Integrated Global Radiosonde Archive V2.2

Product Description Supplement

National Centers for Environmental Information (NCEI)

Prepared by

Imke Durre, Bruce Hundermark, Xungang Yin, Bryant Korzeniewski

January 2023

Contact: ncei.igra@noaa.gov

Table of Contents

Acronyms and Abbreviations

Acronym	Meaning
AIRS	Aerological Information and Resource System
ARL	Air Resources Laboratory
BAS	British Antarctic Survey
BUFR	Binary Form for the Representation of meteorological data
CARDS	Comprehensive Aerological Reference Data Set
CDMP	Climate Data Modernization Program
DOI	Digital Object Identifier
DSI	Data set identifier
ECMWF	European Centre for Medium-Range Weather Forecasts
EPA	Environmental Protection Agency
ERA-CLIM	European Reanalysis of Global Climate Observations
ETH	Eidgenössische Technische Hochschule
FD	First difference
FIPS	Federal Information Processing Standard
FTP	File Transfer Protocol
GCOS	Global Climate Observing System
GFDL	Geophysical Fluid Dynamics Laboratory
GPS	Global Positioning System
GTS	Global Telecommunications System
GUAN	GCOS Upper Air Network
HMSC	Hydrological and Meteorological Service Center
hPa	Hectopascals
ICAO	International Civil Aviation Organization
IACS	Institute for Atmospheric and Climate Science
ID	Identifier
IGRA	Integrated Global Radiosonde Archive
IGRA ID	IGRA station identifier
IORGC	Institute of Observational Research for Global Change
JAMSTEC	Japan Agency for Marine-Earth Science and Technology
LKS	Lanzante/Klein/Seidel
NCDC	National Climatic Data Center
NCAR	National Center for Atmospheric Research
NCEI	National Centers for Environmental Information
NCEP	National Centers for Environmental Prediction
NMC	National Meteorological Center (the former name of NCEP)
NOAA	National Oceanic and Atmospheric Administration
NWS	National Weather Service
NWSTG	National Weather Service Telecommunications Gateway
PW	Precipitable water
QC	Quality Control
RATPAC	Radiosonde Atmospheric Temperature Products for Assessing Climate
RH	Relative humidity
STD	Standard Deviation
TAC	Traditional Alphanumeric Form
TD	Tape deck
U.S.	United States

USAF	U.S. Air Force
UTC	Universal Coordinated Time
WBAN	Weather Bureau, Army, Navy
WMO	World Meteorological Organization

1. Introduction

1.1 Purpose

This document is a supplement to the Product Description for version 2.2 of the Integrated Global Radiosonde Archive (IGRA). Its primary purpose is to provide additional background and technical details regarding the data, derived products, and metadata that constitute IGRA 2.2. Specific sections inform the user of caveats and potentially significant characteristics of these products, describe the source data and how they were processed, and provide a high-level guide to understanding the processing system and its component algorithms. Interested readers may include IGRA users, personnel responsible for processing IGRA, and dataset developers in general.

While much of the information in this document references IGRA v2.0, most of the procedures described also apply to the current version of IGRA, v2.2. Changes made in the transition from v2.0 to v2.2 are summarized in Section 1.4.4 and are indicated, as needed, in other parts of this document.

Several other documents provide additional information about IGRA. The scientific context and rationale for the creation of IGRA 1 and 2 are documented in peer-reviewed papers (Durre et al. 2006, 2008, 2018). At each NCEI-owned URL where IGRA is available, a readme file orients the user to the folder and file structure used, and the exact contents and format of the IGRA-related data and metadata files are described in various format documentation files within these folders.

The actual algorithms are defined by the computer scripts and programs that are stored and version-tagged in NCEI's internal Git repository. Basic information on the IGRA processing system is documented in the IGRA Operations Document. Guidance for how to locate, set up, and execute these programs is provided in the IGRA Operations Document's Supplement. The Supplement to the Operations Document also contains details about each data atd metadata source used in IGRA. The Operations Document and Supplement are available from NCEI upon request.

1.2 Document Maintenance

This document will be maintained in a manner consistent with version control practices. When a new version of IGRA is ready to be deployed, this document will be edited to ensure its continued consistency with the operational version of the dataset.

Document Version	Description	Month/Year Completed
1.0	Created from Sections 1-4 and 9 of the IGRA v2.0 Dataset Description	January 2023

	Document. Incorporates changes made for IGRA v2.2 as well as various typographical corrections.	

1.3 IGRA Synopsis and Access

IGRA is a collection of historical and near-real-time radiosonde observations from around the globe. The archive consists of five components: quality-assured individual soundings, monthly means, sounding-derived humidity and stability parameters, station history information, and the homogeneity-adjusted time series of temperature anomalies that constitute RATPAC. All of these components are updated on a regular basis and are made available in ASCII format.

The IGRA sounding data, derived parameters, monthly means, and station history information can be accessed from the links provided on the IGRA webpage at https://ncei.noaa.gov/data/integrated-global-radiosonde-archive/ .

At each download location, a readme file provides an orientation to the respective files and subfolders as well as a quick start guide for locating the desired data. Various *-format.txt files located throughout both folders contain an exact description of the format of each group of files.

All RATPAC-related files can be found at https://www.ncei.noaa.gov/data/ratpac/ and ftp://ftp.ncei.noaa.gov/pub/data/ratpac. The readme.txt file, located in the directory, describes both the file and folder structure as well as the formats of the various RATPAC files.

The following table summarizes the access methods, dataset identifiers, and key publications for both IGRA and RATPAC.

	IGRA	RATPAC
Contents	Soundings, derived parameters, monthly means, station histories	Bias-adjusted monthly-mean temperature time series
DOI		https://doi.org/10.7289/V5SF2T7J
Webpage	https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive	https://www.ncei.noaa.gov/products/weather-balloon/radiosonde-atmospheric-temperature-products
Data access	https://www.ncei.noaa.gov/data/integrated-glo	https://www.ncei.noaa.gov/data/ratpac/

	bal-radiosonde-archive/	
Alternative data access (FTP)	ftp://ftp.ncei.noaa.gov/pub/data/igra/	ftp://ftp.ncei.noaa.gov/pub/data/ratpac/
NCEI DSI	6353_01	6354_01
Dataset citation	Durre et al. (2022a)	Durre et al. (2022b)
Key publication	Durre et al. (2018)	Free et al. (2005)

1.4 IGRA History

IGRA and RATPAC are maintained, archived, and distributed by NOAA's NCEI, formerly NCDC. The following sections describe the evolution of these data products.

1.4.1 Version 1

Version 1 of IGRA (Durre et al. 2006, 2008), consisting of quality-controlled soundings drawn from 11 data sources and monthly means, was developed at NCDC in the early 2000s and released in 2004 as a successor to NCDC's CARDS (Eskridge et al. 1995).

During the same timeframe, NOAA's ARL, GFDL, and NCDC developed a set of homogeneity-adjusted time series of temperature anomalies for 85 individual high-quality stations as well as for large areas of the globe. The 1958-1996 portions of these time series were derived from CARDS (Lanzante et al. 2003a,b), while post-1996 data were taken from IGRA (Free et al. 2005). This collection of time series, known as RATPAC (Free et al. 2005), was released in 2005 and has been updated operationally at NCDC/NCEI since then.

