The International Surface Pressure Databank (ISPD) land component version 2.2

NOAA's National Climatic Data Center, Product Development Branch

Xungang Yin, Byron Gleason, Russell Vose

University of Colorado, CIRES, Climate Diagnostics Center
NOAA's Earth System Research Laboratory

Gilbert Compo, Nobuki Matsui

November 13, 2008

ISPD

The International Surface Pressure Databank (ISPD) land component

Instruction Manual
Version 2.2

How to Reference this Publication:

Yin, X., B.E. Gleason, G. P. Compo, N. Matsui, R. S. Vose, "The International Surface Pressure Databank (ISPD) land component version 2.2", National Climatic Data Center, Asheville, NC, pp.1-12, 2008.

About the Cover Image:

The image on the front cover is taken from NOAA's National Weather Service Hydrometeorological Prediction Center Product Suite (www.hpc.ncep.noaa.gov).

Table of Contents

Introduction

The International Surface Pressure Databank (ISPD) is a large global collection of surface and sea level pressure data. The ISPD 2.2 represents an update to a previously developed but unreleased version of ISPD (version 2.0). The result is that users will find a greatly enhanced and expanded product from the last ISPD release (version 1.1). ISPD 2.2 upgraded many stations with more current data and expanded by adding new stations. In addition new duplicate checks were employed to assist in assembling all of the data into a single dataset.

Table 1 below highlights some statistics of the ISPD 2.2:

Number of total stations	33,628
Number of total files:	543,144
Number of total records:	1,199,332,691
Number of total pressure obs:	1,751,291,205
Number of total sea level press. obs:	1,095,599,956
Number of total surface press. obs:	655,691,249
First recorded year of data:	1768
Last recorded year of data:	2008

Table 1

Figure 1 depicts the total number of sea level pressure (blue line) and station pressure (grey line) observations per year. (Note: data begin in 1768)

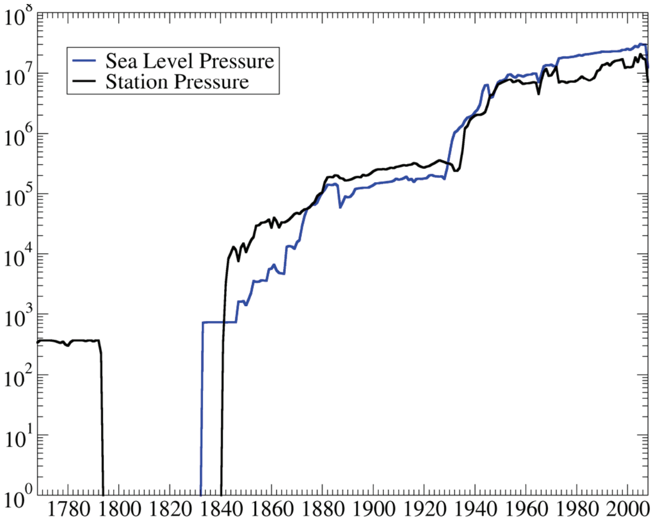


Figure 1

Duplicate Detection

Duplicate data represent the most significant challenge to assembling the ISPD 2.2. The existence of duplicate data within a dataset can lead to inflated analysis bias, incorrect trend calculations, and significant increase in the overall size of the dataset. Thus techniques were developed to identify potential duplicate data and remove the least desirable sources of duplication. The ISPD 2.2 employed a two step duplicate removal process. First exactly duplicated monthly data from among all of the stations were removed and second similar stations (based on metadata comparisons) were removed for “common” days.

Duplicate Removal Step #1 (removal of identical months):

The first duplicate removal step focused on identifying identical months of data from not only within a station, but also among every other station within the ISPD 2.2. Of course direct comparison of every single month involves a significant amount of processing, and therefore was impractical due to time constraints. Therefore, a technique was needed to decrease the “brute force” method of direct comparison down to a simpler method that scaled linearly with time rather than at some more expensive non linear rate. In order to accomplish this task in a reasonable time, a similarity hash was developed. The hash was generated based upon the date and pressure values for the entire month. More specifically the hash is a vector with both magnitude and direction. The magnitude was a floating point number that would be identical for duplicated months, and very similar in numerical space for similarly or near duplicated months. The direction was represented by an angle that could vary between 0 degrees and 90 degrees. Thus, for each month of data within the ISPD 2.2 a floating point number and an angle were calculated. This technique ensured that only one “read” or “pass” through the dataset was needed in order to generate all of the station-month hash values. After calculating all of the hash values, they could then be easily sorted for direct inspection of exact or similar values.

