Last Updated: February 2012

LTER and U.S. Forest Service Climate/Hydrology Database Guidelines

i. Table of Contents

1.0. Introduction

The National Science Foundation's Long-Term Ecological Research (LTER) program and many U. S. Forest Service Experimental Research Stations collect and maintain extensive, long-term ecological databases including streamflow and meteorological measurements. These databases have been widely used in intersite comparisons, modeling studies, and land management-related studies. To facilitate intersite research among the network of LTER sites, information managers have developed a prototype to provide climatic summaries dynamically over the Internet (http://www.fsl.orst.edu/climhy/), and serves as one model for improving access to data across sites (Baker et al. 2000, Henshaw et al. 1998). Individual sites maintain local climate data in local information systems while a centralized site continually harvests, updates, and provides access to all sites' data through a common database. Common distribution report formats and graphical displays have been established to meet specific needs of climate data users.

Funding from the U. S. Forest Service has allowed the climate data prototype (ClimDB) to be improved and expanded to include hydrologic variables (HydroDB). Mechanisms for capturing appropriate metadata essential for discovery and interpretation of the hydroclimatological records are also developed. Report formats and graphical displays have been updated for the hydrological data. Enhancements to the existing harvester allow the prototype module to truly function as a production module. Most recent enhancements have combined the two modules and have made capturing and accessing the data seamless.

1.1. ClimDB Overview

Long-Term Ecological Research (LTER) sites have generally followed established LTER Climate Committee guidelines (Greenland 1986) for collecting baseline meteorological data. Standardized measurements provide a basis for coordinating meteorological measurements at two or more sites and enable intersite comparisons. However, access to comparable datasets from multiple sites is problematic. While most sites make climate data accessible via the World-Wide Web (WWW), the data are displayed in a variety of formats, are aggregated using different methods, and are often not easily located.

A project to conduct climatic analyses of the LTER sites (CLIMDES) gathered individual site temperature and precipitation data (1960-1990) and created on-line monthly summaries for each site (Greenland et al. 1997). While the CLIMDES project satisfied an immediate need for access to monthly site climate data, no mechanisms were established for updating these summaries. With synthesis groups needing ready access to current climatic summaries, a system to provide climatic summaries dynamically over the WWW is needed. ClimDB was developed in response to this science-driven need.

1.2. HydroDB Overview

Twenty-three Forest Service experimental sites with long-term hydrologic and associated meteorologic data have been funded to establish web access to existing long term data sets. Access will facilitate use of these data to improve estimates of postfire flood risk and other scientific and practical uses. Forest Health Monitoring is seeking to increase the accessiblity of long-term data on-line by funding linking long-term electronic data sets to a central "web harvester". This central portal can provide direct access to long-term data sets via the world wide web for a variety of uses including Fire Evaluation Monitoring.

Long-term data sets of interest include streamflow (l/sec) for gaged watersheds with corresponding precipitation (mm) and ambient air temperature (^{o}C) data that were collected simultaneous with hydrologic data and represent conditions in the watershed. Data collected at daily or more frequent intervals with a data record longer than ten years are preferred. Shorter data records will be considered if they are part of a current program designed to collect data for longer than ten years. Metadata describing site conditions and methods of data collection and processing will also be required and must conform to specific content and format standards that are under development.

1.3. USGS Data Overview

ClimDB/HydroDB how has the ability to harvest streamflow data from any real-time USGS gauging station and processing it for submission on a weekly basis. For more information, visit http://gce-lter.marsci.uga.edu/lter/research/tools/usgs_harvester.htm.

1.4. Project History

In the fall 2003, it was decided to merge ClimDB and HydroDB. The back-end database has always been seamless, but the front-end interfaces have been different. Therefore, there is only one place to go to get data and another for participants to harvest data and update metadata. For more information, see ClimDB/HydroDB Progress Reprot (17 Mar 2003) and ClimDB/HdyroDB (ClimHy) Database Update (Spring 2005).

In June 2010, the ClimDB/HydroDB database was migrated to a server at the LTER Network Office (LNO). The migration occurred to relieve the support and administrative burden from the Andrews LTER site and LNO welcomed hosting this key Network Information System (NIS) module. LNO hosting should improve efficiency in the eventual integration of this module into the new NIS architecture. For more information, see ClimDB/HydroDB (ClimHY) Database Migration to LNO (Fall 2009).

1.5. Required Steps for Site Participation

To participating the site will:

1) Provide the names of research areas, meteorological stations, gauged watersheds, and gauging station names and code names to the ClimDB/HydroDB administrator. These names must be in the central database before any test harvest can proceed. Additionally, provide the names, addresses and email addresses for a data contact person as well as all interested principal investigators.

2) Use the online metadata forms to provide metadata for overall research area, for every weather station and for every parameter measured at each station, watershed characteristics of gauged watersheds, and every gauging station. (See section 4 for metadata categories and descriptors).

