A DESCRIPTION OF THE WEATHER SOURCE COMPREHENSIVE GLOBAL WEATHER OBSERVATION DATABASE

In recent years, demand for databases of high quality weather observations has intensified. While there are many causes for this, increases in the number of climate studies and in the number of businesses employing weather risk management methods are two principal causes driving this increased demand. In response to this need, domestic and foreign governments, academic institutions and a few businesses have made efforts to improve the quality and access to such databases. This paper describes a comprehensive global weather observation database developed by Weather Source, LLC. The Weather Source database (WSDB) consists of daily, hourly and special weather observations from tens of thousands of weather stations from across the globe. The period of record for these observations in most cases covers the period from station inception to real-time if available. A series of quality control and data improvement methods are employed on the data to increase quality; these include both meteorological consistency and statistical methods which perform various functions. These functions include: identification of errors, production of high quality estimates to correct errors and fill missing values, and data homogenization (adjustments for observation discontinuities and long-term trends). Discussed herein is a description of the WSDB data, data improvement methods and data access options.

Mark J. Gibbas*

Craig Gilbert

Weather Source, LLC

1. WSDB OVERVIEW

In the simple sense, the WSDB is the fusion of computer hardware, software, algorithms and weather data. The computing platform is a PC-based Linux cluster running the MySQL 4.0 database server. Key software includes an array of data decoders and ingestors which collect, decode, perform quality control checks and ultimately import the weather data into the database.

The data sources for the WSDB are numerous. Historical data sets from NCDC include the TD3200 (Cooperative Summary of the Day), the TD3210 (First Order Summary of the Day) and the Integrated Surface Hourly datasets. In addition the NCAR DS512 (CPC Global Summary of Day/Month Observations) dataset is also used to create the historical component of the WSDB. The real-time data sources include global metar and synoptic reports as well as daily climate reports from many US weather stations. Station metadata are obtained from various sources including the NCDC Master Station History database, the Greg Thompson[1] station information file, and the NCAR DS472 and DS512 station libraries.

The data are organized into the database by station identifier[2], data source, time, and weather variable. In many cases a single station will be redundantly represent by the various data sources (e.g. TD3200, TD3210, DS512, etc). The benefit of this is that while each set of source data for a station may have errors and data gaps, a data merging process allows for the creation of a unified and more complete data set for the station in question. Data processing algorithms (primarily written in 'R') are then used for a variety of tasks such as data cleaning and homogenization and other analysis such as computing probability distributions from station time-series.

2. WSDB FIELDS

Table 1 contains all weather variables contained in the WSDB as of this writing.

[1] Greg Thompson's "stations.txt" file can be found at: http://www.rap.ucar.edu/weather/surface/stations.txt

[2] The WSDB uses a modified version of the NCDC SHIPS identifier. This identifier is associated with physical station locations, and therefore aids in methods which are used to account for station location changes

* Corresponding authors address: Mark J. Gibbas Weather Source, LLC, Amesbury, MA. Email: mark@weather-source.com

Weather Variable	Speci	Hourly	Daily
Wind Speed & Direction	Y	Y	
Gust Wind Speed	Y	Y	
Peak Wind Speed & Direction		Y	
Maximum Wind Speed & Direction			Y
Maximum Gust Wind Speed			Y
Maximum Peak Wind Speed & Direction			Y
Precipitation	Y	Y	
Hourly Accumulated Precipitation		Y	
6-Hourly Accumulated Precipitation		A	
12-Hourly Accumulated Precipitation		A	
Daily Accumulated Precipitation			Y
Daily Accumulated Snow			Y
Dew Point	Y	Y	
Temperature		Y	
6-Hourly Max/Min Temperature		A	
Daily Max/Min Temperature			Y
SLP	Y	Y	
Present Weather	Y	Y	

Table 1 WSDB Weather Variables. 'Y' denotes weather variable is represented. 'A' denotes weather variable is represented as available.

3. DATA PROCESSING ALGORITHMS

The data processing algorithms range from simple to complex and are applied as a series of steps that ensure that the data contained in the database is of highest quality possible. While not all weather variables are subject to each step, the following is a listing of the processes that most weather data are subject to.

Data Quality Tests

There are three data quality checks that each observation is subject to. In each case the test results in the setting of a flag to indicate whether the data passed or failed the test. After the tests have been completed, a test evaluation method reviews the tests to determine if a value is good, suspect or bad. If the value is deemed good, it is allowed to stay in the database. If the value is deemed suspect, a message is emailed to a meteorologist to further evaluate the validity of the data. If the value is deemed bad, it is removed from the database. The following is a brief description of the three quality control tests applied to the observations.

A. World record bounds test. Thresholds are set to know world records plus a small delta to allow for the possibility that a new world record may occur. Data outside the bounds of the thresholds are flagged as failing.

B. Internal test. Check to see if value is consistent with other observed values (e.g. dew point greater and air temperature, or max temp less than min temp). Also check to see if the value is statistically consistent with station values observed in the past for the same time of year.

C. External test. Using surrounding stations that pasted the first two tests, a set of regression estimates are produced, which are then used to test the data at the station of interest.

After the tests have been completed, an evaluation is done to determine the status of the data in question. Passing all three test means the observation was deemed to be a good value. Failing any single test marks the values as being suspect, and is queued to be evaluated by a meteorologist. Failing two or more tests, the value is deemed to be bad and is removed from the data base.