The hash generation algorithm involved reading all records for a particular month and storing the date and pressure value (in record order) for every record in one long sequence. The sequence was then compressed to just integer values ranging from 0 to 9, so decimal points, symbols (e.g. “+”), etc. were set to “nine” for consistency. The integer values in the sequence could then be used to calculate the properties of the vector.

Example (Duplicate Vector Calculation):

Consider a monthly record string as follows, whereby the record date, time, and pressure information for each observation (record) are extracted as follows:

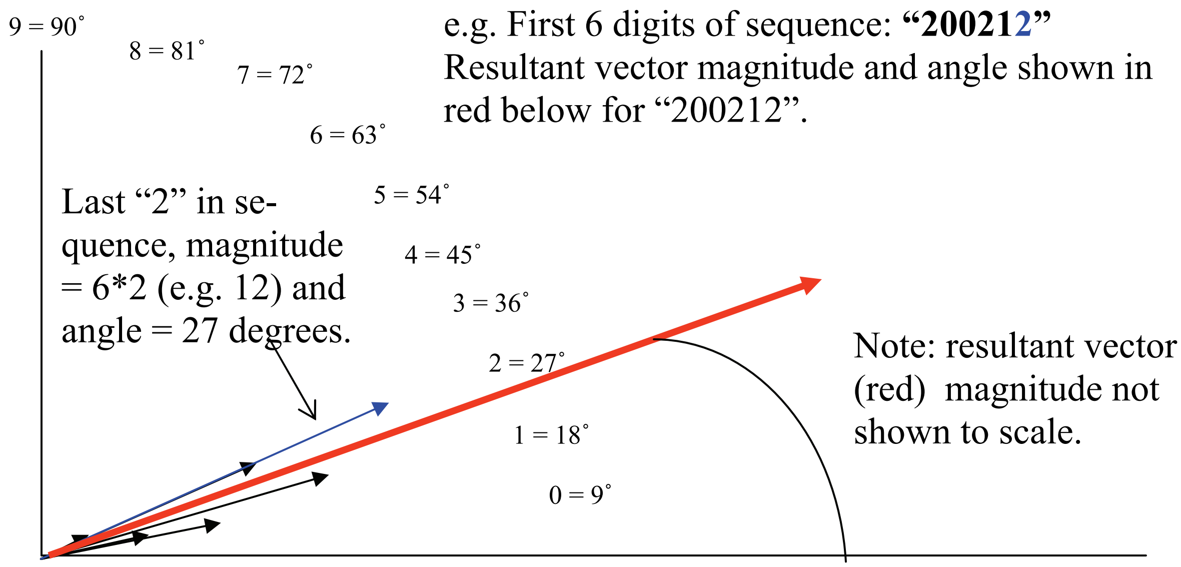
“2002121400011001.2”

the deciphered data represent a pressure of 1001.2 on December 14, 2002 at 00 U.T.C. and 01 minute past the hour.

The string is compressed (e.g. substitute decimal point with “nine”) to ensure all values are integers between 0 through 9. The resultant string becomes,

“200212140001100192”.

To calculate the “hash” vector magnitude and angle, calculate the resultant vector (using parallelogram rule) obtained by plotting each individual integer within the sequence as a vector whereby the magnitude of each vector is its integer value multiplied by its integer position within the sequence and an angle corresponding to one of the 10 angles (0 through 9) listed in Figure 2.



e.g. First 6 digits of sequence: **“200212”**
Resultant vector magnitude and angle shown in red below for “200212”.

Last “2” in sequence, magnitude = 6*2 (e.g. 12) and angle = 27 degrees.

Note: resultant vector (red) magnitude not shown to scale.

Figure 2

Thus, extending the example in figure 2 to include the whole sequence for the entire month would allow for the calculation of the resultant vector's magnitude and angle for the entire sequence. The scalar magnitude and angle would serve as the "hash" used as a basis of comparison for that month. These hash values are calculated for every station data month. It is important to mention for this step within ISPD 2.2, only exactly equal scalar magnitudes and angles were considered duplicates. In the future, since the algorithm can provide results which measure similarity it may be possible to redefine "duplicates" and delete close but not exactly duplicated months.

<u>Duplicate Step Removal #2 (removal of similarly located stations)</u>

Station latitude and longitude coordinates were compared for all stations and similarly located stations were identified. These "clusters" of stations were then examined for duplicate data within each cluster. A ranking scheme was used to identify and remove all but the highest ranked data. More specifically, clusters were identified by finding stations whose difference in both latitude "and" longitude were no larger then 0.1 degrees. Over 37 million observations were identified as duplicates using this method. Figure 3 below depicts a cluster of stations within the Barrow, AK region reduced to non duplicate data (shown in red).

