NOTE: This is a PDF version of the following publication:
Woodruff, S.D., 2001: COADS updates including newly digitized data and the blend with the UK Meteorological Office Marine Data Bank. *Proceedings of International Workshop on Preparation, Processing and Use of Historical Marine Meteorological Data, Tokyo, Japan, 28-29 November 2000*, Japan Meteorological Agency and the Ship & Ocean Foundation, 49-53.

Quality Control in Recent COADS Updates

Scott D. Woodruff
NOAA/OAR Climate Diagnostics Center, Boulder, CO, USA

1. Introduction

Data quality and quality control (QC) have been key considerations since initial development of the Comprehensive Ocean-Atmosphere Data Set (COADS; Slutz et al., 1985). This paper provides some background on the sources of data problems and inhomogeneities impacting surface marine data, an overview of the QC processing applied to COADS, with a focus on "trimming" (data screening), and future plans for QC improvements.

2. Marine data problems and inhomogeneities

Data errors can arise from many different sources (e.g., during observation, digitization, communication, or processing). Also, many changes have occurred since the late 18th century in observational practices and instrumentation (e.g., Diaz and Isemer, 1995; Parker et al., 1995); historical metadata to address these changes often are nonexistent or incomplete. COADS processing has not yet sought to make observational "bias" adjustments, because the amounts and methods of adjustment are difficult to define and in some cases still the subject of debate among the experts. Moreover, the data mixture contains many sampling inhomogeneities, including historical changes in ship tracks (Fig. 5 in Woodruff et al., 1987) and observing time (Fig. 4 in Woodruff et al., 1998).