Data Filling

For many applications, users prefer to work with weather data that has no gaps in the time series. Of course data gaps occur frequently either because there was data missing in the first place, or because certain data were found to be invalid during the data testing process and were removed. To fill in the data gaps caused by missing and erroneous data, Weather Source has developed a suite of methods to produce estimates for the purpose of filling data gaps. While it is beyond the scope of this paper to discuss these in detail, the following is a brief description of the more well used methods.

Popular Data Filling Methods

A. Filling from multiple data sources. As mentioned above, the WSDB is composed of many data sources, which often times results in having multiple time-series for a single station. After the data for each of these time-series has been tested and processed to remove erroneous data, the time-series are then collated to produce a single unified time-series. This simple process often results if filling many data gaps to produce a time-series that is more complete than time-series from the original sources.

B. Regression estimates. When suitable nearby station data are available, regression estimates are produced to fill data gaps. In practice it has been found that estimates produced from the neighboring station with the highest correlation to the station being filled provide the most accurate estimates for daily temperature values. If a strongly correlated (greater that 0.95) station does not exist, estimates from the top ten highest correlating stations are use. In this case the high and low estimates are thrown out and the averaged of the remaining eight are used to produce the final estimate. While the regression method was develop for produce temperature estimates, it has also been shown to work well with a number of weather parameters including precipitation, which due to its discontinuous (in space and time) nature is the one of the most problematic parameters to estimate.

C. Diurnal Phasing. While hourly temperature can be filled using other methods, this particular method is used exclusively for the production of estimates for hourly temperature. In this method, the mean hourly temperature for a given day of the year is phased in to splice a data gap between two known observations. Figure 1 offers a graphical description of this method.

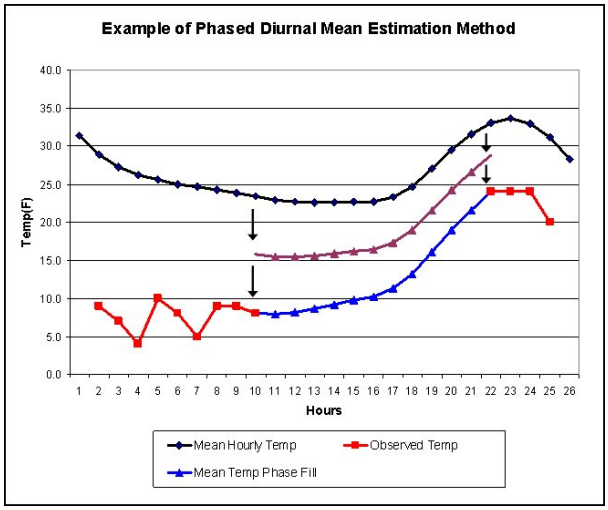

Figure 1 A graphical example of the phased diurnal estimation method. This method uses observed hourly temperature data (red square line) and mean hourly temperature (black diamond line). By knowing the mean diurnal profile of temperature for a given day and knowing the beginning and end points of a data gap, this method splices the data gap by phasing in the appropriate portion of the mean diurnal temperature into the data gap.

Data Homogenization

The subject of data homogenization will be discussed only briefly here. The authors intend to produce a more in-depth paper of this subject in the not too distant future.

The data homogenization process is composed of two parts: data discontinuity adjustments and data trend adjustments. The former is applied first, then the latter. Depending on then need of the users, the data can be provide with only the discontinuity adjustment.

Since the variance in parameters such as temperature can be larger than the amount of variance caused by a station change, statistical methods which rely only a stations own data to find discontinuities are often not effective. For this reason many individuals and teams (Nelson et al

(1979), Karl and Williams (1987), Easterling and Peterson (1995), Allen and DeGaetano (2000)) have shown that the use of a reference time series can be employed to produce effective methods to reveal discontinuities within an inhomogeneous time-series. Similar methods are used to produce the discontinuity adjustment for WSDB data. The following are some of the key features of the WSDB discontinuity identification method:

A. Uses identified station changes whenever possible, but relies on data tests to identify non-documented discontinuities.

B. Method allows for the identification discontinuity effects throughout the year. For instance a station moved closer to a body of water would be cooler in the summer and warmer in the winter, but the net averaged adjustment may be zero and may be missed by methods that only test yearly data segments.

C. Method produces a series of candidate solutions. Each solution is tested to identify the solution which best improves the data homogeneity.

4. DATA ACCESS

As of the time of this writing, access to the WSDB is conducted by request and by subscription. Primary data deliver methods are via CD/DVD and FTP. In early 2005 it is anticipated that an online web interface will be in place to allow users to query and retrieve data online. While the WSDB is a commercial service, academic research interests are supported with significant discounts.

5. CONCLUSIONS

The WSDB is a comprehensive database of high quality surface weather data suitable for many commercial and research purposes. Currently the database offers temperature, wind, precipitation and other weather observations for tens of thousands of weather stations worldwide. Future plans call for the inclusion of cloud observations and the development of an online interface to enable users to query and retrieve data online.

6. REFERENCES

Nelson, W.L., R.F. Dale, and L.A. Schall, 1979: Non-climatic trends in divisional and state mean temperatures: A case study in Indiana. *J. Appl. Meteor.*, 18, 750-760.

Karl, T.R. and C.N. Williams Jr., 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Climate Appl. Meteor.*, 26, 1744-1763.

Easterling, D.R. and T.C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.*, 15, 369-377.

Allen, R.J. and A.T. DeGaetano, 2000: A method to adjust long-term temperature extreme series for nonclimatic inhomogeneities. *J. Clim.,* 13, 3680-3695