THE GLOBE PROGRAM

Improving Scientific Value Through Quality Control of Data Collected According to the GLOBE Atmosphere Investigation Protocols

Graeme Stephens, Ian Wittmeyer and Debra Krumm

Colorado State University

Department of Atmospheric Sciences

Foothills Campus

Fort Collins, CO, 80523-1371

Telephone: (970) 491-8541 Fax: (970) 491-8449

Email: stephens@atmos.colostate.edu

P.O.Box 3000 / Boulder, CO / 80307-3000 / www.globe.gov

Sponsored by: NASA Supported by: NSF NOAA

Introduction

The Department of Atmospheric Science at Colorado State University has assumed responsibility for the quality control of data collected following the basic Atmosphere Investigation protocols. This includes detailed assessment of the data routinely collected as part of the ongoing GLOBE protocols, as well as data already archived. The establishment of a data quality control and assurance plan to carry out these assessments is underway. In addition to statistical analyses, assessments will include comparisons of the data with similar information derived from outside sources including climatological data and model-based forecast fields. Once the quality control and assurance plan is in place, the concept is to analyze the current data as frequently as possible, with a goal of once per week.

Temporal Tests and Quality Assurance

In addition to its excellent educational goals, GLOBE also maintains strong scientific goals. By providing quality control for the atmospheric data through the use of statistical tests to identify outliers and through comparison to outside databases, a high level of quality assurance can be provided to scientists. Improving the scientific value of the GLOBE atmospheric data should give scientists the confidence and encouragement to use these data in their own research.

Temporal tests will be used to check for obvious outliers in the archived GLOBE atmospheric data. This involves comparing individual values with valid data for the same location, and testing the data for consistency over prescribed time periods. Any value that falls outside the prescribed range of acceptable values will be denoted as an erroneous datum. Many minor measurement errors, however, will likely go undetected due to the low temporal frequency of data collection (just once per day) unless an independent measurement can be used to compare with. Examples of the sources of such suspect data may include transcription or typographical errors, and temperatures recorded in degrees Fahrenheit instead of Celsius.

An example of obviously erroneous data for a particular school is given in Figure 1. Two data points in the figure are obviously not consistent with the rest of the data shown. The goal of our efforts is to eliminate points such as these from the dataset. The tests to be used are based on the premise that

individual values, such as monthly-means, should be statistically 'similar' to values for the same month for other years. As such, the tests are most meaningful when applied to relatively homogeneous data like temperature. The tests thus serve as a simple and gross way of identifying obvious outliers in the measured temperatures as recorded under the GLOBE protocol.

Close Proximity Independent Data Comparisons

To some degree it is possible to use the existing GLOBE records to test for data consistency and possible outliers, but the data record is generally insufficient to carry out climatological comparisons, so the plan is to use climate data compiled by the National Climate Data Center (NCDC) for this purpose. We plan to conduct these tests on monthly mean temperature data (maximum/minimum and mean) in an effort to identify gross biases in the temperature data as well as compare daily mean values to the monthly mean climatologies. Since daily values can vary significantly about the monthly means we are looking for outliers above predetermined 'weather noise' expressed in terms of the standard deviation of daily mean, maximum and minimum temperatures as obtained from the climate records of NCDC.