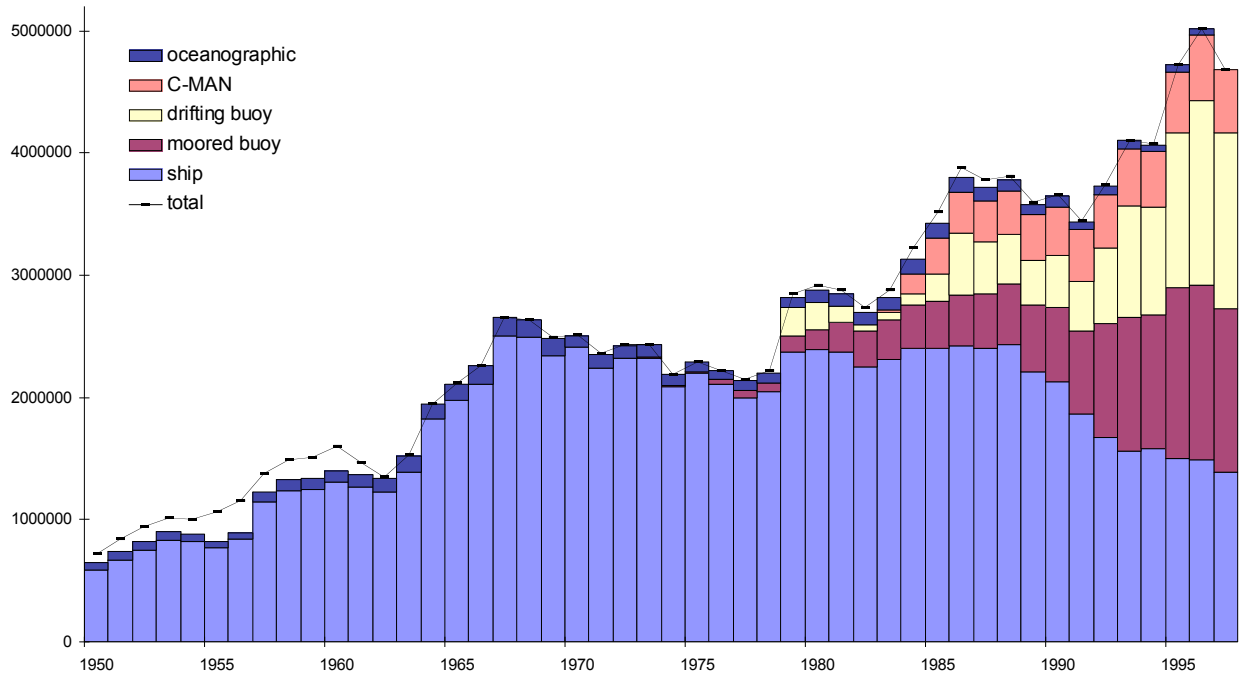

Fig. 1: Annual bars show marine reports output for 1950-97, with the numbers from different platform types plotted in descending order, from bottom to top, of total magnitude: ship, moored and drifting buoy, Coastal-Marine Automated Network (C-MAN), and oceanographic research vessel reports. The total (line) is

sometimes larger due to miscellaneous additional platform types or reports for which platform type was undetermined.

Recently, for example, ship data volume has been declining, whereas there has been substantial growth in the numbers of drifting and moored buoy reports (Figure 1). Thus data coverage has been maintained or expanded, but the data mixture is more heterogeneous.

Another important data quality consideration is that reports digitized from ships' logbooks in delayed-mode are often of better quality and completeness than those transmitted in real-time over the Global Telecommunication System (GTS). However, as shown in Figure 2, the GTS contributed a relatively stable amount, but increasing fraction, to the total ship data mixture in recent decades, and with delayed-mode data still composing over half the mixture by 1997 (these results are based on the final COADS output where delayed reports are generally selected over GTS reports when exact or approximate duplicates are detected).

