend the program to include empty records in the file export to provide contiguous calendar dates.)*

The *Export Daily Data* option is only available for stations in the user's database (highlighted in yellow.)

- **Repair:** This option will delete and re-import this station's data completely; use if in doubt over a station's data reliability or integrity, caused by circumstances like the software or the PC has been closed or has crashed during importing or indexing.
  The *Repair* option is only available for stations in the user's database (highlighted in yellow.)

- **Import NOAA only**: Imports any available raw NOAA data. The log panel reports the years along with the number of daily records imported.
  Missing readings (e.g. 999.9 in the NOAA data) are recorded as *NULL* values. The import function will skip any years already imported. If you suspect any corrupted data in the database, it is best to delete the station using the delete function described below, then re-import and calculate averages.
  (The use of this function is for maintenance and diagnostic purposes only. Instead you are most likely to use the combined Import NOAA and Calc Chart Avgs function described above.)
  This menu item is disabled if the NOAA scan/indexing is in progress, or was unsuccessful (see discussion of error messages at start-up above).
  In offline-mode the relevant GSOD data is extracted from each of the years' gsod_YYYY.tar files found on the user's storage. In online-mode the relevant file for each station and year is downloaded from the NOAA website, and will usually be slower.

- **Calc Monthly Avgs for Charts only:** Extracts daily data from database and calculates average values for charting, and stores them in the *monthlyavg* table. The year range (normally from 1980 to last year) is derived from the available NOAA data. Monthly averages for Jan to Dec are recorded as well as annual averages. Only if the number of valid readings exceeds a certain threshold (this is 90% as of March 2010) will an average be recorded, otherwise null will be inserted into the database. I.e. only if there are 27 or more days with readings in any given month will a monthly average be recorded. Only if there are 328 or more days with readings in a given year will an annual average be recorded. Missing average values show up as gaps in the charts.
  Months and years that haven't got a single eligible average reading with regards to the 90% threshold described above will not generate a record in the *monthlyavg* table (as of v 1.0.3.0). The number of average data records is reported in the log panel. There are 13 records per year, i.e. 12 monthly and one annual average.

This function also calculates WBGT and UTCI heat stress indexes and inserts them into the database. As both indexes require temperature and dew point readings, only if the required temperature(s) <u>and</u> dew point readings have passed the threshold test (of currently 90% data density) will the heat index be calculated. Otherwise a null will inserted and this value will not be plotted inn the charts.

(The use of this function is for maintenance and diagnostic purposes only. Instead you are most likely to use the combined Import NOAA and Calc Chart Avgs function described above.)

- **Add Solar Data:** Will add data from a solar file (*-DLY-NNNN.txt) containing NASA Surface meteorology and Solar Energy data. Navigate to the appropriate file and click OK. (~16 s per 1000 records). Will honour years entered in "Start Year" and "End Year".
  (The development on this feature has been suspended and the function is no longer supported.)

- **Delete:** Deletes all daily data records and all monthly chart averages for the selected station(s) from the database. The number of records deleted is reported in the log panel.
  The *Delete* option is only available for stations in the user's database (highlighted in yellow.)

Charts Tab

Sort station table by any column (click heading, click again for reversing the order). Select any combination (using Shift- or Ctrl-click if necessary) of **stations, values** or **time modes**, then click one of the buttons to create charts.

Time modes will produce the following charts:

- **Annual:** <u>One averaged value per year</u> for each reading and/or heat index (one data point per year)

- **Jan, Feb, Mar…**: <u>The averaged Jan or Feb or Mar… value for each year</u> (one data point per year)

- **Monthly**: <u>Each month's average</u>, (12 data points per year)

Buttons

- **Chart Averages:** Produces a chart of averaged selected readings and/or heat indexes applied to the selected time modes and selected stations, between the years displayed in **Start Year** and **End Year** boxes at the bottom.

- **Chart Thresholds:** Produces a chart of the number of days where the selected actual daily readings and/or heat indexes exceed a threshold, applied to the selected time modes and selected stations, between **Start** and **End Year**. You will be prompted to enter a threshold value in a dialog box after clicking this button. *(For a future version it is envisaged to allow entering*

multiple threshold values for multiple value choices, e.g. specify threshold 26 for WBGT and 30 for UTCI in the same graph.)