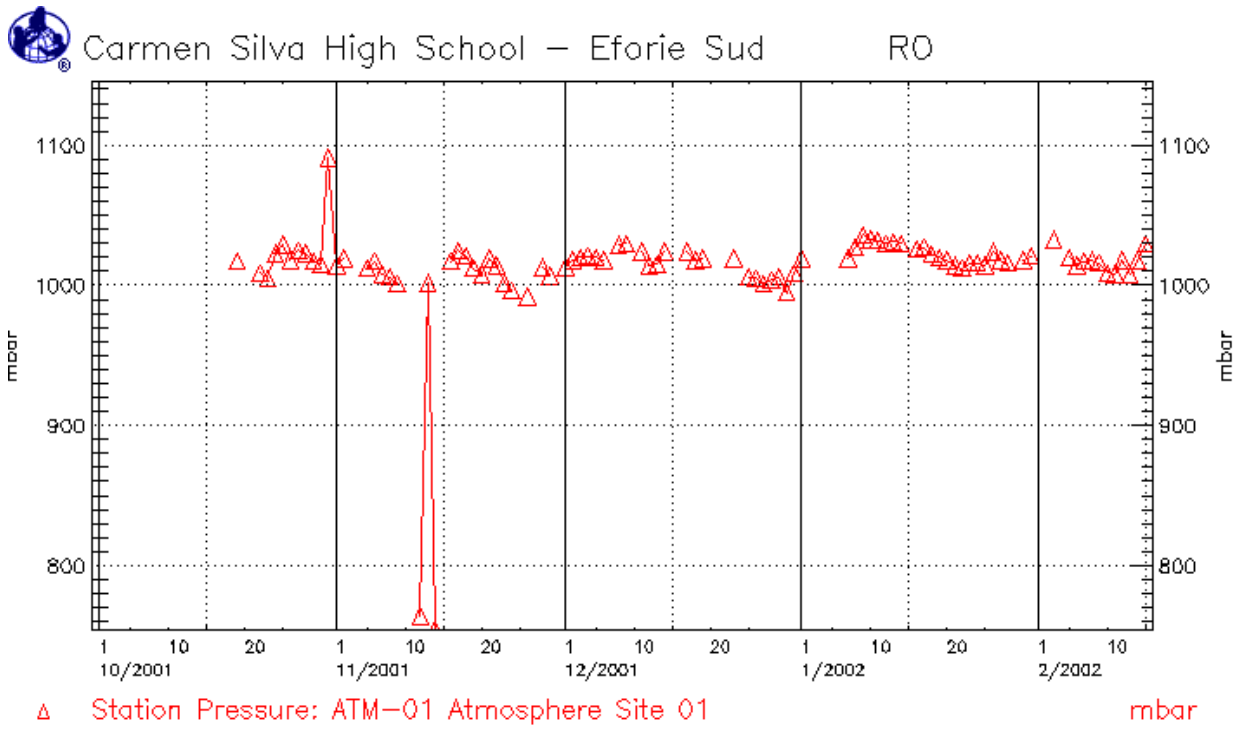

Figure 1. The time series of daily temperature from a single GLOBE school.

Figure 1 shows a time series of GLOBE temperature data from a school where a few measurements are obviously erroneous and would be filtered by such a simple screen. Daily values will also be examined internally by comparing values with previous and next day values if reported data frequency allows. These daily changes can be compared with maximum daily changes derived from the entire time series.

In some cases we can use close proximity independent data for examination of biases and obvious errors. The GLOBE site in Hays, Kansas, is near a National Weather Service surface observing station that has a long record of hourly data. Figure 2a shows a time series of daily maximum temperatures from both datasets during 2001, while Figure 2b illustrates the difference in those daily maximum temperatures. A number of days show significant differences that can be related to a host of issues, including site location inconsistencies, differing time periods for daily maximum and minimum temperatures, and integrity of the measurements. In these cases, it will be our task to try and determine the predominant reasons for these types of differences.

We intend to focus our initial efforts on data collected by schools that report most frequently, and make the results available to the schools in an effort to provide motivation for further high quality data collection.