3) Provide appropriate quality assurance parameters for every measured parameter as part of the metadata for central database validation checking (See Section 3.0). Otherwise the global defaults are assumed (section 3.3).

4) Restructure local site data into a standardized daily exchange format (See section 2.0). This process can occur on a scheduled basis into static files, or can be created dynamically during the harvest process. The web service, last_harvest, can be used to obtain the last date in the ClimDB/HydroDB database (See Section 1.8).

5) Provide an Internet address (URL) to identify the location of the exchange format data file. The address will link to a static file or a dynamic script. This is entered using the online metadata forms under the research area category.

6) Harvest data. (Data is in the exchange format and located at or generated from one of the harvest URLs.) A web page providing a mechanism for self-harvest is provided. Please resolve any error or warning messages that are reported, and then re-harvest. The ClimDB/HydroDB administrator can be contacted if there are unsolvable problems.

For sites wishing to add their USGS maintained stations, they need to provide:

1) USGS station number and name (see attached file for a complete listing of USGS stations)

2) Provide a station code (10 characters or less - you can use the USGS number or not)

3) Provide a watershed name and watershed code (can be the same as station or not) for streamflow sites

4) Provide a list of measured parameters at this USGS site (or we will screen for any valid HydroDB parameters, e.g., precipitation, stream temperature, etc.)

5) Adjust the QC min-max ranges in the metadata web pages to prevent harvest failures due to excessive WARNING(101) warnings (section 7.3).

Note: General quality assurance criteria (min-max ranges) for all stations by variables can be entered in the metadata. Before set up in the automated system, the station's data are pre-screened with broader upper limits on gage height, discharge, precipitation, air temperature, etc. if provided. This is a mechanism for eliminating bad values that might cause ClimDB/HydroDB harvest to fail.

Visit the USGS NWISWeb data page (http://waterdata.usgs.gov/nwis/rt) to see USGS maintained stations with Real-time data.

1.6. Implementation

1.6.1. Do-it-yourself harvest

ClimDB/HydroDB allows participating sites to trigger a harvest of their site's data from the central site webpage. The newest implementation is allowing sites to control their data harvest URL from the online metadata forms and two options per site are allowed. Therefore, sites will need to specify which harvest URL they would like to use. Additionally, the site will be able to wait and see any error or warning messages appear directly onto the screen. The success or failure of the harvest will be known immediately, and data files can be harvested in an iterative process until all changes or corrections can be made. The error log will be posted to the screen at the conclusion of the entire process, which might take several minutes. Additionally, the error log file is automatically emailed to the site's data set contact person and the ClimDB/HydroDB database administrator.

Another change in this implementation is the preservation of previously harvested data. If data has been previously harvested, it does not have to be harvested again (although it is ok to do so). However, if changes need to be made, simply re-harvest an edited exchange file containing corrected data or both new and corrected data.

The harvester mechanics are divided into 3 phases: harvest, ingestion, and population.

1. In the harvest portion, the central harvester checks to make sure the URL address is valid and active, and then captures the exchange file from this URL address. An error message will be logged if the harvest fails.
2. The ingestion phase does all of the data screening for errors and generates warning and error messages to the log file. Header line and data set compatibility and consistency, as well as quality assurance checking are done here. This process is generally fairly quick unless there are massive numbers (10,000's) of data records.
3. The population phase takes a few minutes. Here, the data is placed into the relational database. Even if you abandon the web page during this process, the process will complete itself, and the resultant error log file will be posted to the screen as well as emailed to the data set contact and the ClimDB/HydroDB database administrator. The log file should be checked to make sure the data was successfully harvested. The data will be instantly available on the download portion of the webpage for users to check.

1.6.2. Do-it-yourself metadata

Site climatic and hydrologic metadata descriptors are entered using a password protected web entry form. Metadata can be entered in piecemeal fashion and edited again at a later time. Metadata is separated into various categories by their descriptors. (See Section 4.0. and associated webpages for more on metadata.)

1.6.3. Measurement Parameters

The valid implementation variables follow. Please refer to section 3.1 for the valid variable names to be used in the exchange format.