Context Menu

Click any combination (using Ctrl- or Shift-click) of values, time periods or stations, then access the *context menu* by right-clicking anywhere in station list. Note that the values in the **Start** and **End Year** boxes at the bottom of the screen apply to all of the following functions:

- **Analyse (Basic)**: Shows how much data is available for charting: Complete years, partial years (giving percentage of completeness) and number of years with no data.

- **Analysis (Detailed)**: Shows a grid of how many values (dew point, temp max, temp min etc.) are available for charting for the yearly chart and each month per year. A "7" means all values have averages for this month/year. Whether a value appears here as fit for charting depends on the threshold applied (currently 90%) when calculating the averages in the Daily Data section (above). A "0" means no reading is available at all. Any number in between means that some values are missing; e.g. temperatures are often recorded while dew point or wind speeds are sometimes not. You will have to plot the graphs to find out which values are available.

- **Delete**: This will remove the calculated averages for the selected station(s), effectively removing the ability to plot this station. The station will no longer show in the list. This function is for maintenance and diagnostics only as it leaves the raw data in the database, and is not available in the public version of the software. If you want to remove a station completely, use the Delete function in the *Manage Daily Data Tab* (page 10).

The following functions calculate values from the database and store the results in the clipboard in a format (tab-delimited plaintext) ready for pasting into a spreadsheet, for example. The results have a header-row. All calculations contain the following common columns, e.g. for Manila:

Country	Station (Type)	USAF-WBAN	Long	Lat	From	To	....compl
RP	MANILA (Annual)	984250-99999	120.983	14.583	1980	2010	61.30%

- The columns are largely self-explanatory. The last column is added at the end of each calculated value to indicate the data density ("completeness") of the underlying data.

- **Export Average (daily** and **monthly)**: This function calculates the average and standard deviation (SD) for the selected stations between selected Start and End Year. The averages are calculated according to the following table:

Selected *Time-Value*:	**Annual**	**Jan, Feb…**	**Monthly**
Export Average (daily):	One avg. of all daily readings	One avg. of daily readings for each of the selected months	n/a
Export Average (monthly):	One avg. of annual averages	One avg. of monthly averages for each of the selected months	Avg. of all monthly averages

Note that the following statistical calculations, except for the Regression (Thresholds) are derived from averaged (monthly and yearly) values. As outlined above, if any month or year didn't reach the data density threshold (of currently 90%) this data point is not recorded and therefore excluded from these statistical calculations. For example, if a station had only 200 daily temp max readings for a particular year, no annual value would be stored in the average table for that year, and these 200 temperature readings would be excluded from the averages and regression statistics. The reason is that it is not obvious which days of the year are missing; if the remaining recorded days were seasonally biased, the result of the calculated statistic would also be biased.

The threshold regression, on the other hand, includes all daily readings, irrespective of gaps in the data. Missing readings, when counting the number of days where a specified threshold was exceeded, were not seen as being able to unduly bias the regression.

- **Export Regression (Averages)**: Calculates the slope (a), offset/intersect (b) and standard error (SE) for the selected station(s), values and time-modes. Slope and standard error are given in unit/year. "Unit" here is the obvious unit, i.e. °C in most cases, mm and m/s for rainfall and wind speed respectively. (Rainfall is actually in mm/day.)
  This is an example output (the common station columns have been omitted):

Temp Max (a)	(b)	SE	compl	WBGT(max) (a)	(b)	SE	compl
0.03796	11.417	0.0161	100.00%	0.04766	8.8222	0.0144	100.00%

- **Export Regression (Thresholds):** The program will prompt for a threshold value. It will calculate and produce the slope (a), offset/intersect (b) and standard error (SE) for the selected station(s), values and time-modes, relative to the number of days above the specified threshold. The output is comparable to that shown for average regression above. The slope and errors are given in days/year, the intercept is in days.

- **Export Slope Residuals:** This sums the differences between the data values of the last 10, 8, 6 and 4 years and the trend line/slope of the series.

- **Export Average Residuals:** This sums the differences between the data values of the last 10, 8, 6 and 4 years and the series' average.

Chart SD Tab

The name "Chart SD" has its origin in "Standard Distribution", and can now be associated with charting "statistical distribution" of the data contained within the user's database.