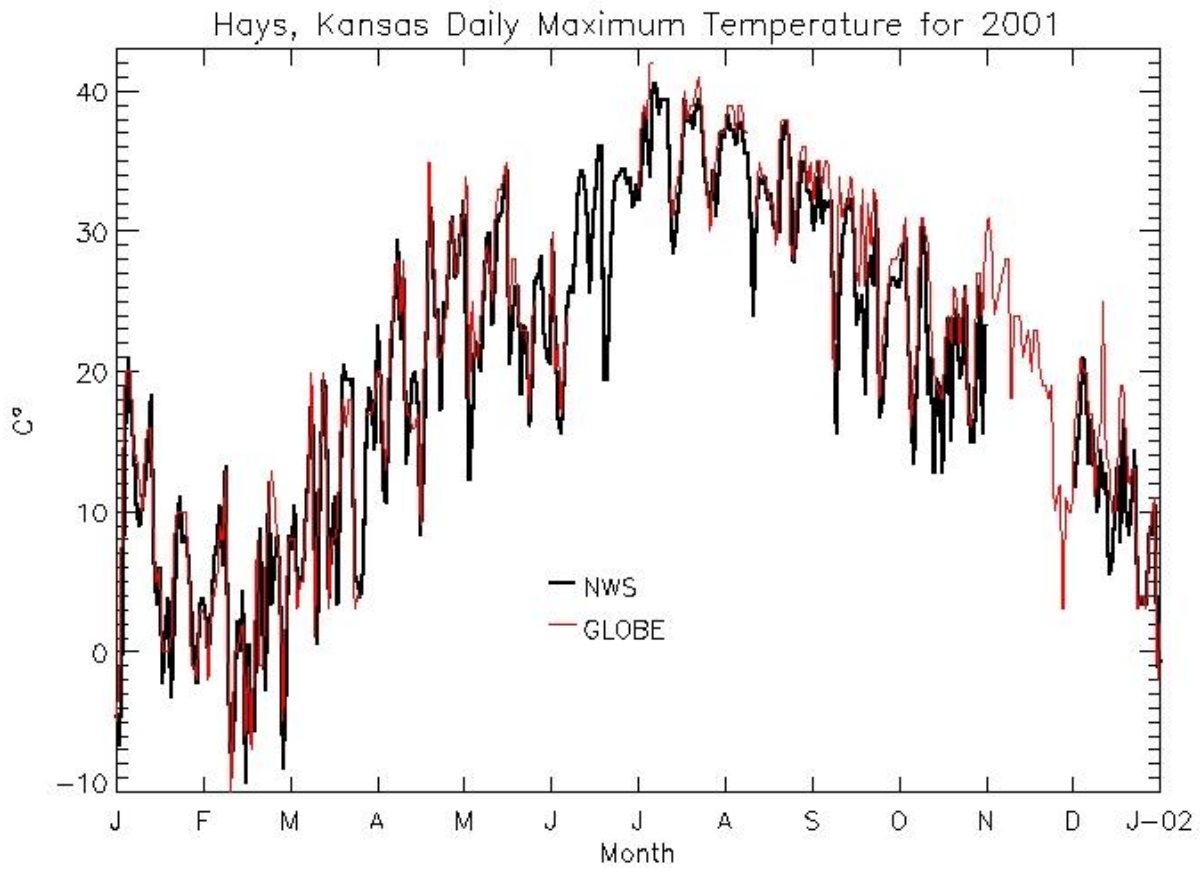

Figure 2a. Hays, KS: Daily maximum temperatures for 2001 - GLOBE data versus NWS records.

