Digitizing and quality control of US Marine Meteorological Journals

National Marine Data and Information Service

25 Nov. 2000

1 Introduction

According to the Joint Implementation Plan between the National Oceanic and Atmospheric Administration (NOAA), National Environmental Satellite, Data, and Information Service (NESDIS), National Climatic Data Center (NCDC), and the National Center for Atmospheric Research (NCAR), of the United States of America and the State Oceanic Administration (SOA), National Marine Data and Information Service (NMDIS), Chinese National Oceanographic Data Center (CNODC), of the People's Republic of China signed in May 1998 CNODC and NCDC cooperate to digitize the US Marine Meteorological Journals for the period 1878--1894. For the Chinese side CNODC will prepare the records for transcription, transcribe, digitize, and quality control of the data from the provided paper copies and will complete the project within three years of the receipt of the first copies of the journals. For the US side NCDC will furnish CNODC with paper copies of each frame containing Marine Meteorological journals (MMJ). NCDC will develop a draft-digitizing format prior to the beginning of the digitizing effort; define keying formats and relative codes.

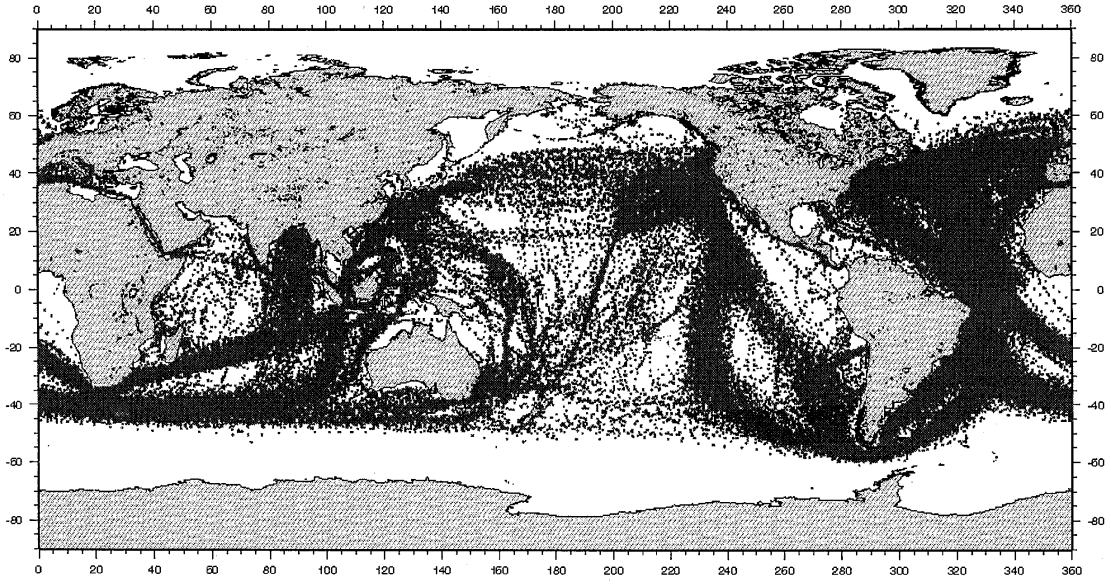
The Joint Coordination Panel for Data and Information Cooperation made this implementation plan based on US--PRC Protocol on Cooperation in the Field of Marine and Fishery Science and Technology. This is the second cooperative digitizing project pursued between the US and PRC under this protocol; the first project of the Murry Collection digitization was successfully completed

2 Summary of Marine Meteorological Journal(MMJ)

The Marine Meteorological Journals contains marine weather observations within the period of 1878--1894. The main parameters include Current Speed and Direction, Wind Direction and Mean Beaufort Force, Barometer, Air Temperature, Water Temperature, Form of Clouds, Clear Skies, Sea State and weather. Most of the above parameters were observed in every two hour. There are 1.98 million records, 156,000 stations. 19 countries in the world observed these marine weather data. ships