Sort the station table by any column (click heading, click again for reversing the order). Select any combination (using Shift- or Ctrl-click if necessary) of **stations, values** or **year ranges**, then click the *Chart Distributions* buttons to create charts.

A list of *year-ranges* allows charting data distributions over various time periods:

- Use the **Add** button to add a new year-range to the list. The resulting dialog box allows you to enter start and end years for the period. (Start and End year inputs are not validated other than Start year cannot by greater than the End year. If you enter years outside of the data available you will get strange or empty charts.)

- **Double-Click** on a year-range entry to edit its **Start** and **End years**.

- Click **Delete** to remove unused year-ranges.

Maintenance Tab

Most maintenance features are disabled in the public version of the software. These functions are necessary for creating and initialising the database, and are documented below for the benefit of the administrators of the software. The explanations refer to a number of files that need to be available in order to perform the database creation. The explanations of the disabled items are intentionally greyed out.

Importing of stations should only be necessary once. For the first time (or to override the existing data) choose *Import Timezone CSV*. We need the *Timezone Calcs.csv* (headings: ISO, Lat, Long, Name, UTC, DST, Comment), approx. 400 records, derived from Timezone Calcs.xlsx, massaged data from a website (Wikipedia?). This loads the records into memory only, and is used to add UTC time offsets to station data.
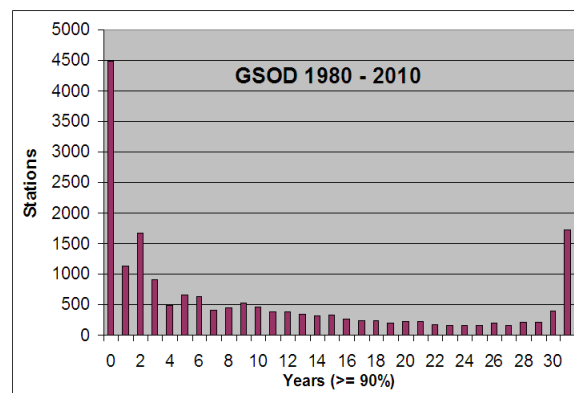
Then run *Import Station CSV*, opening *ish-history cleaned.csv* (headings: USAF, WBAN, STATION NAME, CTRY, FIPS, STATE, CALL, LAT, LON, ELEV(.1M)., approx. 26,600 records. This file is derived from NOAA's ish-history, purged of bogus, duplicate and buoy stations. The data is imported into the database table *station*. The program will calculate and assign an approx. UTC offset for each station by locating the closest time zone imported above (using Lat/Long and Pythagoras). This might need optimising, perhaps weighting longitude over latitude?!

Even after purging bogus, duplicate and buoy station in the *Import Station CSV* function above, many stations contain no data at all. The *Purge 0-Data Stations* button scans the NOAA index in the database to locate data-less stations in the range of years entered at the bottom of the form (e.g. 1980-2012). This may take 5-7 minutes. The purpose is to prevent locating dud stations during analysis. The

program will show a message box showing you how many stations with 0 data have been found and asking for confirmation to delete them. It is likely that this has already been performed in the database version you are using.

(I have been hesitant implementing the Purge 0-Date Stations functionality: While in the time-span specified at the bottom of your form, say 1980-2010, there is no data for a given station, and in the future you will make more NOAA data available, either from future or past years, it is theoretically possible that there may now be data for this station. Once purged with this button, you cannot locate the purged station any longer, import or graph it. You could however, re-import the Station CSV as explained above, possibly reporting a lot of duplication errors as the majority of stations have already been loaded.)

The **Station Statistics** function searches the stations in the database for available readings. It counts the number of stations that have near-complete readings (more than 328, i.e. >= 90% of 365, annual readings of any sort, e.g. temperature, dew point etc.) for each year within the specified range (start year, end year) and displays the accumulative results in the output pane. Depending on the year range specified this analysis may take several hours! An analysis carried out in 2011 resulted in the graph shown here (before applying the Purge 0-data Station function described above.)

The **Delete Null Avgs** button removes any record in the monthlyavg table that contains only nulls. (The monthlyavg table is used for most of the time charts produced by this software.) From version 1.0.3.0 onwards monthly avgs records don't have to be contiguous and removing empty records makes the database smaller (up to 20%) and faster. From version 1.0.3.0 onwards any new importing will not generate null records in the monthlyavg table.