1. Air temperature; daily minimum, maximum, and mean in degrees Celsius (°C)
2. Atmospheric pressure; daily mean in hectopascals (hPa)
3. Dewpoint temperature; daily mean in degrees Celsius (°C)
4. Global solar radiation; daily total in MegaJoules per square meter (MJm^{-2})
5. Precipitation; daily total in millimeters (mm)
6. Relative humidity; daily mean in percent (%)
7. Snow depth (water equivalence); daily instantaneous observation in millimeters (mm of water).
8. Soil Moisture; daily mean in megapascals (MPa)
9. Soil temperature; daily mean in degrees Celsius (°C)
10. Stream Discharge; daily mean in liters per second (l/sec)
11. Vapor pressure; daily mean in hectopascals (hPa)
12. Water Temperature; daily minimum, maximum, and mean in degrees Celsius (°C)
13. Wind direction and resultant wind direction; daily mean in degrees azimuths (deg)
14. Wind speed and resultant wind speed; daily mean in meters per second (m/sec)

1.7. Disclaimer and Caveats

While every effort will be made to assure the integrity of the ClimDB/HydroDB central database, complete accuracy cannot be guaranteed. Users of ClimDB/HydroDB will take responsibility for subsequent use of any data retrieved. Data providers understand that ClimDB/HydroDB datasets are public.

1.8. Web Services

Experimental web services exist for accessing data in the ClimDB/HydroDB database. A useful web service for building the exchange file is called last_harvest, which returns the date of the last harvest for a given site, station, and ClimDB long or short variable name. For more information see Web Services for ClimDB/HydroDB Database (Fall 2011).

2.0. Exchange Format

2.1. Exchange Format Specification

The exchange file is fundamental to the operation of ClimDB/HydroDB. The following are some basic guidelines for the exchange file:

1. The exchange file is comma-delimited ASCII.
2. The exchange file must be made internet-accessible at a specific publicly accessible URL at the local site.
3. The exchange file can be static (created ahead of time) or dynamically created on the fly by a web script.
4. The exchange file contains a header line that describes the sequence of variables and their associated flag name contained in the data. The "!" (bang or exclamation point) character must initiate the header line and is reserved for this use only.
5. The header line is followed by the comma-delimited data.
6. A data quality flag directly follows each variable.

Note: Current valid variable names are listed in Section 3.1 along with their data limits. The data quality flag uses the same variable name preceded by the word "Flag_"

Here is an example header line for air temperature and precipitation data. Note, the header line could be one long continuous line, but this example uses continuation characters (further described in section 2.3.1):

```
!LTER_Site, Station, Date, Daily_AirTemp_Mean_C, Flag_Daily_AirTemp_Mean_C, \
#Daily_AirTemp_AbsMax_C, Flag_Daily_AirTemp_AbsMax_C, Daily_AirTemp_AbsMin_C, \
#Flag_Daily_AirTemp_AbsMin_C, Daily_Precip_Total_mm, Flag_Daily_Precip_Total_mm
```

Examples of variable names are defined as follows:

LTER_Site	A three-letter LTER/Research Area site code assigned by ClimDB/HydroDB database administrator
Station	Local site name for the weather station or gauging station (10 character max)
Date	An 8 character field, yyyymmdd
Daily_AirTemp_Mean_C	Mean daily air temperature
Flag_Daily_AirTemp_Mean_C	Data quality flag for mean daily air temperature.
Daily_AirTemp_AbsMax_C	Daily absolute maximum air temperature.
Flag_Daily_AirTemp_AbsMax_C	Data quality flag for daily absolute maximum air temperature
Daily_AirTemp_AbsMin_C	Daily absolute minimum air temperature.
Flag_Daily_AirTemp_AbsMin_C	Data quality flag for daily absolute minimum air temperature

Daily_Precip_Total_mm	Daily total precipitation
Flag_Daily_Precip_Total_mm	Data quality flag for daily total precipitation
Daily_Discharge_Mean_Lps	Mean daily discharge
Flag_Daily_Discharge_Mean_Lps	Data quality flag for mean daily discharge

2.2. Data Quality Flags

Here is the list of valid codes for data quality flags:

G or blank	Value is a good value (blank is preferred)
E	Value is estimated
Q	Value is questionable
M	Value is missing (in this case, it is preferred to leave value field null or blank with the data quality flag = “M”. It will be allowed to assign the value of “9999” to the data field with the data quality flag = “M”, but not preferred.)
T	Trace value (For precipitation only. Values must be assigned to the data field (e.g., assign a zero or 0.1). DO NOT leave the data field null or blank.

2.3. Detailed Notes and Examples of the Exchange Format

Here is a precise example of the daily exchange format including the header line from the Andrews Forest (AND) site’s Primary Meteorological Station (PRIMET). Note: (1) The data has been aligned for readability, but fill spaces are not necessary; (2) the header line could be one long continuous line, but this example uses continuation characters (described below in section 2.3.1).

```
!LTER_Site, Station, Date, Daily_AirTemp_Mean_C, Flag_Daily_AirTemp_Mean_C, \
        #Daily_AirTemp_AbsMax_C, Flag_Daily_AirTemp_AbsMax_C, \
        #Daily_AirTemp_AbsMin_C,Flag_Daily_AirTemp_AbsMin_C,  \
        #Daily_Precip_Total_mm, Flag_Daily_Precip_Total_mm
AND,PRIMET,19960101,6.8, ,10.8,Q,4.5, , 0.0,T
AND,PRIMET,19960102,5.3, ,10.6,Q,0.8, , 4.3,
AND,PRIMET,19960103,7.7, , 9.7, ,4.1, ,20.6,
AND,PRIMET,19960104,4.2, , 6.7, ,2.4, ,11.4,
AND,PRIMET,19960105,4.8,E, 7.4,E,2.7,E,    ,M
AND,PRIMET,19960106,5.7,E, 9.7,E,1.3,E,    ,M
```