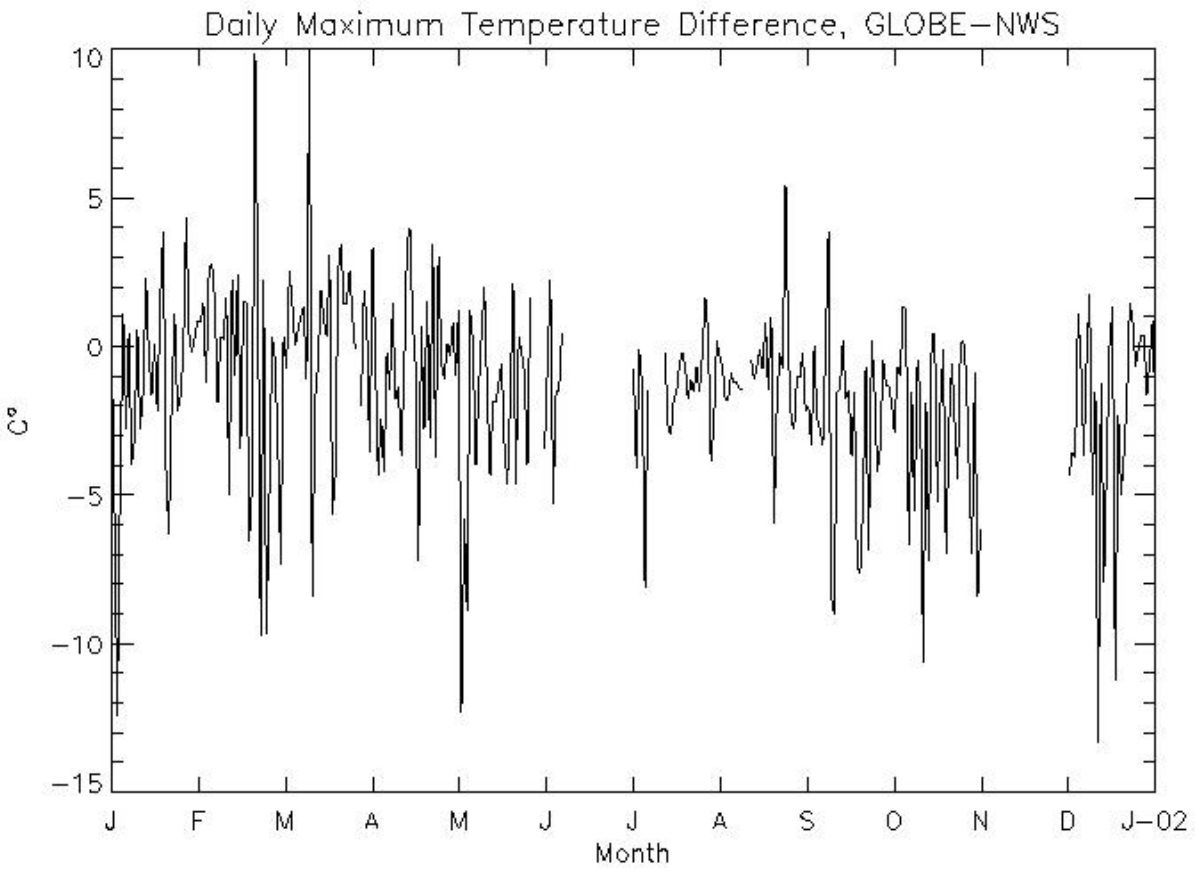

Figure 2b. Hays, KS: Daily maximum temperature differences between GLOBE data and NWS.

Since the GLOBE measurements are taken near solar noon, there is a different 24-hour period during which a maximum or minimum temperature may be recorded as compared to the NWS, whose 24-hour period ends earlier in the morning. Large temperature changes during the morning can potentially cause the datasets to have an event maximum or minimum temperature reported in a different day. As an example, Figure 2b provides a number of cases where the difference in maximum temperature was largely positive one day and largely negative the next, which is likely due to this difference in how maximum and minimum temperatures are reported.

In addition, some GLOBE sites have been relocated within that school's property. Figure 3 shows a time series of daily maximum temperatures from the Hays, Kansas GLOBE sites ATM-01 and ATM-02, both located at the high school but over a different surface type. Site ATM-01 exhibits more extreme values in a number of instances, which is likely due to the surface type over which the observations were taken. Site ATM-01 is located near the school over a rock surface, while ATM-02 is located over a grassland surface (see Figure 4). Understanding the nature of the environment around the site is important for understanding possible measurement discrepancies (e.g. Hanamean et al., 2002). Photographs of the sites and the environment around the sites will provide useful metadata for this purpose and schools will be encouraged to document their sites in this way, in the comment section already provided by GLOBE under Data Entry.