As of version 1.0.2.0 CRU data has been included for comparison and quality control. **Import CRU Data** needs 3 Tab-delimited text-files, one for each of the 3 values: mean (avg), max temperature and vapour pressure. The files are expected to be reduced to relevant land-based grid-cells and limited to the years between 1980 and 2009 (at the time of writing). These files (on the developer's file system) are called tmp1980onLand.txt, tmx1980onLand.txt and vap1980onLand.txt. The import function creates a database table and indexes for the CRU data and will take several (12-20) hours.

The **Calc CRU Averages** button calculates one annual average record for each year and grid-cell and inserts it into the cru table.

The **Add WBGT/UTCI (CRU)** button calculates and adds 3 WGBT and 3 UTCI values to each cru record (assuming these fields have already been added to the table structure). This process took around 24 hours (on an admittedly slow machine).

Note 1: The sequence during development of executing **Calc CRU Averages**, then **Add WBGT/UTCI (CRU)** resulted in:
Annual WBGT & UTCI for CRU = WGBT&UTCI (annual avg Ta & Td).
This is in contrast to the way heat stress indices are calculated for the main (GSOD) database, where:
Annual WBGT & UTCI for GSOD = average (monthly WGBT&UTCI)
A separate routine in "DB_Utilities" has corrected this, and re-calculated heat stress indices to be:
Annual WBGT & UTCI for CRU = average (monthly WGBT&UTCI)
(This process took around 12 hours.) Cru tables distributed from v 1.0.3.1 onwards have these corrected heat stress indices.

Note 2: Up to version 1.0.3.0 the cru table has been organised by: Year, Month, Lat, Long. This was based on the original organisation of the CRU data and is useful for GIS type processing where you provide a year/month combination and after that you need speedily access to all available grid-cells. However, this makes for poor performance (multiple *seeks*) in the scenarios that this software is used for: look up a 30-year time frame for one given grid-cell. As of v 1.0.3.1 the cru table is organised by Lat, Long, Year, Month which greatly improves performance, especially for "Monthly" charts. (CRU data processing is included in a separate piece of software called "DB_Utilities". Re-arranging the large CRU data table in the way described above can also be achieved using standard SQL statements, which can run in a reasonable timeframe, e.g. a few hours, on powerful computers e.g. TALOS at NMIT.)

Optimise Database button issues the relevant command to the underlying database engine for the daylyrecords and monthlyavg tables. This will remove empty blocks and rebuild the indexes. It is useful after a lot of deleting and re-importing NOAA data into the database.

Re-Import All Stations will go through all your stations in the database and try and re-import NOAA data and calculate monthly & annual averages where necessary. This is useful once new NOAA data has become available, e.g. after a new year's dataset has been downloaded. This function will not do anything if all your data is up to date with the available NOAA data. This function will be called if new NOAA data is detected either at start-up of the software, after changing the NOAA Data Path or after re-indexing NOAA date (see below). This process is spreads the workload over several *background threads*. a) This means you can carry on doing other things with the software (and the software will remain responsive) while it is re-importing, and b) it will adapt the processing pattern to the available number of *cores* in your computer and balance the processor load by creating one background process for every core. (In theory this should maximise performance on your machine while not locking it up.) An estimated processing time will be displayed.

Re-index NOAA data: This button is disabled in Online-mode. In Offline-mode the NOAA data is expected as a series of downloadable files from the NOAA website. The files are in the form of **gsod_YYYY.tar**, with YYYY the year of the recording. At the start of the software these NOAA files are scanned and an index recorded in the database. This index is necessary for the rapid access of any station in any year. It is likely that your database already contains a complete NOAA file index, in which case this process will quickly report that 0 entries have been indexed (as in the screenshot example above). However, if you experience problems with the NOAA data (analysing, importing, see below) you

can use the ***Re-index NOAA data*** button to repair the index. (Essentially it deletes the entire index on the database and recreates it. This is not a background process, so the program will be unresponsive for minutes or hours, depending on the computer's processing speed, number of years and size of the NOAA files.)

If re-indexing NOAA data was successful the software prompts the user to re-import all station (see above).