One comma-delimited header line is followed by an indefinite number of comma-delimited data records (lines). ClimDB/HydroDB is coded so that a data record (line) value is based on the immediately preceding header line. Here is a more generic example of the exchange format file.

```
!Lter_site, station, date, field1, flag_field1, field2, flag_field2,\
#field3, flag_field3, field4, flag_field4
ABC,MY_STATION,19970228,111.1,,222.22,E,333.3,,444.4,
ABC,MY_STATION,19970304,,,,,,,34,Q
```

(Note: the next line will cause an ERROR(101) to be logged and this one data record will be ignored because the header and data do not match.)

```
ABC,MY_STATION,19970305,27,E
```

In this example, field names are fictional to demonstrate the generality of formats. In practice the field names would be known names such as daily_airtemp_mean_c in place of field 1. See section 3.1 for the current valid variable names.

Also in this example, the value 222.22 corresponds to the variable name "field2" for 19970228 due to the number of commas that precede 222.22 on that line. According to the format, there must be 5 commas before field2.

2.3.1. Exchange Format Header Line

The generic example above has a format header line, denoted by the reserved character "!" (bang or exclamation point) followed by data lines. This header specifies 11 comma-separated fields.

Note: you can continue lines (any lines, header and/or data) if you end the previous line with a '\' and then begin the next line with '#'.

Multiple header lines can appear within an exchange file. That is, if the data variables in the data set change (e.g., different variables included, a change in the order of variables, variables added or removed, station changes, etc.), a new header line can be inserted followed by the corresponding data set. If only the station name changes with the variable list remaining the same, a new header line is not necessary. This will produce a WARNING(107) and the data will be successfully harvested. However for better interpretation of the log file, multiple headers should be included with a new header line for each station.

Note: no other data delimiter may be used other than a comma. **All variable names that appear in a format header are pre-assigned names.** It is assumed that no variable names shall ever be devised that are not restricted to A-Z, 0 - 9, and underscore. (No non-standard characters such as %, /, etc. will be accepted.) However for convenience, case sensitivity, underscores, and spaces will be ignored in evaluation of the variable names. (Thus daily_airtemp_mean_c could be represented as DailyAirTempMeanC if desired and still be recognized).

2.3.2. Missing Data

It is recommended that small gaps in the record be filled in with records and 'M'issing flags as appropriate. However, it is not necessary to pad the fields and flags with M where data is missing. If all the data fields specified in the format are missing for a date, the record does not have to appear at all, as in the gap between Feb 28 and March 4 in the example. Large gaps should be noted in the metadata comment field. If only some data fields are missing, the specified data fields must appear with the appropriate number of preceding commas but missing values can be blank or null. In line 2 of the example, the program can tell that field1, field2, and field3 are missing, but field4 is present with a value of 34 and a flag of Q.

Note that 9999 may be supplied as a placeholder for a missing value, even if the variable is not numeric. (We may have non-numeric data variables in the future.)

The following representations are equivalent and acceptable:

```
ABC,MY_STATION,19970304,,M,,M,,M,34,Q          (preferred method)
or
ABC,MY_STATION,19970304,9999,M,9999,M,9999,M,34,Q
or
ABC,MY_STATION,19970304,,,,,,34,Q
```

2.3.3. Exchange Data Format Rules, Errors and Warnings (Also see Appendix A.)

1. It MUST be the case that a flag field will immediately follow its respective data field. Flag fields are not optional and MUST be specified in the header. Failure to follow each variable with a flag variable will cause a fatal error and complete rejection of the file.
2. Improper or unrecognizable variable names in the header will cause an error to be logged, and those data fields will be ignored throughout the ingestion process.
3. The number of fields present in the header must EXACTLY match the fields present in the data set. Records failing to conform will be ignored and an error logged.
4. In the data following any given header line, LTER_site, station, and date constitute a "unique key" (e.g., no duplicate dates). Duplicate violations will cause a fatal error.
5. ClimDB interprets data based on the format header immediately preceding the comma-delimited data set. In all cases the "primary key" fields, in this case, LTER_Site, station, and date MUST appear. For the rest of the fields, the data and flags, it uses the number of commas to judge where every data value belongs. Where there are fewer commas than "there should be", all data items on this record are ignored.
6. Html code is accepted but not preferred. Html tags will be removed and a warning message logged. Lines that contain both the "<" or ">" character anywhere in the line will be discarded (assumed to be non-data html code). Files where html is included should be sure that the html does not share any lines in common with actual data.
7. Blank lines will be discarded.
8. ClimDB assumes that data values will have a specific value if known. Thus if precipitation is known to be 0, it must be given as 0 and not null. Null or blank values are assumed to be missing.
9. Data values are assumed to have a valid number and such things as "<80", ">42", or "89-95" will result in an error being logged and that record ignored.