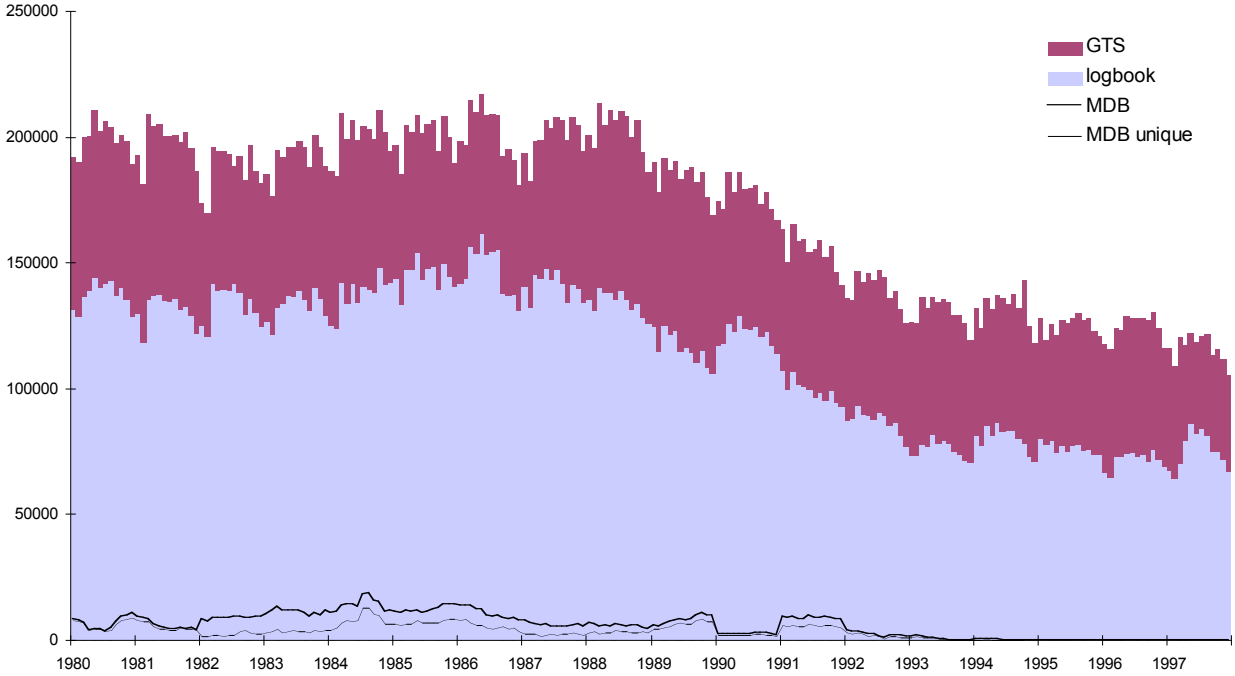

Fig. 2: Monthly bars show ship reports output for 1980-97, received via the Global Telecommunication System (GTS) or in delayed mode (generally keyed logbook data). Within the logbook category, curves show the total retention of UK Marine Data Bank (MDB) data (heavy), and within that the number of unique MDB reports (light); i.e., "total retention" includes unique MDB reports plus others that were considered preferable duplicates (in comparison to COADS, plus possibly to other MDB reports).

3. QC processing overview

The general goals of COADS QC processing are to detect and flag suspect or erroneous observational data. The flags can then be used by users to interpret the observational data, or in later COADS processing to eliminate outliers from sets of monthly summary statistics (for 2°×2° and 1°×1° latitude/longitude boxes).

The QC processing is applied in several stages. First, all data are translated into the common Long Marine Report (LMR) format used for production processing. This format has the advantage of "attachments" for flexible storage of QC and other information. The supplementary attachment is used to store original data values (e.g., prior to conversion into modern units). Any physically unrealistic data values encountered during translation are

stored in the error attachment. Working with variable-length data records containing supplementary and error attachments can be complicated, but it provides the ability to reconstruct the original data if necessary.

After the LMR records are sorted into a uniform order, "data preconditioning" is applied. This process is used subjectively to eliminate some reports with data problems, and to adjust individual weather elements or data indicators in some cases to help improve data quality and continuity. As part of this processing, for example, efforts are made to establish a platform type (ship, drifting or moored buoy, etc.) indicator for each report.

In the next stage of processing, QC flags are assigned within each LMR record. A first set of flags is assigned according to a complex "NCDC-QC" procedure developed at the NOAA National Climatic Data Center (NCDC) (Slutz et al., 1985, supp. J), originating from logic NCDC developed decades ago for FGGE. Tests are made as part of this procedure among weather elements (for internal consistency), against an older set of climatological limits using 5°×5° latitude/longitude boxes, and for landlocked data.

A second set of flags is assigned according to the "trimming" (data screening) procedure, which was developed for COADS (Slutz et al., 1985, supp. C). The primary observed variables (sea surface and air temperature, wind, pressure, and humidity) are tested against trimming limits for three climatological periods: 1854-1909, 1910-49, and 1950-79 (or using limits from the closest period for data prior to 1854 or later than 1979). The limits were calculated at ±3.5 standard deviations (σ) about the smoothed median for each period, month, and 2°×2° box (provided there were sufficient data to create limits). But to increase the usefulness of the trimming flags within each LMR record, they are set to indicate the relationship of a given observation to the trimming limits using three thresholds (±2.8σ, ±3.5σ, and ±4.5σ) about the smoothed median.

Additional QC flags are available from a few data providers, e.g., for drifting buoy data track-checked and assembled in delayed-mode by Canada's Marine Environmental Data Service (MEDS). A selection of 23 flags from all the available information is later made available to users in the Long Marine Report Fixed-length (LMRF) format.

The final stage of QC processing involves elimination (or flagging of "uncertain") duplicate marine reports detected within 1°×1° boxes (Slutz et al., 1985, supp. K). Seven weather elements are compared between reports, and the NCDC-QC results plus other criteria are used to select the "best" duplicate. For example, Figure 2 illustrates the results of applying duplicate elimination to UK Marine Data Bank (MDB) data for 1980-97.

4. Trimming problems and remedies

The trimming flags on each observational record are used for computing the monthly summary statistics for COADS. Studies of events such as the 1877-78 (Figure 3) and 1982-83 El Niños revealed that the trimming threshold used to create Release 1 statistics (rejecting observations outside ±3.5σ about the median) was too conservative, resulting in the distortion or elimination of some large climate signals (Wolter, 1997).