Charts Window

The chart feature of the software produces X, Y line charts in separate, resizable windows. You can customise certain properties of the chart as outlined below and your preferences will stored in the HothapsSoftConfig.xml file.

The x-axis (time continuum) is governed by the **Start** and **End Year**s entered at the bottom of the main window. The yearly divisions mark the <u>beginning</u> of a year. For annual averages data points are plotted in the middle of the year; monthly averages are positioned relative to their

If any month or year contains fewer readings than the threshold (currently 90%) of the year or month, that data point will not be plotted. If a graph appears completely blank (heading only), no data is available for the selected combination of values, time modes and station(s). Go to the *Analysis* option (Context Menu, above) to find out why.

Graph Menu

- **Lines, Markers, Heading:** These are on/off options and quite self-explanatory. Note that turning **Markers** on may show data points in patchy data series that are not visible in Lines-only mode. Lines will only show if they can connect at least 2 existing consecutive data points (years).

- **Legend:** You can turn legends on or off. If they are on, you can **Customise Legends** with regards to the position and font size.
  Legends show a combination of the following. Information common to all series is displayed in the chart title:
    - Station name
    - Reading type (e.g. Jan, Feb, Annual or Monthly)
    - Reading value (e.g. Temp Max, Wind Speed) and unit (e.g. [°C], [mm], [m/s]). Note that rainfall is displayed as an average of daily readings, so it is to be taken as mm/day.
    - The percentage value in brackets is the data completeness of the series found in the raw NOAA data.
    - If trend lines are turned on (see below), the type of trend line is displayed along with the slope value in units per year.

- **Fixed Y-Axis (NN - MM):** If this option is turned **off**, the program will derive a suitable range for the Y-axis from the plotting data. If you want to fix the Y-axis (e.g. to produce a number of graphs with the same scaling), turn this option **on**. You can customise the Y-axis's maximum and minimum values as well as both axes' interval and font size, as well as showing or hiding gridlines.
  - The X-axis's maximum and minimum are determined by the Start and End Year values entered on the program's main window.
  - You can set the axes' font sizes even if Fixed Y-Axis is turned off.
  - You can adjust the X-Axis Interval even if Fixed Y-Axis is turned off.
  - The Y-Axis's *automatic* box applies to the Interval only.
  - Changing the Y-Axis Min or Max turns Fixed Y-Axis on when you click OK.

- **Colour Mode:** Turn this option off if you need to publish graphs in black & white. The software will select at this stage non-customisable markers and lines thicknesses.
  Note: With colour mode on or off (i.e. even in black & white mode), you can use the **Set Colours…** submenu to customise the colours for each data series.

- **CRU Reference:** This option overlays a line graph of the CRU data for the relevant grid-cell of the station(s) shown. The latitude/longitude of the matching CRU data is shown in its legend. CRU data is available for Temp Min, Temp Max, Temp Mean, Dew Point and all WBGT and UTCI indices. At the time of writing (Apr 2014) CRU data is available from 1980 to 2011. The colour of the (dashed) CRU reference graph initially matches the colour of the original Station graph but can be customised using the colour mode settings described above.
  In Online-mode a working Internet connection must be established, as the CRU data is obtained online (via the CLimateCHIP web-service.) As a result, depending on the quality of the Internet connection, producing charts may be slower than in Offline mode.

Trend Lines Menu

Rather self-explanatory, you can turn on/off any of Linear, Last Decade (linear), Logarithmic and 3 degrees of Polynomial trend lines.

Export Menu

You can export charts in **JPG** or **PNG** format. While JPG is popular, the PNG format uses lossless compression and is the preferred format for line graphs. The program will prompt for the location of the files and remember your preference. The file name is taken from the chart's heading.

Note that the resolution of exported graphic equals the current display size. So if you shrink the window the resulting graphic file will also be smaller. If you need to export a uniform set of graphs for publication, either stick to the default size of each chart window (i.e. do not resize windows) or maximise each chart window before exporting.

All graphed data can also be exported in **Numeric** form to a csv file. After choosing the Export/Numeric menu the software will ask for a destination folder and save the file using the chart's heading as the

filename. Each graph is represented by 2 text rows: one for the X, another for the Y values, each prefixed with a suitable row label.