2.4. Data Aggregation Rules

Values flagged with “Q” or “M” will not be included in monthly or yearly aggregation. Values tagged with “E” will be included. The number of valid values used in the aggregation will be displayed. Sites are encouraged to estimate data values rather than reporting questionable or missing data.

If all data values (e.g., data values listed in the header line) are all missing for a period of days, it is not necessary to “fill in” these periods with null data and “missing” flags.

Each field in the data is parsed and has its leading and trailing spaces removed before inspection. Then in this order these operations occur:

- If a data value of 9999 is encountered, its flag will be forced to M.
- If an invalid flag code is encountered, an error message will be logged and the record ignored.
- If a data value of NULL (nothing) is encountered, the flag will be forced to M
- If a flag value is G, the flag will be forced to NULL.
- If a flag value is M, the data value will be forced to NULL,
- In the case of precipitation, if the flag is T but the data value is NULL (e.g., blank), the flag will be forced to M and a warning message will be logged.

2.5. Guidelines for Units of Measurement and Precision for Each Variable

The units of measurement are listed along with the variable naming conventions in section 4.0. Alternative units are not allowed. For example, mean, maximum, and minimum temperatures will be reported in degrees Celsius and precipitation will be reported as millimeters. Values should be reported only to the number of significant figures. The central site reserves the right to report summary values with altered decimal placement.

Some useful conversions:

1 hectopascal (hPa) = 1 millibars (mbars)
1 bar = 0.1 megapascal (MPa) = 100 kilopascal (kPa)
1 Langley = 4.187 x 10^{-2} megajoules per square meter (MJm^{-2})
1 megajoules per square meter (MJm^{-2}) = 10^{-2} Joules per square centimeter (Jcm^{-2})
1 cubic foot per second (cfs) = 28.31685 liters per second (lps)

3.0. Quality Assurance and Control

This section details guidelines and implementation procedures for data quality assurance within ClimDB/HydroDB. While every attempt has been made to assure data integrity, complete accuracy cannot be assumed. Data users should be made aware of quality procedures and potential errors in the database (See section 1.7).

Note that quality assurance checking is intended to provide added assurance of data integrity; however, **primary responsibility for data quality assurance rests with the individual sites.** Section 7.0 lists the errors and warnings for the types of checks within the harvester.

3.1. Guidelines for General Network QA

Data quality assurance checks will be carried out at a general network level at the time of harvest. These QA checks provide a consistency check on the data (that is, the data in those fields are what was intended), provide very general checking for outliers, provide logical consistency checks of measurements when possible (e.g., min<mean<max), and check for errors in data transmission. Parameters will be checked against threshold limits for each day

3.2. Guidelines for Parameter-Specific Range Checking

The following notes from the Climate Standards (CLIMSTAN) meeting (Greenland et al. 1997) provide simple guidelines in the determination of validity ranges for climatological data at each site. The exact form of these ranges will depend on both the measurement variable and the temporal aggregation in question. Although the process should be guided by climatological data, it is recommended that sites incorporate expert knowledge into the development of thresholds. For example, rather than simply designating record high or low values as the thresholds, some allowance should be made for the possibility of a valid record-setting measurement.

Air Temperature (maxtemp, mintemp, meantemp) - Error thresholds should be based on extreme values modified by expert knowledge of conditions. For example, a site might determine that a valid range for mean daily temperature in January is -5 to 15 C.

Precipitation - Thresholds should be determined from modified monthly extremes.

Relative Humidity - The minimum threshold should be based on monthly climatological values modified by expert knowledge. Use of zero as a minimum threshold should be avoided. Sites should take into account the type of instrument used when setting maximum threshold values. Hygrometers based on electrolytic resistance sensing elements (such as Vaisala or Phys-Chem humidity probes) are generally unreliable at humidities exceeding 99%.

Global Radiation - Thresholds based on modified monthly extremes. Maximum values should show a fairly uniform annual progression for these data, but minimum values are likely to be quite variable due to the effects of clouds.

Mean Wind Speed - High observed variability makes setting thresholds for wind speed more difficult than for other variables. Sites should consider use of annual, rather than monthly data in setting maximum and minimum thresholds.

Vector Mean Wind Direction - Also a highly variable measurement. Thresholds should consist of a range of vector directions typical of the site. Sites should consider using longer time periods for deriving threshold values.