Since 2006, station history information for IGRA stations has also been made available. Initially, this information was based on the Gaffen (1996) historical records of equipment, observing and reporting practices, processing conventions, and station locations for many upper-air stations. The first release of IGRA station history information also incorporated updated information for a subset of the stations that had been collected by NCDC during 2003-2005. Subsequently, four updates were made by Steve Schroeder of Texas A&M University in 2007-2009 and by Liz Zhang of NCAR in 2010.

In an effort to facilitate studies of variations in the vertical structure of the atmosphere, NCDC scientists also collaborated with colleagues at ARL and the EPA to calculate a set of relevant parameters from IGRA soundings. The resulting set of IGRA-derived sounding parameters (Durre and Yin 2008) was released in 2008 and expanded to include additional stability and moisture parameters in 2011 (Durre and Yin 2011). The sounding parameters have been updated at NCDC/NCEI on a daily basis since their first release.

1.4.2 Version 2.0.0

After the release of the original IGRA data, additional historical observations

were digitized, previously existing datasets not included in Version 1 became more easily accessible, and new radiosonde stations started operation. Hence, between 2008 and 2014, NCDC scientists worked on a new version of the sounding data with the goals of enhancing spatial and temporal coverage, simplifying the process of adding new data sources to the dataset, and incorporating the most commonly received suggestions from users of Version 1. In addition, the various processes that update the sounding data and monthly means, calculate the sounding-derived parameters, and produce the RATPAC time series were consolidated into one processing system. The result of these efforts was IGRA 2.0.0 (Durre et al. 2018, 2022a).

At the time of its release, IGRA 2.0.0 contained nearly twice as many stations and 30% more soundings than IGRA 1. The increase in data was achieved through the incorporation of more data sources, 33 in total, including 10 of the 11 data sources used in IGRA 1. This change improved the spatial coverage over Europe before the 1970s, China in the 1970s and 1980s, and Africa and Latin America at all times. In addition, data from a subset of fixed and floating ocean-based ships and platforms were included. Further, the additional variables of reported RH and time elapsed since launch were included in the sounding data whenever they were available, and vapor pressure was added to the set of variables available as monthly means.

1.4.3 Version 2.1.0

In 2018, IGRA was rebuilt using a revised station matching algorithm. In addition, additional data from the U.S. Air Force data source that had previously been excluded due to a lack of station names were included.

1.4.4 Version 2.2.0

During 2020-2022, IGRA was upgraded to allow for the incorporation of GTS reports received in BUFR format. Traditionally, GTS reports adhered to a standard format referred to as Traditional Alphanumeric Code (TAC). Starting in 2014, however, upper-air stations began to transition from this decades-old format to the newer BUFR format, which had been approved by the WMO for operational use in 2007. Stations usually commence their transition from TAC to BUFR by disseminating their observations in both formats. This transition has been increasing at an accelerating rate, and BUFR is expected to eventually completely replace TAC at all sites. By December 2022, the number of stations that had switched to only transmit in the BUFR format had increased to more than 100. Consequently, without the inclusion of BUFR data, the records for these stations would no longer be updated in IGRA.

For IGRA 2.2, two significant changes were made to the IGRA processing system. First, parallel processing was implemented to increase the overall system performance in the context of the much larger volume of BUFR data compared to TAC data. Second, a mechanism was developed for using BUFR GTS reports from NWSTG and ECMWF to supplement reports in the TAC-only NCEI/NCEP GTS data source. This was accomplished by first downsampling the high-vertical-resolution BUFR soundings to the historically typical set of standard and

significant pressure levels and then using this low-vertical-resolution version of the decoded BUFR data to fill the TAC GTS gaps. While the native high vertical resolution of the BUFR records is of interest to some users, the use of the low-vertical-resolution versions of the BUFR data in lieu of the original higher-resolution data allowed for greater homogeneity in the IGRA time series.

This gap filling approach allows for the retention of maximum continuity in the TAC records while they exist and while stations experiment with the BUFR format. The gaps in the TAC GTS are mainly caused by stations that ceased TAC GTS dissemination, but may also be due to some soundings only appearing in the BUFR data, even while the corresponding TAC stream was still continuing at the same station, e.g., due to technical reasons. Since the BUFR GTS soundings were integrated into the previously existing NCEI/NCEP GTS data source, no new source codes were assigned to the BUFR sources; they are encompassed by the source code "ncdc-gts" in IGRA.

Together, the upgrades to the IGRA system allowed BUFR-only stations to continue to be updated in IGRA, while also reducing the daily processing time by 75%. As of early 2023, the switch to IGRA v2.2 allowed for the resumption of updates to the records of more than 100 IGRA stations. There was no change in the format or file naming conventions for any of the IGRA -related output files, and no new data source code was added.

1.5 Dataset Version Identification

1.5.1 Definition

Beginning with IGRA 2.0.0, a single version identifier is assigned to the entire suite of IGRA 2 sounding data, sounding-derived parameters, monthly means, and RATPAC time series. The version identifier of each day's IGRA update consists of the version number and the date on which the version was produced.

The version number is of the form "X.Y.Z", where X, Y, and Z are defined as follows:

X is incremented when a major change to the source data or processing procedures results in the modification or addition of a large percentage of the sounding data. A change in X is intended to be accompanied by a peer reviewed manuscript.

Y is incremented when the replacement of an existing data source, the addition of a new data source, or a change to the processing software results in significant modification or addition of a portion of the sounding data values or in the derived products. Such a change should be accompanied by some form of technical note or an update to this Dataset Description Document.

Z is incremented during minor revisions, such as modifications to the processing software that have a minor impact on a small percentage of values in

the sounding data or in one or more of the derived products.

The version date is of the form yyyymmdd, where yyyy is the year, mm is the month, and dd is the day on which the data were last modified in any way, regardless of whether the modification is accompanied by a change in version number. Regular daily updates and routine quasi-annual reprocessing are examples of modifications that are tracked only by a change in the version date.

1.5.2 Version Tracking

The method for tracking the version identifier that accompanies a particular version of the dataset differs between the public-facing data download locations and NCEI's archive.

In NCEI's data download locations for IGRA, the current version identifier can be found in file igra2-version.txt, and all changes in the version number are tracked in the file called igra2-status.txt.

While the version number itself is the same, different means for identifying the version number are required for storage in the NCEI Archive. The approach used is described in the internal data submission agreement for IGRA. To avoid greater divergence of documents, it is not reproduced here.

2. Station Metadata

2.1 IGRA Stations

2.1.1 Definition

The data within IGRA and its derived products are generally organized according to the IGRA station. An IGRA station may represent a single observing location, may consist of a series of nearby locations over time, or may represent a ship or other mobile station. IGRA 2 consists primarily of land-based stations whose location has remained relatively fixed over time. Other "fixed" IGRA stations include moored ships and buoys. Mobile IGRA stations, on the other hand, include Volunteer Observing Ships and ice islands.