Charts SD Window

The statistical distribution window show box and whisker plots for each month of the year range selected in the Chart SD Tab. Within the plotted year range and selected station each box and whisker plot shows:

- Whisker percentiles (top and bottom), default 10%
- Box percentiles (top and bottom), default is 25%
- Average/mean (solid line)
- Median (dashed line)
- Outliers

All of these parameters can be customised:

Graph Menu

- **Heading:** Turn headings on/off (some publishers like to use their own headings, or headings are shown in a caption below the graph)

- **Legend:** You can turn legends on or off. If they are on, you can **Customise Legends** with regards to the position and font size.
  Legends show a combination of the following. Information common to all series is displayed in the chart title:
    - Station name
    - Reading value (e.g. Temp Max, Wind Speed) and unit (e.g. [°C], [mm], [m/s]). Note that rainfall is displayed as an average of daily readings, so it is to be taken as mm/day.
    - Year range, e.g. "1990-2000"

- **Colour Mode:** Turn this option off if you need to publish graphs in black & white. The software will select at this stage non-customisable markers and lines thicknesses.
  Note: With colour mode on or off (i.e. even in black & white mode), you can use the **Set Colours…** submenu to customise the colours for each data series.

- **Fixed Y-Axis (NN - MM):** If this option is turned *off*, the program will derive a suitable range for the Y-axis from the plotting data. If you want to fix the Y-axis (e.g. to produce a number of graphs with the same scaling), turn this option *on*. You can customise the Y-axis's maximum, minimum values and the interval, as well as showing or hiding gridlines. Changing Min, Max or Interval turns Fixed Y-Axis on when you click OK.

- **Percentiles:** Customise the percentile distribution displayed in the box & whisker charts.

- **Show Median/Average/Outliers.** Turn these chart options on or off.

Export Menu

You can export charts in **JPG** or **PNG** format. While JPG is popular, the PNG format uses lossless compression and is the preferred format for line graphs. The program will prompt for the location of the files and remember your preference. The file name is taken from the chart's heading.

Note that the resolution of exported graphic equals the current display size. So if you shrink the window the resulting graphic file will also be smaller. If you need to export a uniform set of graphs for publication, either stick to the default size of each chart window (i.e. do not resize windows) or maximise each chart window before exporting.

All graphed data can also be exported in **Numeric** form to a csv file. After choosing the Export/Numeric menu option the software will ask for a destination folder and save the file using the chart's heading as the filename. Each graph is represented by 8 text rows, each prefixed with a suitable row label, one for each station and year range. There is one labelled column value for each month. The rows are:

- Highest (outlier), only if "Outliers" are turned on in the graph window
- Whisker high
- Box high
- Box low
- Whisker low
- Lowest (outlier), only if turned on in the graph window
- Mean
- Median

Supporting (Configuration) Files

HothapsSoftConfig.xml

This file is read and written to during various phases of running the program. In successive versions more parameters were added. It is important to always have the latest HothapsSoftConfig.xml file with new releases of the program. This file is expected to be in the same folder as the exe.

<StationUpdateURL>

New in v 1.0.5.0, this parameter contains the url of a file (see stationupdate.dat below) needed in online mode when NOAAReadings column in the station table needs to be updated. This will only be triggered (in online mode only) if the current year is greater than the entry in <YearLastUpdated>. (In offline mode the presence of a new gsod_YYYY.tar file causes NOAA re-indexing and a recalculation of the NOAAReadings column that way.) Currently (as of 2014) this is
http://www.climatechip.org/resources/stationupdate.dat

<YearLastUpdated>

New in v 1.0.5.0, this parameter contains the year when the last online update of the NOAAReadings column (see above) was successful. It was introduced only to coordinate this update process, i.e. to answer the question: has the NOAAReadings column been updated to reflect the last full calendar year.

<NOAA_URL>

New in v 1.0.5.0, this parameter contains the url of the GSOD files needed in online mode. Currently (as of 2014) this is ftp://ftp.ncdc.noaa.gov/pub/data/gsod

<OnlineMode>

New in v 1.0.5.0, this parameter (true or false) remembers the online/offline mode setting.

<GridLinesY>, <GridLinesX>, <ShowBoxOutliers>, <ShowBoxAverage>, <ShowBoxMedian>, <WhiskerPercentile>, <BoxPercentile>, <Markers>, <Lines>, <ColourMode>, <LegendPosition>, <IntervalX>, <IntervalY>, <AxisFontSize>, <LegendFontSize>, <MaxY>, <MinY>

These parameters remember many of the user's graphic settings.