3.3. Parameter-Specific Default QC Threshold Values

Default QC threshold values for the ClimDB/HydroDB measurement parameters:

Variable name	Variable Code	Low warning threshold	High warning threshold
daily_atmpressure_mean_hpa	ATM	960	1050
daily_dewpoint_mean_c	DEW	-50	50
daily_discharge_mean_lps	DSCH	0	20000
daily_globalrad_total_mjm2	GRAD	0	40
daily_precip_total_mm	PREC	0	150
daily_rh_mean_pct	RH	0	100
daily_reswinddir_mean_deg	RWDI	0	360
daily_reswindsp_mean_msec	RWSP	0	50
daily_soilmoisture_mean_mpa	SM	0	0.3
daily_soiltemp_absmax_c	SMAX	-5	25
daily_soiltemp_mean_c	SMEA	-5	25
daily_soiltemp_absmin_c	SMIN	-5	25
daily_snowh20_instant_mm	SNOW	0	1200
daily_airtemp_absmax_c	TMAX	-50	50
daily_airtemp_mean_c	TMEA	-50	50
daily_airtemp_absmin_c	TMIN	-50	50
daily_vappressure_mean_hpa	VAP	0	100
daily_winddir_mean_deg	WDIR	0	360
daily_watertemp_absmax_c	WMAX	-10	40
daily_watertemp_mean_c	WMEA	-10	40
daily_watertemp_absmin_c	WMIN	-10	40
daily_windsp_mean_msec	WSP	0	50

The general set of limit checks is intended to provide a quick warning if an instrument or data collection system has completely failed. The threshold-based range tests are intended to catch more subtle measurement errors resulting from instrument damage, miscalibration, or error in data retrieval.

3.4. Changing Threshold Values for QC Checks

Threshold values for each variable are stored in the metadata for that measurement parameter. Sites should adjust the QC threshold values descriptors (e.g., qc_min, qc_max) for each measurement parameter and these values will be used to check all data uniformly on a daily basis. If no value is provided, then the default values, outlined above, are used.

3.5. Implementation of QA/QC Guidelines

3.5.1. General Harvest

The harvester reports three types of errors and warnings (see section 7.0 for full description). If a ***Fatal Error*** is encountered, the program halts and the data is not accepted. If an ***Error*** is encountered, the program continues and just the particular data point or record that produced the error is not accepted. When ***Warnings*** are logged the program also continues and the data point or record that produced the error may be accepted or ignored, depending on the warning.

Climate observers at the site should evaluate warnings. If the data are determined to be correct (despite exceeding threshold values), no action is required. If a measurement is confirmed to be bad, a missing value indicator may be entered, or the data may be estimated using some proxy value. In either case, the corresponding flag should be inserted into the database, and the data re-harvested. At the discretion of site personnel, a datum of unknown quality may be included in the database and harvested by the network. Such data should be marked with a 'Q' flag, indicating questionable data. Data with the 'Q' flag will be excluded from monthly or yearly aggregates at the network level; however 'Q' data will be available in daily files.

3.5.2. Parameter-Specific Guidelines

General default threshold quality control limits are set in the database (section 3.3). However, station parameter-specific checking is also available. Data limits for each parameter at each station can be specified in the metadata. If provided, the harvest program checks all data versus these qc_min and qc_max metadata descriptors for each station by parameter. Warnings are logged and the log file emailed to the data set contact at each site.

When mean, maximum, and minimum exist for a parameter, the following relationship must hold to prevent a warning (note: this is only implemented for air, soil, and water temperature.):

daily_airtemp_absmin_c ≤ daily_airtemp_mean_c ≤ daily_airtemp_absmax_c

Failure of any data limits test will result in a warning message to be logged, but the data will be retained.

3.5.3. QA Warnings Using Data Quality Flags

The following codes and flags will be used by the sites to notify users of potential errors in the data. See Section 2.2 for additional explanation of codes.

Code	Code description
9999	Missing data (enter in data field, this code is optional)
M	Missing data (enter in flag field)
E	Estimated data
Q	Questionable data

4.0. Metadata Database

Web forms are available for individual sites to provide metadata. Metadata descriptors (elements) have been grouped into categories for ease of entry and implementation on a web form. To view the metadata descriptors for each category see http://wwwdata.forestry.oregonstate.edu/climhy/variable_desc.pl. The metadata provides a place for users to obtain information related to the data or how it was collected. The fields can be filled in with text or a URL provided to link to further information. Also, it is within the Research Area Information category where sites input/edit their harvest URL.