A listing of all IGRA stations, including their IGRA IDs and current names and locations, can be found in file igra2-stations.txt. See file list-format.txt for a description of the format of this IGRA station list. The subset of 85 stations whose data are used in RATPAC are listed in a corresponding file in the RATPAC FTP directory.

2.1.2 Station Identifiers

Each IGRA station is identified by an 11-character IGRA identifier (IGRA ID) that consists of a two-character country code, a one-character station network code, and an eight-character station identifier.

The country code, which is based on FIPS, indicates the country in which the station is located. A country code of "ZZ" identifies an ocean location and is presently used for moving ships and ice islands as well as for moored ships and buoys. A complete list of country codes used is provided in file igra2-countries.txt. The current country code is used in all of a station's data and metadata records even if the station has been under the control of different countries throughout its history. Some significant changes in control are noted in station names or comments in the IGRA metadata file (see Section 2.2).

The network code specifies what type of station identifier follows. IGRA station IDs for sites identified by WMO station numbers contain a network code of "M" followed by "000" followed by the five-digit WMO identification number. For example, the IGRA ID for Key West (WMO# 72201) is USM00072201. If more than one WMO ID has been assigned to a particular site over time, the most recent WMO ID is used in the IGRA ID. Other types of station IDs used include call signs for Volunteer Observing Ships (network code "V"), WBAN numbers ("W"), ICAO call signs ("I"), and specially constructed identifiers ("X"). A specially constructed ID is used only when none of the other types of ID have ever been assigned to a site and is typically constructed from the nonstandard ID used in one of the data sources contributing to the IGRA record for that site.

Mobile IGRA stations have station identifiers beginning with ZZV for Volunteer Observing Ships and ZZX in the case of ice islands.

2.1.3 Station Locations

For fixed stations, the coordinates provided in the IGRA station list, igra2-stations.txt, represent the most recent available location of the station. In the IGRA sounding data for fixed stations, the same locations are also provided along with each sounding. Since a sounding-by-sounding accounting of the location of fixed stations is generally not possible, the coordinates provided along with each sounding at those stations are also set equal to the most recent known location of the station. Historical locations and WMO IDs, to the extent that they have been documented, can be found in the IGRA station history information, igra2-metadata.txt (see Section 2.2).

Mobile IGRA stations, whose location is constantly changing, carry a special code in their latitude and longitude fields in the IGRA station list (see file list-format.txt). Their actual positions are shown as part of each sounding in the IGRA sounding data.

Although coordinates are provided with four decimal places, the actual position of the coordinates is typically less. Common precisions are hundreds of degrees, minutes, and seconds. The accuracy of the coordinates is even more variable. Comparisons of locations provided in different sources of information and mapping of the coordinates of certain sites suggest that while most stations are likely to be located within 5 km of the location provided, the accuracy can range from less than 1 minute to 10 km or more.

2.1.4 Methodology

IGRA stations, as well as their station identifiers, names, and locations, were obtained as part of the creation of the IGRA sounding dataset. Initially, a station list is created for each data source. For stations indexed by WMO numbers or six-digit NCEI station numbers, an algorithm has been designed that searches for the most appropriate location and name within a hierarchy of station catalogs maintained by the WMO, NCAR, and NCEI. The names and locations of stations not indexed by WMO number were typically obtained from station information provided along with the data source. As part of further processing, the appearance of station names and other aspects of the station lists created for each source were further standardized, certain station locations and names were corrected, and stations with unresolvable data or metadata problems were removed from possible inclusion in IGRA 2. The sources that were combined to form a particular IGRA station were identified during the station matching process and further refined when certain nearby IGRA stations were composited into a single station. Further details about these steps are provided in Sections 3 and 6.

All in all, a station must meet all of the following requirements in order to be in IGRA:

A. It is required to have a valid location and name or, in the case of a mobile

station, only a valid name.
B. It also has to have a minimum of 28 soundings from at least one of its data sources in at least one year and a minimum total of 100 soundings following the application of all QC procedures.
C. Its data must be substantially different from the data at other IGRA stations.
D. Its data and metadata must not exhibit any major obvious and systematic problems that cannot be resolved.

The IGRA station list is re-created only when the entire dataset is reprocessed. During the daily updates, only the fields for the last year of record and the total number of soundings available are updated.

2.2 Documented Station History Information

2.2.1 General Information

The documented station metadata provided as part of IGRA contains records of instruments, ground equipment, observing and reporting practices, and processing conventions that have been used over time at many of the IGRA stations. Some changes in station locations and station identifiers are also included. All available records, also referred to as "metadata events" or simply "events", are stored in one file and are organized alphabetically by IGRA station ID and chronologically by date within each station. Each record in the "igra-metadata.txt" file contains information for one "event" as identified by the station identification number, date, and event type. Event types are either dynamic (i.e., reporting a change) or static (i.e., reporting practice or equipment in use at the time of the event date). The format of the fields is described in the file igra2-metadata-readme.txt.

2.2.2 Caveats

Users of the documented station history information should keep in mind the following caveats.

A. Radiosonde station history information is generally received in a variety of formats and to varying degrees of specificity. Consequently, the completeness, precision, and accuracy of IGRA metadata records also vary among stations and over time.
B. Although station history information is listed by IGRA station identifier, no attempt has been made to reconcile locations and dates in station history information records with station locations and data availability in the IGRA 2 data. The presence of the station history event in the station history file does not necessarily imply that the data are available in IGRA at the time of the event.
C. Gaffen (1996) aimed to reveal inconsistencies and errors in sources of metadata but not to correct them, and the IGRA metadata are a reflection of this approach.
D. Different station catalogs often disagree about the exact locations of stations, and the number of small location or elevation discrepancies is too

large to hope to resolve. While a large number of discrepancies can be traced to differences in rounding methods, it is probable that many small location errors arise from interpreting a surface observing location as an upper air location or vice versa.

E. The definition of ground equipment is often ambiguous. The ground system and radar may carry separate names, or the entire ground system may be referred to by the name of the radar, or the radar may be referred to by the name of the ground system. In addition, a radiosonde is sometimes referred to by the name of the ground system it uses (such as an AN/GMD-1A radiosonde), implying that any radiosonde compatible with that type of ground equipment may have been used.

F. While every effort has been made to standardize the spelling of event types to conform with documentation, the so-called "before information" and "after information" have not been fully standardized; in particular, records added as part of IGRA 2 processing have not been checked to ensure standardization of the contents of the before and after information fields.

2.2.3 Sources of Information

The documented events originate from four principal efforts to collect station history information which are described below.

2.2.3.1 Gaffen (1996)

The basis of the IGRA station metadata is the digital version of a vast collection of station history information compiled from WMO Member Countries by Gaffen (1993) and constitutes the primary source of information. Records from this source are identified in the IGRA metadata file by an update date of 99/1996. Records associated with IGRA 2 stations that were not available in IGRA 1 further carry an update comment of "from Gaffen (1996) for IGRA 2".

2.2.3.2 Updates for GUAN and RATPAC Stations

Updates to records for many GUAN and RATPAC stations (Free et al. 2005) were collected by NCDC between 2004 and 2007. Most of these records are identified by a source attribution of Hammer (personal communication) and are based on information received through communication between the NCDC GCOS Lead Center and GCOS focal points at HMSCs in WMO member countries. Additional sources of information gathered during this effort were Joyce (personal communication) and NOAA/NWS.