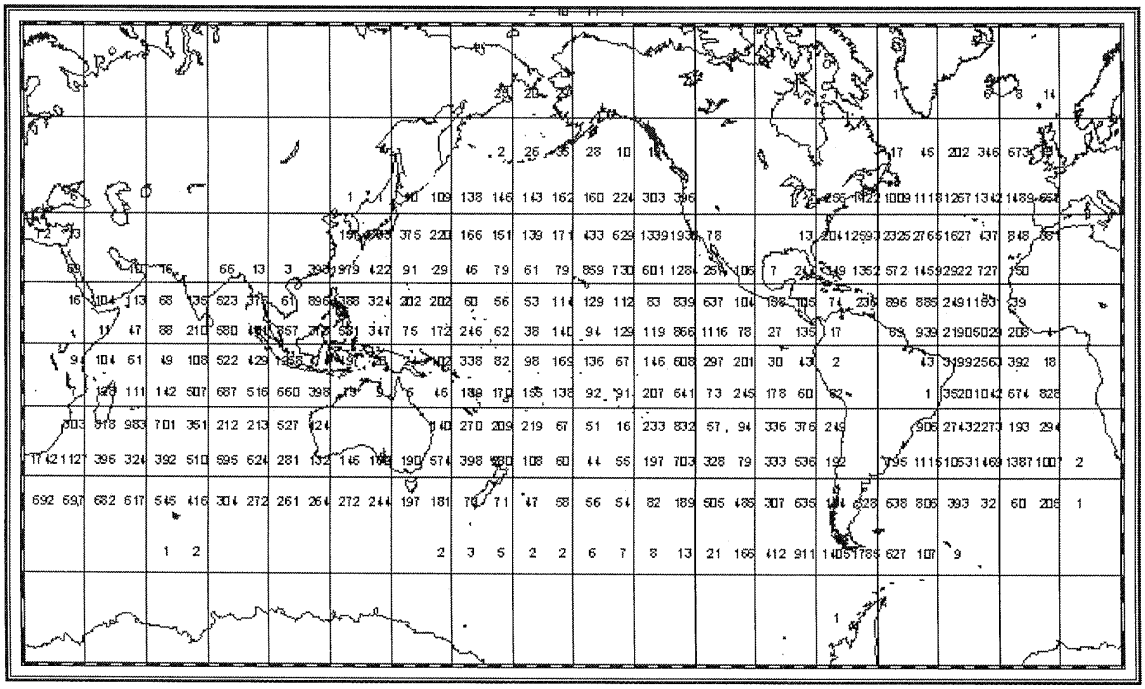
17 years

STATION DISTRIBUTION of MMJ



National Marine Data and Information Service,SOA,P.R.China

Distribution of MMJ Data



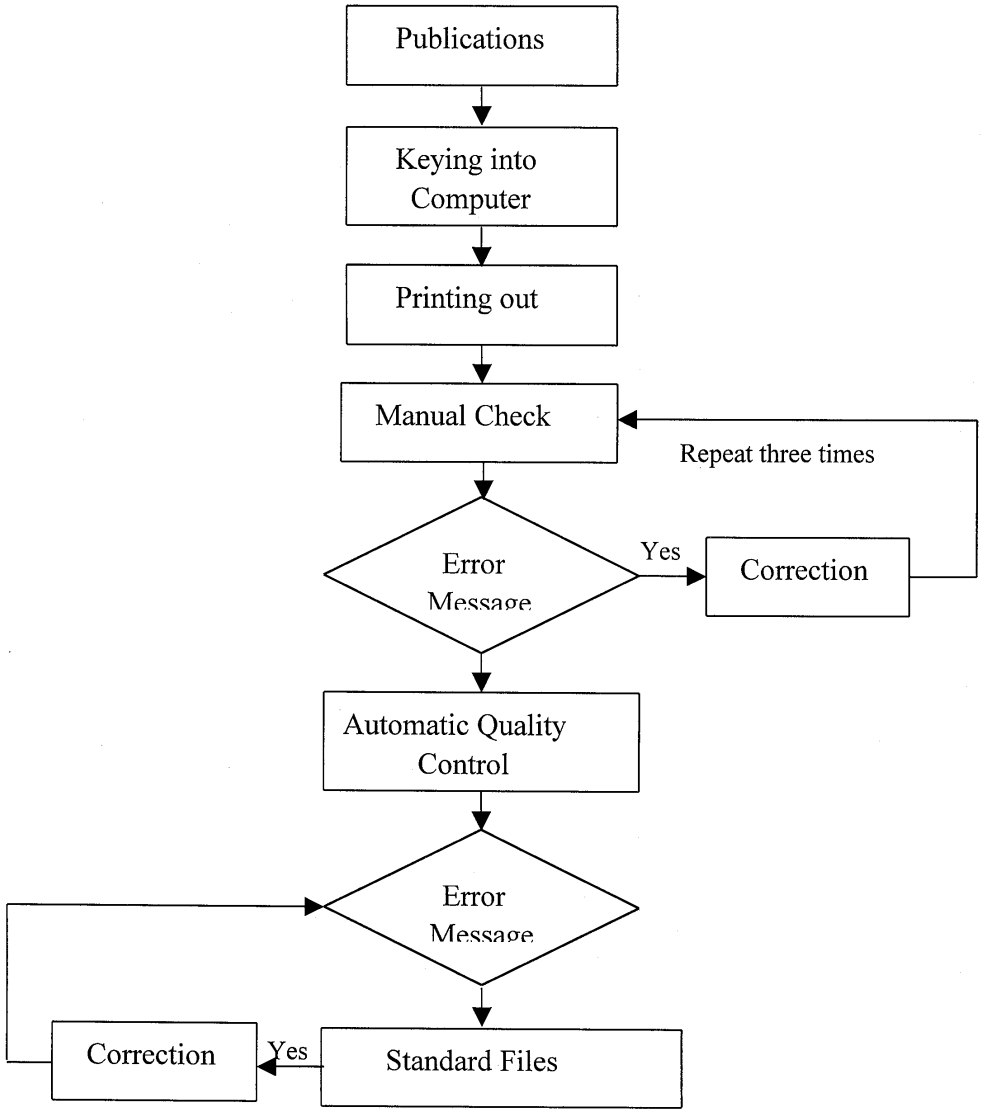
Statistics of Observed Stations (156,022)

Year	Stations	Year	Stations	Year	Stations
1878	396	1884	24108	1890	2360
1879	5952	1885	38010	1891	1129
1880	4069	1886	35738	1892	1133
1881	3687	1887	13514	1893	402
1882	3119	1888	6807	1894	150
1883	11749	1889	3699		

3 Quality Control of US Marine Meteorological Journals

In order to get higher quality we have to do some quality control, especially we have to design some quality control and quality assurance methods. During digitizing the marine meteorological journals the following procedure are was widely used.

Digitizing Flow Chart



3.1 Manual Check

It would be difficult without any errors during keying the data into a computer. But we have to try to reduce the errors to a minimum. The following measure has been taken:

After keying the raw data in designed format we print out them and let one specialist to check as compared with the original journals. After correcting any errors found by the specialist we print out them again and let another specialist to check as compared with the original journals. This procedure is repeated three times. We think this procedure can overcome some systematic errors and any clerical errors. By this procedure we could keep the digitized data as like the raw data and reduce the keying errors.

3- individual checks

3.2 Automatic Check

After finishing the above procedure we need to do some automatic quality control in order to make sure that the digitized data was reasonable. The following quality control methods are used:

a) Data field verification
b) Data relationship verification
c) Statistical test

3.2.1 Data Field Verification

The data field verification procedure performed a variety of automatic checks on individual data fields to assure that the characteristics of any given field matched characteristics defined as acceptable for that field. The tests were performed to assure that:

a) Each field contained the expected data type.
b) Data values were written within acceptable ranges.
c) Fields filled with some codes contained acceptable code values
d) Mandatory fields were present

The following table is a sample for data field verification:

Latitude	0 <= X <= 90
Longitude	0 <= X <= 180
Year	start date <= X <= end date
Month	1 <= X <= 12
Date	1 <= X <= 31
Hour	0 <= X <= 24
Time	0 <= X <= 60

Where X = Observation Value

3.2.2 Data Relationship verification

The data relationship verification procedure was used to evaluate the appropriateness of data values in relation to other data values. The following quality control methods are used:

1) The minimum control distance between two adjacent stations.
2) The maximum control distance between two adjacent stations.
3) The maximum time interval between two adjacent stations.
4) The maximum changed angle between two adjacent stations.
5) The maximum ship speed between two adjacent stations.
6) The position cannot be on land.
7) The reel number cannot be different if the data come from same reel.
8) The journal number cannot be different if the data come from same journal.
9) The observation time must be less the time of the above station if the data come from the same journal.
10) The page number must be less or equal the above page number if the data come from same reel.
11) The difference of wind or course directions in two hours cannot be grater than 150 degrees.
12) The wind speed cannot be grater than 20 miles per hour. maybe mph / hour
13) The difference of barometers cannot be grater than 2.0 hpa per hour.
14) The difference of air temperatures cannot be grater than 2.0 C per hour.
15) The difference of water temperatures cannot be grater than 1.6 C per hour.
16) The difference of air and water temperatures at the same time cannot be grater than 12 degrees in F.
17) The form of clouds and the clear of skies cannot be conflicting with the weather at the same time.
18) The phenomena of weather at the same time cannot be conflicting with each other. We also used other methods to do data relationship verification.

By the above procedure we find many problems rose from the original journals and resolved them. We think that the corrected data will be more reasonable.

3.3 Data Statistical test

The statistical tests are used to determine the acceptability of data values in comparison to existing values. We used the following test models: Water Temperature, Air Pressure, Air Temperature and so on.

The following formula was used: how were M and SD determined

$$M-3.5*SD < X < M + 3.5*SD$$

Where M = Mean value

SD = Standard Deviation
X = Observation value

3.4 Error message and analysis

During automatic check some error messages can be found. We have to design some methods to judge the messages in order to correct them. The following procedures are used.

1) To see if the value is same as the value of the raw data. If it is not same as the original, we just correct it. If it is same as the original, we must do some analysis and find the reason of the errors.

2) To see if it is a kind of systematic errors. For example, we find some barometers of one journal are too low. After analysis we found that this is a systematic error. In this case the pressure values were in "Paris inches" not English units. To correct these values we have to multiply the value by the conversion factor (1.066). In some cases we find that the weather is rain, but the clear sky is 10(that means no cloud in the sky). Here the observer was reporting the amount of cloud and not the proportion of clear skies. This is a systematic error. In this case the clear sky should be 10 - observer's value.

3) To see if it is a clerical error. If it is a clerical error, we must correct them. For example, we find that at one station the location is 1809 S and 2632 W, but the location of the adjacent station is 2526 S and 1832 W. This is a clerical error. In this case the observer confused latitude and longitude. The correct value should be 1832 S and 2526 W.

After automatic quality control we think we can get higher quality data.

4 Special Problems

During digitizing the Marine Meteorological Journals we find some special problems, which need to be resolved in order to make the digitized data be right. The followings are main problems.

4.1 Barometer

Pressure value was reported as 30.50/100. In this case the pressure should be keyed as 30.05.

Barometer value was reported as: 81. In this case: means '29'. So: 81 should be '2981'.

Barometer value was reported as 283/12. In this case the pressure values were in 'Paris inches'. To correct these values we should multiply the value by the conversion factor 1.066.

Barometer value was reported in characters such as 'Steady'. In this case characters describe the barometer. Here 'Steady' means that the barometer value is the same as above value.

Barometer value was reported as 305. In this case we have to check the above records. After analysis the barometer should be 3005.

4.2 Temperature

Attached thermometer was reported in Fahrenheit, but the Dry bulb, wet bulb and water temperatures were reported in Centigrade. In other cases the attached thermometer was reported in Centigrade, but the Dry bulb, wet bulb and water temperatures were reported in Fahrenheit. In order to express temperature value of each parameter at one record we used temperature indicator. For example,

'3' = attached thermometer was reported in Fahrenheit, with Dry bulb, wet bulb and water temperatures in Centigrade.

'4' = attached thermometer was reported in Centigrade, with Dry bulb, wet bulb and water temperatures in Fahrenheit.

4.3 Course Direction

Course direction is reported as 'SE1/41/2E'. In this case we should code it as SE18E (based on 1/4 of 1/2 = 1/8).

4.4 Proportion of clear skies

Observers were reporting the amount of cloud instead of the proportion of clear skies. In this case we should do some conversion by subtracting this reported value from ten.