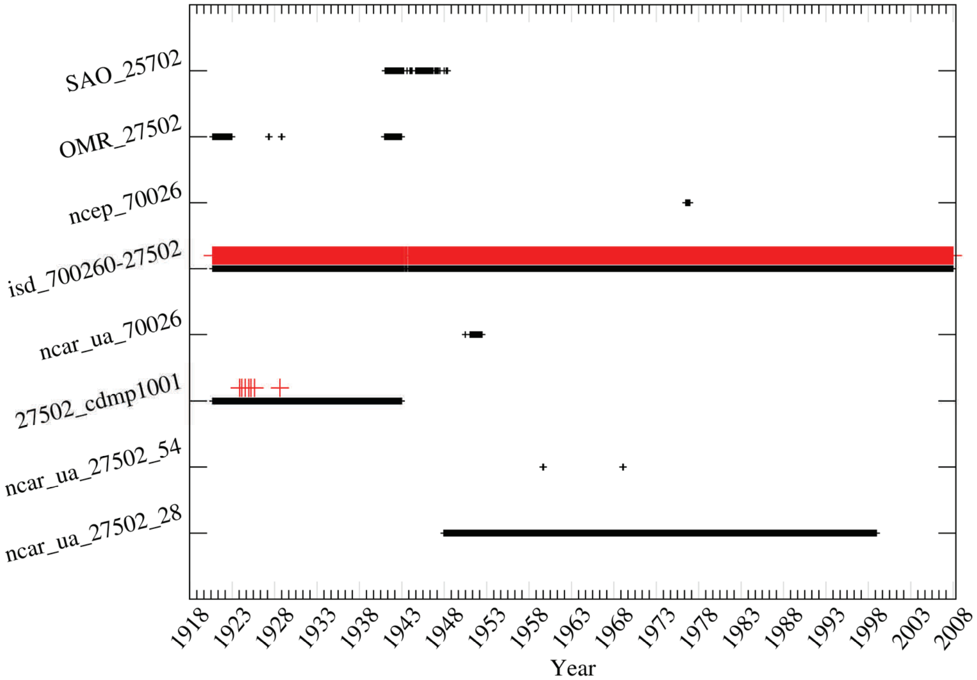


Figure 3: Barrow, AK stations with individual pressure observations for each station indicated by the black crosshairs. The red crosshairs indicate the "retained" data after the duplicate removal process.

Format Compliance

The following format checks were used to verify correct entries for various fields and ensure the integrity of the dataset. For more information on each individual field please see the dataset format documentation (Appendix B). Note, not all fields were checked, since some fields could have a near infinite number of ways of being expressed (e.g. “original data”).

1) no record longer than 402 character record limit.

2) no blank fields

3) The following fields were checked to ensure they were of the proper “type”, (integer, floating point, or character):

Fields 04, 05, 06, 07, 08, 11, 12, 13, 14, 16

4) The following fields were check to ensure they only contained a valid entry (e.g. Field 17 must be one of “M”, “0”,”1”,”9”):

Fields 02, 03, 10, 15, 17, 27, 29, 34, 36, 38, 40, 42, 43, 44, 48

5) The following fields were checked to ensure all values were within a certain range:

Field 04,	1700 ≤ YEAR ≤= 2008
Field 05,	01 ≤ MONTH ≤ 12
Field 06,	01 ≤ DAY ≤ 31
Field 07,	00 ≤ HOUR ≤ 23
Field 08,	00 ≤= MINUTE ≤ 59
Field 11,	-90.00 ≤= LATITUDE ≤ 90.00
Field 12,	0.00 ≤= LONGITUDE ≤ 359.99

6) no record containing missing sea level pressure “and” missing station pressure

Appendix A: Contact

The main contact for the ISPD 2.2 is Dr. Gilbert Compo of NOAA's Earth System Research Laboratory and the University of Colorado/CIRES Climate Diagnostics Center. Dr. Compo is one of the original founding project members of the ISPD effort, the number one user of the dataset (at the time of this writing), a significant contributor of source data, and so far the ISPD’s most consistent data set champion. Please contact Dr. Compo for any questions related to the ISPD 2.2.