2.2.3.3 Events from Steve Schroeder

Additional events were supplied by Steve Schroeder of Texas A&M University between 2007 and 2009.

Identified by a source attribution of Schroeder (2007), these 2763 events cover stations in the Russian Federation, India, Japan, China (including Hong Kong and Taiwan), and Antarctic stations operated by those countries.

2.2.3.4 Metadata Updates for U.S. and Australian Stations

Additional updates for U.S. and Australian stations were added by Liangying Zhang of NCAR during the summer of 2010. The 475 added or edited records are identified with an update date of 06/2010 and a reference of "Updated by NCAR/ERL". Updates are based on information received by NCAR from the NWS and the Australian Bureau of Meteorology.

2.2.4 Methodology

During the various updates that were made to the metadata for IGRA 1 stations prior to the development of IGRA 2, the original Gaffen (1996) digital file of station history information was reformatted to include "day," "hour," "update date," and "update comment" field, to convert longitude to decimal degrees ranging from -180 to 180, and to increase the lengths of the "before information" and "after information" fields to 40 characters each, and a variety of additional minor format inconsistencies were corrected. In addition, two sets of records that appear to have been inserted automatically but did not add to the value of the station history information were removed. The first set included records identifying a change in a station's geo-political affiliation that carried a year prior to 1900 or whose year was equal to 9999. The second set consisted of all "Station opened" and "Station closed" events attributed to "NOAA NCDC" since they appear to be based on often spurious changes in coordinates. Finally, as information was updated over time, some of the original Gaffen (1996) records were removed if they were deemed unnecessary given previous records or given newly received information.

During the creation of IGRA 2, for IGRA 2 stations that were not represented in IGRA 1, station history information from Gaffen (1996) was added to the IGRA one metadata. During this process, the format of existing and new records was adjusted to allow for the 11-character IGRA 2 station identifiers, eliminate the country code and country code flag fields (since the country code is now contained in the station ID), and change the elevation from an integer in whole meters to a real number in meters to tenths. In new records, the process also adjusted the spelling of event types to comply with the event types listed in the documentation (see igra2-metadata-readme.txt). The records added during this process were assigned an update date of 99/1996 since they originate from Gaffen (1996) and an update comment "from Gaffen (1996) for IGRA 2". This process has been automated in program igra1to2metadata.txt, will be repeated as part of each full reprocessing of the IGRA 2 sounding data, and requires as input the IGRA one metadata file, the original Gaffen (1996) metadata file, and the post-compositing IGRA 2 mingle list (mingle-list-composited.txt, see Section 3).

2.3 Instrument and Equipment Codes

2.3.1 General Information

Radiosonde reports transmitted via the GTS typically include codes for the instrument and equipment used during the sounding. Instrument codes identify the type of radiosonde used, while the equipment code specifies the type of equipment used to track the balloon and is used to report equipment problems. Times series of the codes transmitted in soundings for a particular station therefore can serve as an indication of the instrumentation and equipment used at that station over time. To supplement the documented station history information described in the previous section, IGRA 2 includes files listing the instrument and equipment codes extracted from the so-called "31313 groups" of GTS messages that have been received at NCEI via NCEP since 2000. These files are updated through the end of the previous calendar year whenever the IGRA 2 sounding data are reprocessed, typically annually.

2.3.2 Structure

The instrument and equipment codes are provided in two separate files, each of which contains all available records for all IGRA stations. File "wmo-sonde-history.txt" contains the two-digit radiosonde codes defined in WMO Common Code Table 3685, and "wmo-wndeq-history.txt" contains the equipment codes defined in WMO Common Code Table 0265. Since these tables are updated by the WMO whenever a new code is added, the user is referred to the WMO for the latest versions of these tables.

Each record in a file indicates the period of consistent usage for one code at one station as determined from the available GTS messages. The interpretation of the records is best illustrated by way of an example for Grand Junction, Colorado.

```
USM00072476 2000 01 01 00 2006 09 03 12 52
USM00072476 2006 09 18 12 2011 09 13 00 87
USM00072476 2011 09 13 12 2011 09 13 12 -1
USM00072476 2011 09 14 00 2013 10 30 12 87
USM00072476 2013 10 30 18 2013 10 30 18 82
USM00072476 2013 10 31 00 2013 10 31 12 87
USM00072476 2013 11 01 00 2013 12 31 00 82
```

The first record above shows that instrument 52 (last field) was in use at station USM00072476 (first field) between the 0000 UTC on 1 January 2000 and 1200 UTC on 3 September 2006. This was followed by a gap in the data until 1200 UTC on 18 September 2006, when instrument 87 began to be in use. As shown in the second through fourth records, Instrument 87 remained in use virtually consistently until 1200 UTC on 30 October 2013. The code of "-1" in the last field of the third record shows that the instrument used at 1200 UTC on 13 September 2011 is unknown. Between 1800 UTC on 30 October 2013 and 1200 UTC on 31 October 2013, it appears that both instruments 82 and 87 were in use before the instrument code switched to 82 beginning at 0000 UTC on 1

November 2013.

2.3.3 Caveats

Users who wish to refer to the instrument and equipment codes are advised to note the following caveats:

A. No attempt has been made to check the supplied codes for accuracy or for consistency with other station history information, except that invalid non-numeric codes are not shown.
B. The WMO routinely reassigns radiosonde codes as old instruments are retired and new ones are placed into operation. Code 52, for example, refers to a different instrument in 2013 than in 2000. Reassignments are indicated in WMO Common Code Table 3685.

2.3.4 Methodology

The instrument and equipment codes are extracted from the GTS reports as part of the reformatting of the ncdc-gts data and are then written to files of supplementary information during the further standardization of the source data (Section 3.5). In these files, as in the original GTS reports, there often are multiple records for the same date and time. The files that are distributed as part of IGRA 2 are created from these supplementary files using a program that, depending on the user specified command line argument, processes either instrument totals or equipment codes.

Working on one station at a time, first all of the codes available for that station are read in. If a nonnumeric code or a missing value code is found, the code for that particular station, date, and time is said to be missing. If more than one instrument or equipment code is present for the same station and observation time, a special code of -1 is recorded for that station and observation time.

A station's time series of extracted codes is then analyzed to identify the beginning and end times for each segment during which the same instrument code is reported. Same-code segments may span days on which no valid code was available. However, if a data gap is followed by a code that is different from the one that was in use in the last sounding before the gap, than the endpoint of the previous segment is the last observation time before that gap, and the begin point of the next segment is the first observation time following the gap (see Records 1 and 2 in the example for Grand Junction above).

3. Sounding Data

3.1 Description

3.1.1 Organization of the Data

The IGRA sounding data are organized into station files. Each file is identified by the 11-character IGRA station ID in its filename. For convenience, two types of files are available:

- Full period-of-record (POR) data (files *-data.txt): Each file contains all soundings within a station's full POR. Although these files are updated daily (see below), all data prior to the beginning of the *-y2d.txt files are reprocessed at most once a year.