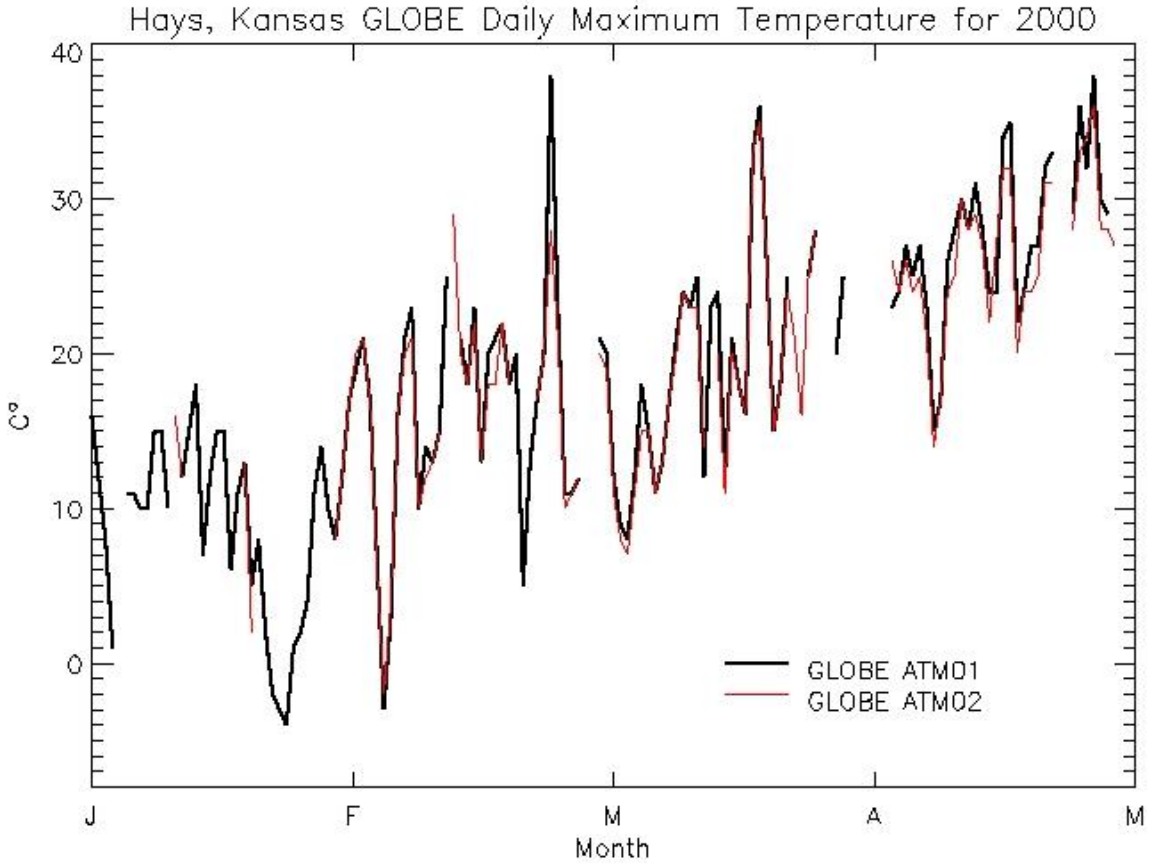

Figure 3. Hays, KS: GLOBE daily maximum temperatures for the year 2000 taken from two different locations on the school grounds.

Figure 4. Hays, KS: View of grasslands from GLOBE station ATM02.

Increasing Participation through Mentoring

The GLOBE site at Hays has been listed on the "Chief Scientist's Honor Roll" for their high volume of data collection. Many other GLOBE sites collect data at a much-reduced level and therefore are of limited value for scientific use. For practical reasons, quality control efforts will focus most heavily, although not exclusively, on the GLOBE schools providing the most continuous and highest quality data. Through a mentoring system already underway, GLOBE schools will be encouraged to increase their level of participation in data collection related to the Atmospheric Investigation protocols.

References

Hanamean, J. R., Jr., R. A. Pielke Sr., C. L. Castro, D. S. Ojima, B. C. Reed, and Z. Gao, 2002: Vegetation impacts on maximum and minimum temperatures in northeast Colorado. *Meteor. Appl*., submitted.