Dr. Gilbert P. Compo
325 Broadway
Boulder, Co 80305-3328
P: (303) 497-6115
F: (303) 497-6449
Gilbert.P.Compo@noaa.gov

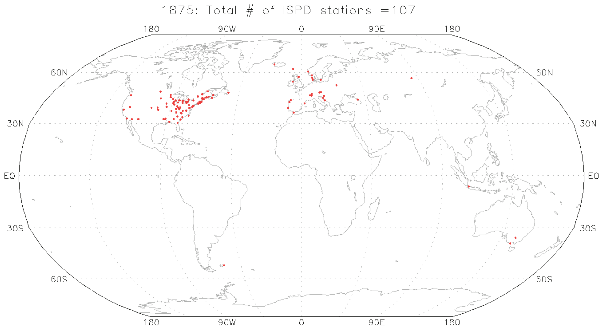
Appendix B: NCDC ASCII Interchange Format Specifications

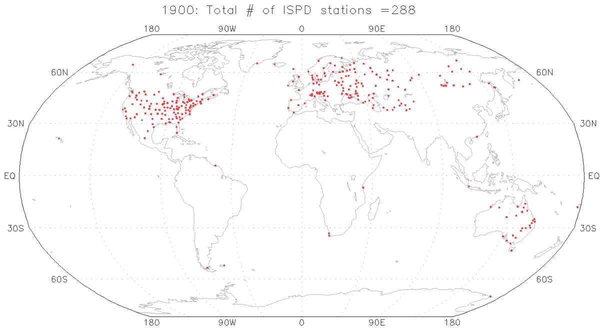
Many different sources of data have contributed to the ISPD 2.2 In order to facilitate processing, quality control, and distribution it was necessary to convert all of these diverse sources into a common format. The “NCDC ASCII Interchange Format” is the official format of ISPD 2.2. The documentation of the format is found here:

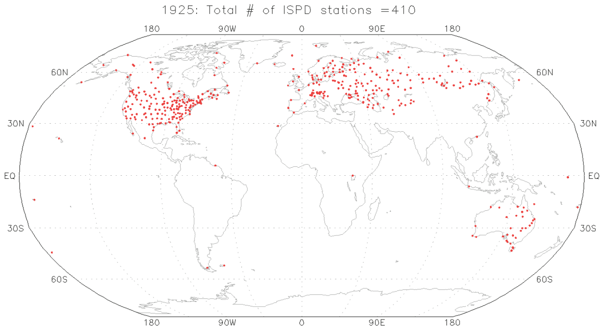
http://www1.ncdc.noaa.gov/pub/ispd/doc/ASCII_transfer_v1_0.pdf

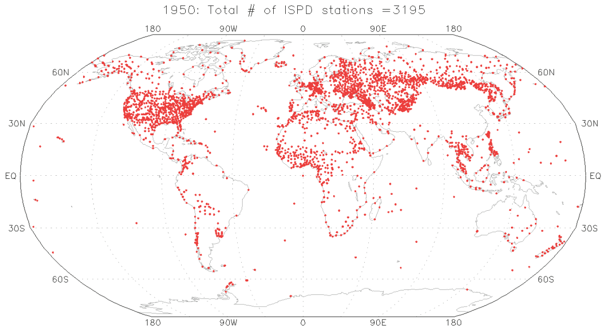
Appendix C: Spatial Coverage of ISPD 2.2 data

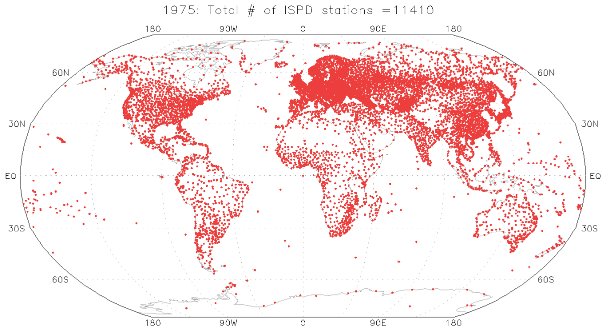
Spatial Coverage of ISPD 2.2 Pressure Data (note: criterion for plotting a station: at least one non-missing annual pressure value)

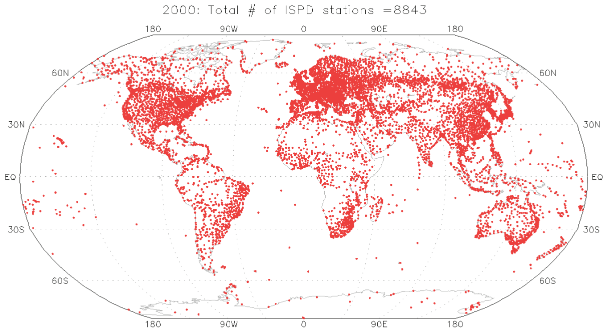












Appendix D: Data License

The ISPD 2.2 is available for research purposes to any person(s) or institutions that contributed data. Please cite the dataset per the citation listed at the beginning of this manual. If you did not contribute to the dataset, please refer to the contact (Appendix A), for specific requests about using the data.