- Recent data (files *-y2d.txt): Each file contains the station's soundings from the current, or current and previous, year. These files are intended for users who are interested only in the more recent data and do not wish to download the full POR of a station. They also represent the portion of IGRA that is reprocessed, and updated, every day.

Both types of files are updated once a day in the early morning Eastern Time. The latest observations usually become available within two calendar days of when they were taken.

Consider, as an example, the data for Key West, Florida, with IGRA ID USM00072201. As of 20 November 2015, USM00072201-data.txt contained soundings beginning with 2100 UTC on 19 May 1919, the beginning of the IGRA record for the site, and ending with the 1200 UTC sounding on 19 November 2015. On the same date, the corresponding year-to-date file, USM00072201-y2d.txt, consisted of only the subset of soundings extending from 0000 UTC on 1 January 2014 through the end of the record at 1200 UTC on 19 November 2015.

Within each data file, soundings are sorted chronologically from earliest to latest. Each sounding consists of an identification, or header, record, followed by a number of data records corresponding to the number of data levels in the sounding. The header record contains the station identifier, date, time, number of data records, data source codes, and coordinates. Each data record contains fields for a level type indicator, elapsed time since launch, pressure, geopotential height, temperature, RH, dew point depression, wind direction, and wind speed, although frequently not all of these variables are available for a particular data record.

3.1.2 Observation Time

The time of observation of radiosonde and pilot balloon reports in IGRA is expressed in UTC. Since 1958, the WMO has prescribed that soundings be taken around 0000 UTC and 1200 UTC each day. Some stations only observe at one of these two times, while others observe at additional times on a regular

basis or during special weather events, most commonly 0600 UTC or 1800 UTC, or both. Before 1958, typical observation times were 0300 and 1500 UTC, with a greater fraction of reports at atypical times than from 1958 on. The typical observation hours are often referred to as "nominal hours".

Since the ascent of a weather balloon tends to take 1.5 to 2.5 hours, it is common practice to release the balloon from the ground at around 60 minutes prior to the nominal hour. However, the actual release time can be as early as three hours before or even some time after the intended observation time.

The IGRA sounding header records contain fields for both the nominal, or observation, hour and the release time. While some of the IGRA data sources provide both types of times, most do not. As a result, both fields are available in only about 19% of the IGRA soundings. However, the majority of the new soundings from GTS reports in BUFR have both times, which has increased this percentage. In about 74% of the soundings, only the nominal hour is available, and in approximately 6% only the release time is available. When the nominal hour or observation hour is provided, the accompanying date refers to the date of that hour, not the date of release. Thus, a sounding identified with year 2015, month 11, day 01, nominal hour 00, and release time 2309 should be attributed to 0000 UTC on 1 November 2015, even though the radiosonde was launched at 23:09 UTC the day before. When only a release time is present, however, it may be assumed to refer to the time of day on the date identified by the accompanying year, month, and day.

In certain countries, data sources, or periods, the observation hour is simply given as the hour closest to the release time or as the next hour after the release time, regardless of whether that hour is one of the nominal times or not. Thus, a release time of 10:05 UTC may be accompanied by an observation hour of 10, 11, or 12. Consequently, when only an observation hour or only a release time is provided, it is not possible to infer with certainty the missing time. Users interested in identifying all soundings for a particular nominal hour are advised to consider all soundings whose observation hour or release time falls within one or two hours before and after the nominal hour. In cases of frequent observations, such as during special weather events, this approach may yield more than one observation on a particular day.

3.1.3 Variables

The IGRA sounding data include the primary variables of temperature, humidity expressed as either RH or dew point depression or both, and wind direction and speed. The vertical location of a data level within the atmosphere is always identified by pressure, geopotential height, or both. In some cases, elapsed time since launch is also available.

Some of these variables are measured directly by a sensor included in the radiosonde, while others are inferred as part of the initial processing at the ground station receiving the radiosonde signal. Temperature and elapsed time are typically measured, while dew point depression is always inferred from

measured RH, temperature, and pressure. Whether an RH value in IGRA reflects the RH measurement or a value computed from the corresponding reported dew point depression depends on whether the value originated from a GTS report or not. Although RH is measured by a sensor included in the radiosonde, reports sent via the GTS are designed to accommodate only dew point depression as a humidity variable.

Wind direction and speed are determined by tracking the balloon. A variety of tracking methods have been used over time, with GPS becoming an increasingly common tracking method. Traditionally, pressure is measured, and geopotential height has been inferred from pressure, temperature, and humidity. In GPS-tracked radiosonde observations, however, pressure is inferred from the observed height. Whether measured or inferred, all data values are provided in IGRA as received; no estimated values are generated by the IGRA process.

Elapsed time since launch is available primarily in soundings for U.S.-operated stations that originate from non-GTS data sources.

The precision with which data values are reported in IGRA varies with time, data source, station, and with other factors. The reporting precision may be higher or lower than the corresponding instrument's measurement precision. Typical precisions for each variable are listed in Table 3.1.

Table 3.1: Common reporting precisions of data values in IGRA.

Variable	**Common precisions**
Elapsed time	6 s, 1 min
Pressure	0.01, 0.10, 1.00 hPa
Geopotential height	1 or 10 m at pressure levels, often less precise at non-pressure levels
Temperature	0.1, 0.2, or 1.0°C
Dew point depression	0.1, 0.5, or 1.0°C, affected by cutoffs
Relative humidity	0.1% or 1.0%, affected by cutoffs
Wind direction	1, 5, 10, or 22.5°
Wind speed	0.1, 1.0, or 2.0 m/s

In addition, humidity observations have historically been subject to a variety of cutoffs, meaning that they were not reported below certain temperature or humidity thresholds. At many stations over many years, humidity was not reported at temperatures below -40°C. At stations operated by the U.S. between the 1970s and mid-1990s, an RH less than 20% was reported in the data as an RH value of 19% and a dew point depression of 30°C. These and other observing practices affecting humidity reports from radiosondes have been reviewed in a variety of papers (Elliott and Gaffen 1991, 1993; Garand et al. 1992; Wade 1994; Elliott et al. 1998).

3.1.4 Types of Data Levels

There are two groups of data records: pressure levels and non-pressure levels. Any sounding may contain one or both groups. When both groups are present in a sounding, all pressure levels appear first, followed by all non-pressure levels. A sounding must have at least two pressure or two non-pressure levels in order to be included in IGRA, and each group of levels is included in a sounding only if there are at least two levels within the group.

Pressure levels are sorted in order of descending pressure and may contain any combination of temperature, humidity, and wind observations. They consist of the standard pressure levels of 1000, 925, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, and 1 hPa as well as significant levels and other levels whose pressure varies from sounding to sounding.

Non-pressure levels are levels whose vertical coordinate is identified solely by height. The heights used may remain constant from sounding to sounding or vary. They are sorted according to increasing height and contain pilot balloon observations of wind direction and speed.

Either group of levels may include a surface level and one or more tropopause levels. A non-pressure surface level, when available, is included only if there is no surface pressure level present in the same sounding. The presence of multiple tropopause levels is consistent with the WMO definition of tropopause levels.