Concerns were also raised about the effects of mixing ship data with data from other platform types such as drifting and moored buoys (e.g., Woodruff et al., 1993). To help mitigate these problems, and so researchers can study the effects, two separate sets of 2°×2° (and 1°×1° since 1960) statistics are available for COADS Releases 1a-1b (1950-97), and will be offered for the entire archive when Release 1c (1784-1949) is completed in the near future: The "standard" statistics are derived from ship data only, using the restricted (3.5σ) Release 1 trimming limits. In contrast, the "enhanced" statistics are derived from ship and

other platform types (e.g., drifting and moored buoys), using relaxed trimming limits (4.5σ) to better preserve climate anomalies.

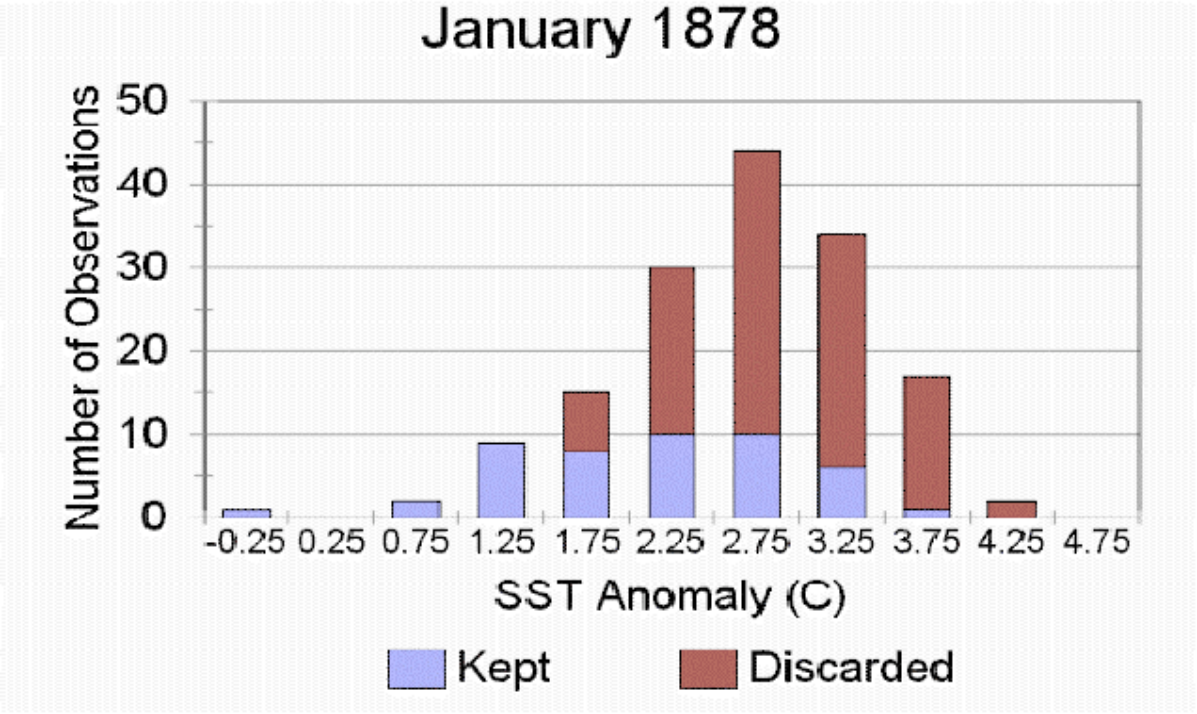

Fig. 3: Sea surface temperature (SST) anomalies and numbers of observations in the region 10°N-10°S, 80°W-180° for the El Niño in January 1878. Colors indicate observations "Discarded" and "Kept" by the trimming procedure (using limits set at 3.5σ around the smoothed median). The total number of observations is kept plus discarded. The observations are binned by 0.5°C.

5. Future plans

To better resolve the trimming problems, Wolter (1997) suggested a more "adaptive" approach, by addressing the scatter of observations around the individual year-month median (assuming sufficient data coverage) rather than around a climatological value. We have begun implementing a revised QC procedure, starting with sea surface temperature (SST), and later to be extended to other primary variables.