4.1. Metadata Categories

1) Research Area Information
2) Watershed Spatial Characteristics
3) Watershed Ecological Characteristics
4) Watershed Descriptions
5) Hydrologic Gauging Station
6) Meteorological Station
7) Measurement Parameters
    a) Air Temperature
    b) Atmospheric Pressure
    c) Dewpoint Temperature
    d) Global Radiation
    e) Precipitation
    f) Relative Humidity
    g) Snow Depth
    h) Soil Moisture
    i) Soil Temperature
    j) Stream Discharge
    k) Water Temperature
    l) Water Vapor Pressure
    m) Wind Direction and Resultant Wind Direction
    n) Wind Speed and Resultant Wind Speed

5.0. Variable Naming Conventions

Section 2.1 displays the currently accepted variables. However, the possibility to add more variables in the future is available. At this time, only daily values are accepted.

The naming convention for variable names in ClimDB/HydroDB specifies a four-part name:

timeresolution_parameter_aggregationmode_units

For example, daily_airtemp_absmin_C is the name for the daily absolute minimum air temperature.

Time resolution refers to the integration period of the measurement. Currently ClimDB/HydroDB only deals with daily time resolution.

Parameter values and associated measurement units include:

Parameter	Parameter Code	Units	Units Code
Air Temperature	airtemp	Degrees Celsius (°C)	c
Atmospheric Pressure	atmpressure	Hectopascals (hPa)	hpa
Dew point Temperature	dewpoint	Degrees Celsius (°C)	c
Global Radiation	globalrad	Megajoules per square meter (MJm^{-2})	mjm2
Precipitation	precip	Millimeters (mm)	mm
Relative Humidity	rh	Percent (%)	pct
Resultant Wind Direction	reswinddir	Degrees Azimuth	deg
Resultant Wind Speed	reswindsp	Meters per second (m/sec)	msec
Snow Depth (water equivalence)	snowh2o	Millimeters (mm)	mm
Soil Moisture	sm	Megapascal (MPa)	mpa
Soil Temperature	soiltemp	Degrees Celsius (°C)	c
Stream Discharge	discharge	Liters per second (l/sec)	lps
Vapor Pressure	vappressure	Hectopascals (hPa)	hpa
Water Temperature	watertemp	Degrees Celsius (°C)	c
Wind Direction	winddir	Degrees Azimuth	deg
Wind Speed	windsp	Meters per second (m/sec)	msec

The aggregation mode includes the following codes:

Aggregation Mode	Code
Mean	mean
Absolute Minimum	absmin
Absolute Maximum	absmax
Total	total
Instantaneous Observation	instant

6.0. Literature Cited

Baker, Karen S.; Benson, Barbara J.; Henshaw, Don L.; Blodgett, Darrell; Porter, John H.; Stafford, Susan G. 2000. Evolution of a multisite network information system: the LTER information management paradigm. BioScience. 50(11): 963-978.

Bledsoe, C., J. Hastings, and R. Nottrott. 1996. Xclimate workshop, Davis, California, USA [Online]. Available: http://www.lternet.edu/documents/reports/Xroots/aclim.htm [1997, September 18].

Greenland, D., T. Kittel, B. P. Hayden and D. S. Schimel. 1997. A climatic analysis of Long-Term Ecological Research sites [Online]. Available: http://lternet.edu/documents/Publications/climdes/index.html [1997, September 18].

Greenland, D. 1986. Standardized meteorological measurements for Long-Term Ecological Research sites. Bulletin of the Ecological Society of America. 67:275-277. http://www.lternet.edu/community/committees/climate/standard86.html

Greenland et al. 1997. CLIMSTAN: Standards for Observation and Archiving of LTER Climate Data. Albuquerque, New Mexico, USA [Online]. Available: http://lternet.edu/community/committees/climate/climstan/standards97.html

Henshaw, Donald L.; Sheldon, Wade M.; Remillard, Suzanne M.; Kotwica, Kyle. 2006. CLIMDB/HYDRODB: a web harvester and data warehouse approach to building a cross-site climate and hydrology database. In: Proceedings of the 7th International Conference on Hydroscience and Engineering (ICHE-2006); Philadelphia, PA. Philadelphia, PA: Drexel University, College of Engineering: [Not paged]. [Online]. Available: http://hdl.handle.net/1860/1434 [Available through iDEA: Drexel E-repository and Archives].