3.2 Data Sources

3.2.1 Identification

Most IGRA station records consist of data from more than one data source. The data source may vary from sounding to sounding. Any one sounding can have up to two data sources, one for pressure levels and one for non-pressure levels. The data source used for each group of levels is specified by an eight-character code in a sounding's header record.

A total of 33 data source codes are represented in IGRA 2.2, representing a total of 42 data sources. Some of these represent combinations of two or more sets of data that either all were transmitted via the GTS or all originated from the same provider and covered mutually exclusive geographical regions. Providers included NCEI's own archive, the CDMP which was managed by the former NCDC, NWS, NCEP, the USAF, NCAR, NSIDC, ECMWF, IORGC, JAMSTEC, Meteo-France, the ERA-CLIM Project, the IACS at ETH Zurich, and the BAS.

The following table lists basic information about each source, and more detailed provenance and processing information is provided in the Supplement to the IGRA Operations Document.

Table 3.2: Basic provenance information for IGRA data sources.

IGRA 2	Title/	Spatial	Years	Provid	Reference

Data Source Code	Description	Coverage	received	er	
bas-data	"UK READER Upper Air Data"	Antarctica	1948-2008	BAS	none
cdmp-adp	African Daily Pilot Balloon Ascent Sheets	Kenya, Malawi, Niger, Tanzania, Zambia	1946-2004	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
cdmp-amr	African Monthly Radiosonde Forms	Malawi, Zambia	1966-1987	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
cdmp-awc	African wind component data from monthly forms	Malawi and Zambia	1956-1985	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
cdmp-mgr	Malawi WMO-Coded Messages from Computer-Generated Forms Malawi		1984-1991	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
cdmp-us2	U.S. Winds Aloft from Daily Computation Sheets	U.S., some U.S.-operated sites elsewhere	1932-1960	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
cdmp-us3	U.S. Winds Aloft from Military	U.S., some U.S.-operated sites elsewhere	1931-1960	NCEI/ CDMP	Dupigny-Giroux et al. (2007)

	Daily Computation Sheets				
cdmp-usm	U.S. Pilot Balloon Observations from Monthly Forms NCEI/CDMP	U.S., Puerto Rico, Pacific Islands, Canada, some U.S.-operated sites elsewhere	1918-1957		Dupigny-Giroux et al. (2007)
cdmp-zdm	Zambia Daily UAMB Ascent Sheets	Zambia	1960-1990	NCEI/ CDMP	Dupigny-Giroux et al. (2007)
chuan101	Comprehensive Historical Upper Air Network V1.01	Many parts of the world	1902-2007	IACS/ ETH Zurich	Stickler et al. (2010)
erac-hud	ERA-CLIM Historical Upper Air Data	Many parts of the world	1899-1972	http://doi.pangaea.de/10.1594/PANGAEA.821222	Stickler et al. (2014)
iorgc-id	Digitized by IORGC from forms obtained from Indonesian Meteorological and Geophysical Agency	Indonesia	1991-1998	IORGC / JAMSTEC	Okamoto et al. (2003)

mfwa-ptu	West African Temperature-Humidity Soundings	Algeria, Cameroon, Central African Republic, Chad, Gabon, Ivory Coast, Madagascar, Mauritania, Morocco, Niger, Senegal, Tunisia	1948-1965	Meteo-France	None
mfwa-wnd	West African Winds Aloft	West Africa	1940-1958	Meteo-France	None
ncar-ccd	C-Cards Radiosonde Data Set	Global land and ships	1949-1965	NCAR	None
ncar-mit	MIT Global Upper Air Data	Global land and ships	1958-1963	NCAR	None
ncdc6210	Marine upper Air data (NCEI DSI-6210)	Ship tracks and coastal sites across globe	1946-1993	NCEI	Eskridge et al. (1995)
ncdc6301	U.S. Rawinsonde Data (NCEI DSI-6301)	U.S., Canada, U.S.-operated stations elsewhere	1945-present	NCEI	Eskridge et al. (1995)
ncdc6309	NCAR-NMC Upper Air (NCEI DSI-6309)	Global land and ships	1970-1972	NCEI	Eskridge et al. (1995)
ncdc6310	Global	Global	1943-1974		Eskridge et al.

	U/A Cards (NCEI DSI-6310)	land			(1995)
ncdc6314	Russian GTS/Aerostab data	Global land	1997-2010	Roshydromet	None
ncdc6315	People's Republic of China Upper Air Data (NCEI DSI-6315)	China	1948-1990	NCEI	Eskridge et al. (1995)
ncdc6316	Argentina National Data (NCEI DSI-6316)	Argentina	1957-1990	NCEI	Eskridge et al. (1995)
ncdc6319	Korea National Data (NCEI DSI-6319)	South Korea	1984-1992	NCEI	Eskridge et al. (1995)
ncdc6322	Australian GTS data (NCEI DSI-6322)	Southern Hemisphere	1990-1993	NCEI	Eskridge et al. (1995)
ncdc6323	Australian U/A Thermo/Winds Merged (NCEI DSI-6323)	Australia and Australian-operated stations in the Southern Hemisphere	1938-1993	NCEI	Eskridge et al. (1995
ncdc6324	Brazilian Upper Air (NCEI	Brazil	1949-1986	NCEI	Eskridge et al. (1995)

	DSI-6324)				
ncdc6326	Global Upper Air Cards (NCEI DSI-6326)	Global land	1963-1970	NCEI	Eskridge et al. (1995)
ncdc6355	Upper Air Russian Ice Island Data V2.0 (NCEI DSI-6355)	Arctic Ocean	1950-1991	NCAR	Kahl et al. (1999)
ncdc-gts	TAC- and BUFR-formatted GTS reports received by NCEI in near-real-time from NCEP (TAC and BUFR) and ECMWF (BUFR only)	Global land, ships, and moving platforms	2000-present	NCEP/ NCEI	None
ncdc-nws	Data received from the NWS in near-real-time	U.S., Caribbean, Pacific islands	2004-present	NWS/ NCEI	None
nsi-hara	Historical Arctic Radiosonde	Arctic land	1948-1996	NSIDC	Kahl et al. (1992)

	Archive				
usaf-ds3	U.S. Air Force Upper Air Data Set (Ds3 Format)	Global land, ships, and moving platforms	1973-2009	U.S. Air Force 14th Weather Squadron	None

3.2.2 Data Source Selection

The primary aim of creating Version 2 of IGRA was to enhance spatial and temporal coverage, particularly prior to the 1970s. To that end, a comprehensive search for suitable data sources was performed within and outside of NCEI. First, NCEI's archive was checked for the presence of any upper air datasets that were not already included in IGRA 1. Second, information about some additional datasets had been communicated to the authors through personal contacts. Other datasets were found by searching the peer-reviewed literature and reviewing the data holdings at NCAR, a major contributor of data to prior reanalysis efforts. All of these efforts combined yielded an initial list of more than 90 potential source datasets.