First, we are computing analyses using optimum interpolation (Reynolds and Smith, 1994) to determine the large-scale SST signal (e.g., for QC purposes, the SST analysis should not be strongly influenced by spatial scales less than 10° and temporal scales shorter than 6 months). Once this signal is removed, the climatological monthly σ can be recomputed. The new σ is smaller, and thus should automatically become more stringent in eliminating "bad" data, because it is not influenced by large-scale signals. The analyses may need to be performed iteratively, and additional complications may have to be considered for other variables, such as impacts on air temperature from daytime heating of the ship's structure.

Further enhancements in QC may be obtained in the future through track checking of individual ships and buoys (when a platform ID field is available). Also, we hope to provide more and better access to ship and buoy metadata (e.g., anemometer height), and move eventually toward some bias adjustments concentrated on the individual observations (see Woodruff et al., 1998). We are also beginning to offer new capabilities for dynamic subsetting (of variables, time, and space) from the basic observational and statistical products

into simple ascii formats, and in the process will offer different levels of QC that can be applied to better satisfy specific research requirements.

Future developments in QC, data enhancements, and data availability can be monitored on the COADS project Website:

http://www.cdc.noaa.gov/coads/

Acknowledgments

I am grateful for the help and advice of S. Worley, S. Lubker, R. Reynolds, T. Smith, X.-W. Quan, E. Kent, K. Wolter, and H. Diaz. COADS is the result of a continuing cooperative project between the National Oceanic and Atmospheric Administration (NOAA)—specifically its Office of Oceanic and Atmospheric Research (OAR)/Climate Diagnostics Center (CDC), its National Environmental Satellite, Data and Information Service (NESDIS)/National Climatic Data Center (NCDC), and the Cooperative Institute for Research in Environmental Sciences (CIRES, conducted jointly with the University of Colorado)—and the National Science Foundation's National Center for Atmospheric Research (NCAR). The NOAA portion of COADS is currently supported by the NOAA Climate and Global Change (C&GC) Program and the NOAA Environmental Services Data and Information Management (ESDIM) Program.

References

Diaz, H.F. and H.-J. Isemer (Eds.), 1995: *Proceedings of the International COADS Winds Workshop, Kiel, Germany, 31 May-2 June 1994*. NOAA Environmental Research Laboratories, Boulder, Colo., 301 pp.

Parker, D.E., C.K. Folland, and M. Jackson, 1995: Marine surface temperature: observed variations and data requirements. *Climatic Change*, **31**, 559-600, 1995.

Reynolds, R. W. and T. M. Smith, 1994: Improved global sea surface temperature analyses. *J. Climate*, **7**, 929-948.

Slutz, R.J., S.J. Lubker, J.D. Hiscox, S.D. Woodruff, R.L. Jenne, D.H. Joseph, P.M. Steurer, and J.D. Elms, 1985: *Comprehensive Ocean-Atmosphere Data Set; Release 1*. NOAA Environmental Research Laboratories, Climate Research Program, Boulder, CO, 268 pp.

Wolter, K., 1997: Trimming problems and remedies in COADS. *J. Climate*, **10**, 1,980-1,997.

Woodruff, S.D., R.J. Slutz, R.L. Jenne, and P.M. Steurer, 1987: A comprehensive ocean-atmosphere data set. *Bull. Amer. Meteor. Soc.*, **68**, 1239-1250.

Woodruff, S.D., S.J. Lubker, K. Wolter, S.J. Worley, and J.D. Elms, 1993: Comprehensive Ocean-Atmosphere Data Set (COADS) Release 1a: 1980-92. *Earth System Monitor*, **4**, No. 1, 1-8. [On-line at: http://www.cdc.noaa.gov/coads/coads1a.html]

Woodruff, S.D., H.F. Diaz, J.D. Elms, and S.J. Worley, 1998: COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth*, **23**, 517-527. [On-line at: http://www.cdc.noaa.gov/coads/egs_paper.html]