Henshaw, D. L., M. Stubbs, B. J. Benson, K. Baker, D. Blodgett, J. H. Porter. 1997. Climate database project: a strategy for improving information access across research sites. In Proceedings of the Data and Information Management in the Ecological Sciences Workshop. Albuquerque, New Mexico, USA [Online]. Available: http://www.fsl.orst.edu/lter/pubs/spclrpts/climdbnm.htm [1997, October 20]

6.1. References

All references available from the ClimDB/HydroDB webpage or archived here: http://intranet2.lternet.edu/documents/scientific-reports/climate-and-hydrology-database-projects

Web Services for ClimDB/HydroDB Database (Fall 2011) (Web Page)

ClimDB/HydroDB (ClimHy) Database Migration to LNO (Fall 2009) (Web Page)

ClimDB/HydroDB (ClimHy) Database Update (Spring 2005) (Web Page)

CLIMDB/HYDRODB Progress Report (17 Mar 2003) (Word Document)

Intersite Hydrological Database (HYDRODB) presentation (March 2002) (slideshow; not downloadable)

HYDRODB Progress Report (17 Oct 2001) (Word Document)

CLIMDB Progress Report (16 Oct 2001) (Word Document)

7.0. Appendix: Errors, Warnings, and Fatal Errors

The following lists errors and warnings generated by harvesting routines:
In this documentation square brackets and their contents, [], will be replaced by the value described in the brackets.

> Flag character [flag] not recognized"
> is printed as
> "Flag character X not recognized"

If applicable, all errors and warning messages are followed by the site, station, and date of the record in question and the file name and line in that file where the warning/error was raised.

7.1. Fatal Error Messages - program halts, data is not accepted.

FATAL ERROR(900):
Message: Fails during attempt to download from [site]
Description: No connection is made to data URL: file cannot be downloaded.
NOTE: This error message is followed by the errors returned from the remote server, or a message stating that the URL is not valid.

FATAL ERROR(901):
Message: [variable] needs to be followed by [flag_variable]
Description: All variables require that a flag_variable directly follow.

FATAL ERROR(902):
Message: Stopped logging errors after [errors] errors
Description: The number of errors exceeds the established threshold.
NOTE: The threshold is currently set to 10 errors

FATAL ERROR(903):
Message: Unknown site code encountered in the data set
Description: Site code is not listed in the central database.
(Contact ClimDB administrator)

FATAL ERROR(904):
Message: Unknown station code encountered in the data set
Description: Station name is not listed in the central database.
(Contact ClimDB administrator)

FATAL ERROR(905):
Message: Continuation line not continued.
Description: Continuation lines end with '\' and next line, following a new line, must start with '#'.

FATAL ERROR(906):
Message: Duplicate found.
Description: Duplicate record by site, station, parameter, and date.

FATAL ERROR(907):
Message: More than [warnings] warnings encountered; Process is aborted.
Description: The number of warnings exceeds the established threshold.
NOTE: The threshold is currently set to 50 warnings.

7.2. Error Messages - Program continues, data point or record is not harvested.

ERROR(001):
Message: Number of data fields[number_of_data] != number of header fields[number_of_headers]
Description: Number of data fields does not match number listed in header. Data record ignored.

ERROR(002):
Message: Flag character [flag] not recognized
Description: Illegal flag. Data point is ignored.

ERROR(003):
Message: [data] is not valid (must be numeric)
Description: Illegal character(s). Data point ignored.

ERROR(004):
Message: Time stamp [date] is in the future
Message: Month is [month]
Message: Day is [day]
Description: Invalid date. Data record ignored.

ERROR(005):
Message: Illegal number of data fields
Description: Commas do not add up properly. Data record ignored.
(See error (001) and fatal error (901))

ERROR(006):
Message: No Header in file
Description: No header present in the file. Program will use the default header line and continue.
(Currently only a warning, see warning (102))

7.3. Warning Messages - Program continues, data points and records may be accepted or ignored.

WARNING(100):
Message: Ignoring UNKNOWN VARIABLE
Description: Variable name is not listed as valid in the central variable database. All values listed for that variable are ignored.

WARNING(101):
Message: [variable] = [value] failed QC test
Description: Data value fails general data limits check. Data is still accepted.

WARNING(102):
Message: No header was supplied. Using assumed header of form:
!SITE_CODE,STATION,DATE,DAILY_AIRTEMP_MEAN_C,FLAG_DAILY_AIRTEMP_MEAN_C,DAILY_AIRTEMP_ABSMAX_C,FLAG_DAILY_AIRTEMP_ABSMAX_C,DAILY_AIRTEMP_ABSMIN_C,FLAG_DAILY_AIRTEMP_ABSMIN_C,DAILY_PRECIP_TOTAL_MM,FLAG_DAILY_PRECIP_TOTAL_MM
Description: No header line is listed. The default header line is inserted, and the harvest continues.

WARNING(103):
Message: File contains HTML
Description: HTML code detected. Record is deleted from the harvested file.

WARNING(104):
Message: Flag = T; data = null. Flag set to 'M'
Description: Flag indicates trace value. Data point is considered missing.

WARNING(105):
Message: (Year<1900) Year is [year]
Description: Questionable year value. Data record is still accepted.

WARNING(106):
Message: Failed (min < mean < max) relationship
Description: Quality assurance failure. Data record is still accepted.

WARNING(107):
Message: Station code changed without a corresponding header change
Description: It is recommended to place another header line in the file when the station changes. Data is still accepted.