The datasets on this list were then further checked for relevance to IGRA, availability, and adequacy of documentation and station location information. Approximately one-third had to be eliminated from consideration because the data turned out to not be available for distribution, because the documentation of the data format was inadequate, because they turned out to contain estimated or modeled values rather than the actual measurements, or because they had been superseded by newer versions of the same data. Another six were set aside when extensive data problems of one kind or another were encountered during the processing. These included the following:

NCEI DSI-6302 (1973-1999) and NCAR DS353.4 (1973-2007, NCEP 1980), consisting of data extracted from GTS reports that had initially been received and processed at NCEP. In these data, the date for 0000 UTC soundings is reported as the date of the release time rather than the date of 0000 UTC.

NCAR TD52 (pilot balloon observations for various parts of Africa and Asia during 1922-1971) and NCAR DS353.0 (radiosonde observations from around the globe during 1962-1972) datasets, which had incorrect level type indicators that NCAR was planning on fixing.

CDMP-digitized, high-resolution data for 1990-1998 at San Cristobal, Galapagos Islands, Ecuador, which did not contain any standard pressure level data, thereby rendering it inconsistent with the types of soundings for which IGRA was designed.

Data for Mexico which had been received on compact disc from Art Douglas,

Creighton University, and turned out to contain primarily data that had originated from the former NCDC and had been processed in unknown ways.

The remaining 53 datasets were grouped into three priority categories. Data were classified as high-priority when they resolved the primary temporal completeness issues affecting IGRA 1, were received from partner programs/agencies, were necessary for replicating data in IGRA 1, and consisted of at least two years of regular (not field campaign) observations. Medium-priority datasets consisted of other datasets of regular observations that had a period of record of at least two years. They included many of the remaining datasets that were used in previous reanalysis projects, all remaining mandatory- and significant-level radiosonde datasets in the NCDC archive that had not previously been quarantined, as well as datasets that enhanced the spatial coverage in certain regions. Low-priority datasets had distribution restrictions, contained fewer than two years of data, or were collected during field campaigns and/or with nonstandard instruments. These datasets could have increased the spatial or temporal density of observations in small regions during a short period of time, but would not have a significant impact on the overall temporal and spatial coverage of IGRA. Considering the often considerable effort that is required for reformatting any one data source, the 13 datasets that fell into this category were not processed for IGRA 2. There then remained a total of 40 datasets.

In the interest of simplicity, some of the sources originating from CDMP were combined since they were sufficiently similar in data format and type and covered mutually exclusive sets of stations. Specifically, five single-country sets of data covering Kenya, Malawi, Niger, Tanzania, and Zambia, had been digitized by CDMP from forms called "daily pilot balloon ascent sheets". These were combined, named CDMP African daily pilot balloon ascent sheets, and assigned the IGRA data source code cdmp-adp. In addition, the CDMP's "Malawi Monthly Radiosonde Data" and Zambia "Monthly UA Temp/Hmr/Additive Sheets (CDMP20ZA)" together comprise the IGRA 2 source that is identified by IGRA 2 source code cdmp-amr. "Malawi Monthly Pibal Data (Form M. O. 34)" and "Zambia Monthly Wind Component Ascent Sheets “MO 34”" form the IGRA 2 source was code cdmp-awc, and "Winds Aloft - Monthly Forms (Phase 1)" and "Winds Aloft - Additional Monthly Forms (Phase 5)" were combined into IGRA source cdmp-usm. Thus, the 40 data sources are identified by 33 IGRA 2 data source codes in IGRA 2 (see Table 3.2).

For IGRA 2.2, two data streams of BUFR GTS reports were incorporated into the NCEI/NCEP GTS source (“ncdc-gts”). One of these data streams comes from NCEP, contains data back to 2017, and is updated daily. The other arrives once a month from ECMWF and contains data going back to 2014 when the transition to BUFR radiosonde reports officially began. Both of these BUFR GTS data sources are archived at NCEI.

3.2.3 Processing

3.2.3.1 Reformatting

To simplify subsequent processing, the datasets are first reformatted into an initial IGRA 2 source format. This format is similar to the format of the final, publicly accessible sounding data files from IGRA 2. Each sounding contains a header, or identification, record followed by a number of data records, each of which represents one level in the sounding. An accompanying station list provides the names, locations, and elevations of the reformatted stations.

However, there are important differences between the format of the data and station list at this stage compared to the final IGRA 2 output: soundings are not chronologically sorted, and data records within a sounding are not sorted according to pressure or height. Stations are identified by 11-character station IDs which are formed from the source station ID, but in most of these IDs, the first two characters are set to "00" since country codes are often not available at this stage, and a slightly different set of network codes is used. Soundings are not chronologically sorted, and data levels within a sounding are also not sorted. The sounding header records contained some instrument and observing system metadata that was later removed. The IGRA 2 level type indicators are not yet used in all data sources. In the station list, there are extra fields for the WMO station number, WBAN number, ICAO call sign, and ship call signs, and these fields are populated to the extent that the information is available.

While rewriting the data from a particular source into this format, soundings were grouped by source station identifier (e.g., WMO station number), data values were converted into a standard set of units, and certain types of obviously defective data records were discarded. Defective records included records whose format or contents did not conform with the accompanying documentation, soundings with nonexistent dates (e.g., April 31) or without a valid observation time, and data values identified as estimated, edited, or interpolated. Some data sources required additional special handling. For example, for some of the U.S. pilot balloon data digitized by CDMP (source cdmp-us2 in Table 3.2), the only available vertical coordinate was elapsed time since balloon launch. These elapsed times were converted to height above sea level using the appropriate NWS standard conversion table as well as the station elevation and balloon weight (30 g or 100 g) that were supplied with the data.

Frequently, the station metadata used originated from the data provider. Many datasets in which stations were indexed by WMO station numbers or ship call signs, however, required reliance on independent station catalogs such as WMO Publication 9 Volume A, the historical station catalogs maintained by the NCAR, or, for moving ships, WMO Publication 47.

3.2.3.2 Standardization

Once both the data and station-identifying metadata had been reformatted, each data source was subjected to an additional set of standardization steps.

These steps are necessary because, in the reformatted data files, codes for instrument types, radiation correction types, cloud information, and level type indicators, when provided, still have the values provided in the original source data; station IDs may not be identified by a network type; soundings may be stored in more than one sounding record, and there may be duplicate soundings; soundings and levels are not necessarily sorted; and there may be duplicate levels. In addition, for data sources that include both land-based stations and ships, there may be more than one station list. The station lists that accompany the reformatted data may include stations whose coordinates or name are not known; may contain inconsistent capitalization of station names; and may not always show a country code, US state code, callsign, or WMO station number.

To remedy all of these issues, an automated algorithm was developed to standardize a reformatted data source. This operation takes all of the following actions:

Combines land stations and ship stations into one station list in which coordinates for fixed stations are in the valid range of -90 to 90 latitude and -180 to 180 longitude, coordinates for mobile stations are set to -98.8888 latitude and -998.8888 longitude, elevations are retained as provided for fixed stations and set to -998.8 for mobile stations, station names are in all capital letters and without any "_" or "," characters, the US state code is set whenever it is reliably available from the input station list's state code field or station name, the call sign is set whenever it is available in the callsign field of the input station list, and the WMO number is set whenever it is available in the WMO number field of the input list or can be reliably inferred from the source station ID.