<NOAA_Path>

Local path of the user's folder containing gsod_YYYY.tar files, or the string "Invalid Data Path!". Needed in offline mode only.

<MySQL_Path />

New in v 1.0.5.0, this parameter contains the location of the mysql folder. In the common USB-stick version this would be **..\mysql**. It is used to launch mysql from within the software (not via batch file up to v 1.0.4.x). In the developer's version (mysql service running), this tag is empty.

<YearRanges>

Holds multiple sets of **StartYear/EndYear clsYearRange** parameters, used by SD-Charts.

Hothaps-Soft.exe.config

Common config file needed by the .NET framework. It contains:

<connectionStrings>

For use with the Entity Framework as well as standard ADO, 2 connection strings are supplied. In the 2014 version this reads for:

Local MySQL service version:

```
<add name="climatedata09Entities"
  connectionString="metadata=res://*/ClimateData09.csdl|res://*/ClimateDa
  ta09.ssdl|res://*/ClimateData09.msl;provider=MySql.Data.MySqlClient;pro
  vider connection string=&quot;server=localhost;User
  Id=root;database=climatedata09&quot;"
  providerName="System.Data.EntityClient" />
<add name="climatedata09ADO"
  connectionString="Server=localhost;Database=climateData09;
  User Id=root;" />
```

USB-stick version:

```
<add name="climatedata09Entities"
  connectionString="metadata=res://*/ClimateData09.csdl|res://*/ClimateDa
  ta09.ssdl|res://*/ClimateData09.msl;provider=MySql.Data.MySqlClient;pro
  vider connection string=&quot;server=127.0.0.1;port=3307;User
  Id=root;database=climatedata09&quot;"
  providerName="System.Data.EntityClient" />
<add name="climatedata09ADO"
  connectionString="Server=127.0.0.1;port=3307;Database=climateData09;
  User Id=root;" />
```

<system.data>

<DbProviderFactories>

Only needed for the USB-stick version (2014):

```
<remove invariant="MySql.Data.MySqlClient"/>

<add name="MySQL Data Provider" invariant="MySql.Data.MySqlClient"
    description=".Net Framework Data Provider for MySQL"
    type="MySql.Data.MySqlClient.MySqlClientFactory, MySql.Data,
    Version=6.6.6.0, Culture=neutral,
    PublicKeyToken=C5687FC88969C44D" />
```

<system.serviceModel>

New in v 1.0.5.0 , the online version can read the CRU data from a ClimateCHIP service. For this we need the following:

```xml
<bindings>
   <basicHttpBinding>
      <binding
            name="BasicHttpBinding_IClimateCHIPService" closeTimeout="00:01:00"
            openTimeout="00:01:00" receiveTimeout="00:10:00" sendTimeout="00:01:00"
            allowCookies="false" bypassProxyOnLocal="false" maxBufferSize="500000"
            maxBufferPoolSize="524288" maxReceivedMessageSize="500000"
            messageEncoding="Text" textEncoding="utf-8" transferMode="Buffered"
            useDefaultWebProxy="true">

            <readerQuotas maxDepth="32" maxStringContentLength="8192"
                  maxArrayLength="16384" maxBytesPerRead="4096"
                  maxNameTableCharCount="16384" />

            <security mode="None">
                  <transport clientCredentialType="None" proxyCredentialType="None"
                        realm="" />
                  <message clientCredentialType="UserName"
                        algorithmSuite="Default" />
            </security>
      </binding>
   </basicHttpBinding>
</bindings>

<client>
      <endpoint address=http://services.climatechip.org/ClimateCHIPService.svc
            binding="basicHttpBinding"
            bindingConfiguration="BasicHttpBinding_IClimateCHIPService"
            contract="ClimateChipService.IClimateCHIPService"
            name="BasicHttpBinding_IClimateCHIPService" />
</client>
```

stationupdate.dat

Tab delimited text file (STN, WBAN, total NOAAReadings up to the previous full year. Instructions on how to create this file are in the procedure header of clsStation.UpdateNOAAReadingsOnline() in the source code.

14 August 2014