The proportion of clear skies was reported as '1/10'. In this case the observer was indicating that over the two-hour observational period the cloud amounts varied from very few clouds to nearly overcast condition. So the lowest numeric value '1' should be entered.

4.5 Current Speed

Current speed was reported as '39' or '45' nautical miles. In this case the nautical miles provided for the current speed was over a 24-hour period. The value should be divided by 24.

Current speed was reported as '100' nautical miles by 96 hours or 4 days. In this case the observer was reporting the current in total nautical miles from one to four days. 96 should divide the value.

4.6 Course steered

The course steered is reported as SzE1/2E, SOzO or SzW. In this case the ship is of German registry so the observer was occasionally switching between his native language

(German) and English. This caused a number of problems in interpreting what was intended. For example 'O' (Osten in German) was used to represent 'East' (English). Also the German word 'zu' means 'to' or 'towards'. So above 'SzE1/2E' should be keyed as 'STE1/2E'. Then above 'SOzO' should be keyed as 'SETE'. 'SzW' should be keyed as 'STW'.

Course steered was reported as 'NW72W' or 'WbyS72S'. In this case it is a clerical error. The observer wrote '1' which looks like '7'. So 'NW72W' means 'NW12W' and 'WbyS72S' means 'WbyS12S'.

Course steered was reported as 'W30N'. The general practice was to report the degree off of north or south in either the east or west direction. However, here the observer was using east and west and varying them in either the north or south direction based on the number of times such entries appeared. The 'W30N' would equal 'N60W' in the generally accepted convention of the time. So both 'W30N' and 'N60W' would approximate 'NWXW.

Course steered was reported as 'N100E' or 'N107W', etc. In these cases the observer was reporting in degrees of up to 180 either east or west of north. Therefore, 'N90E' would be east, 'N90W' would be west, and 'N180E' or 'N180W' would be south. So 'N108E' would be 'S72E'.

4.7 Ship speed

Ship speed was reported as distance traveled in nautical miles over the two hours period between reports. In these cases the ship speed should be half the distance.

4.8 Wind direction

Wind direction was reported as 'N1/8W', 'S1/2SE' or 'NE3/4E'. In this case we used the following codes.

Code	Description
18	1/8
14	1/4
38	3/8
12	1/2
58	5/8
34	3/4
74	7/8

Wind direction was reported as 'ZOTO'. In this case it is a report from a Dutch ship. The following is the four primary directions in a Dutch to English directory:

English	Dutch
North / Northerly	Noorden / Noordelijk
South / Southerly	Zuiden / Zuidelijk
East / Easterly	Olsten / Oostenlijk

West /Westerly Western / Westelijk

Wind direction is 'ON'. In this case it is a report from a Portuguese ship. It is reported in Portuguese. Here 'O' means 'West' and 'ON' means 'NW'.

Wind direction is reported as 'SEXS1/2E'. In this case it represents a wind direction half way between SEXS and SE. To indicate the finer resolution the entry could be 'SEXS1/2E'.

4.9 Form of Clouds

Form of clouds was written by bird symbols. There are two cloud forms:
Primary cloud form:

One bird means Cir
Two birds means Cir
Three birds means cum
Four birds means Nimb

Secondary cloud form:

One bird means Cir
Two birds means Cir, Cum
Three birds means Cir, Str.
Four birds means Cum, Str.

These are very interesting entries where the observer drew in the birds to represent the cloud types.

4.10 Weather

Weather is reported 'H'. According the code of weather 'H' means 'Hail'. But the particulars of the weather during the day are 'Hazy'. In this case the 'H' really means 'Haze' and not 'Hail'.

There are a lot of special problems, which were found during quality control. The above examples make us know that during digitizing any historical observation data we have to do some analysis in order to be sure that the digitized data is correct.

5 Acknowledgments

Now the joint implementation plan has been completed within 2 years. CNODC has provided all copies of the quality controlled digitized records to NCDC. At this time we appreciate NCDC, especially Mr. Joe D. Elms who gave us a lot of help during digitizing the Marine Meteorological Journals and help us resolve many unusual and significant problems associated with the digitization of the original logbooks. Mr. Joe D. Elms works very hard, so that the problems found from the journals can be resolved rapidly and the project can be completed within the period of schedule.