Skips stations whose coordinates are either missing or outside of the valid range, whose names are blank, whose station ID is equal to some generic ID (e.g., 99999) or was determined to be a temporary or invalid ID, where problems are encountered when reading the input data file, or where the input data file is empty.

Standardizes station IDs by inserting the FIPS country code, if available, into the first two characters of the IGRA source station ID and changing the network code in the third character if necessary.

The following special country codes are used in station IDs:

00 = Undetermined
ZZ = non-stationary

After standardization, the following network codes are valid:

A = U.S. Air Force
M = WMO
R = CARDS/NCDC
S = ship callsign

U = unspecified
W = WBAN

Flag soundings with dates in the future; soundings with invalid months, days, or times (e.g., April 31 or hour 37); soundings that have neither a release time nor a nominal hour; soundings with out-of-valid-range coordinates in sounding headers; and soundings for mobile stations whose coordinates are missing or unreadable in the sounding header. The flagged soundings were not considered for inclusion in IGRA 2.

Changes nominal hours of 24 and release times of 2400 to the corresponding 0000 UTC time for the next day.

Fixes observation hours in data received from the NWS and processed at NCEI (sources ncdc-nws and ncdc6301) that are listed as one hour earlier or one hour later than they should be (e.g., for a release time of 1105, the nominal hour should be 12, not 11.)

Source soundings chronologically according to date, nominal hour, and release time.

For source ncdc-gts only, when a correction report is encountered, flags all other sounding records for the same station/date/nominal hour/message type. The report type (CCA, RRA, RRB, etc.) and message type (e.g., TTAA, TTBB, etc.) are taken from positions 45-47 and 53-56 of the sounding header in the reformatted ncdc-gts data.

Flags levels identified as operator-deleted, descending-balloon, or reascending-balloon; levels with insufficient data; non-pressure levels with temperature or humidity; and the pilot balloon levels in certain sources (ncar-mit, ncdc6315, and ncdc6324) in which the height at those levels appears to be expressed in meters above ground rather than meters above sea level. Insufficient data means that either both pressure and geopotential height are missing or both temperature and wind speed and direction are missing.

- Standardizes level type indicators which identify standard pressure levels, variable pressure levels, non-pressure levels, the surface, and the tropopause.
- In wind-only soundings from the U.S. Air Force dataset that contain levels with both pressure and height, removes geopotential height from standard pressure levels and pressure from levels with non-standard pressures, since these height and pressure values were estimated as part of the Air Force's processing of the data.
- Flags physically impossible dew point depression, RH, wind directions, and wind speed values.
- Eliminates or combines levels with the same pressure as well as non-pressure levels with the same height.
- Eliminates or combines soundings with the same date and time.
- Stores metadata that are routinely included in sounding headers, such as instrument types, types of radiation corrections applied by the

ground system, and clouds/present weather, in one supplementary file per station, without any attempt to standardize the respective codes across data sources. Only the instrument types from source NCDC-GTS were processed further as part of the creation of IGRA 2 station history information (see Section 2.3).

The most complex portion of the standardization process involves the resolution of cases in which multiple soundings are present for the same station, date, and time. In archived data sources, such duplication generally is the result of data entry or reprocessing errors. In GTS messages received in real-time at NCEI, however, various portions of the sounding are received in separate sounding records, such that almost every sounding is comprised of more than one record. In addition, when a balloon ascent terminates prematurely, the operator frequently launches another balloon, sometimes resulting in two soundings being submitted for the same date and time. Except when the second of those soundings is identified as a correction record in source ncdc-gts, no information is available on the reason for duplication on a case-by-case basis. Therefore, an algorithm has been designed that analyzes a group of same-time soundings, determines which soundings should be removed, and combines the remaining soundings into one sounding. Factors considered include the total number of levels and total number of standard pressure levels in each sounding, the similarity of data values, and differences in release time, coordinates, or, for U.S. data originating from CDMP, balloon weight. First, each possible pairing within a group of sounding records for the same observation time is classified into one of three categories: complementary, similar, or different. The soundings complement each other if they contain data for mutually exclusive sets of pressure levels or variables, e.g., if one contains pressure levels, and the other contains non-pressure levels. The soundings are considered to be similar if they have identical release times, positions, and, when applicable, balloon weights, and if they overlap in terms of data levels and variables, and at least 90% of the values in common are similar to each other as defined by the similarity thresholds shown in Table 3.3. If the soundings differ in release time, position, balloon weight, or data (with fewer than 90% of the common values being similar), the soundings are considered to be different.

Table 3.3: Similarity thresholds used when comparing data from two sounding records. Values are considered to be similar if the absolute value of their difference is less than or equal to the threshold listed.

Variable	Threshold
Geopotential Height	10 m
Temperature	0.2°C
RH	10%
Dew point depression	0.5°C
Wind Speed	2 m/s
Wind Direction	10°

In a group consisting of only identical and complementary pairings of soundings, all soundings are combined into one sounding. If, on the other

hand, each sounding in the group is different from all other soundings in the group, the sounding with the largest number of data levels is chosen, or, if there is a tie for the total number of levels, the sounding with the largest number of standard pressure levels. Since many groups contain a mix of identical, complementary, and different pairings of soundings, the actual algorithm for choosing which soundings to combine is a complex iterative process. For the steps in the process, the reader is referred to the subroutine choose_items in a module igra2mod.f95.

The same choose_items algorithm is used to resolve duplicate levels within a sounding, e.g., when a sounding contains two sets of data values at 500 hPa or two sets of wind observations at a geopotential height of 300 m. In this case, the level with the largest number of data values is chosen when the data at such duplicate levels differ. Otherwise, the levels are combined into one level.

One other noteworthy standardization step involves the nominal hours assigned to soundings received from the NWS, including those in IGRA data sources ncdc-nws and ncdc6301. In these data sources, the nominal hour of soundings taken around the time of a synoptic hour (0000, 0600, 1200, or 1800 UTC) is frequently listed as one hour earlier (2300, 0500, 1100, 1700) than it should be when the release time is between 31 and 60 minutes before the synoptic hour. These nominal hours are corrected on the basis of the release time that is provided along with the data. For example, for a release time of 1105, the nominal hour should be 12, not 11.

3.2.3.3 Format Checking

A program was developed that checks the format of both the reformatted and standardized data as well as the accompanying station lists. The purpose of this program is to verify that the output from the reformatting and standardization steps conforms with the intended output format as well as with expectations that are based on what is known about a particular data source. This program was executed on the reformatted and standardized versions of all data sources, and its output messages were examined. If issues were found that could be traced to a false assumption about the input data or an error in the reformatting or standardization program, appropriate action was taken to remedy the problem, and the format checker was rerun. These steps were repeated until either no more problems were found or all of the messages written by the format checker could be traced to errors in the original data that could not be addressed at that stage.

3.3 Station Matching

3.3.1 Introduction

Station matching is the term used here to refer to the process of identifying the unique observing sites that are represented within all of the IGRA data sources and the source station whose data should be combined to form the data record for each of those sites. This process is complicated by the use of several

different station identification systems across all data sources, by variations in the precision and accuracy of coordinates and the spelling of station names among the sources, and by the absence of a data overlap between many pairs of source stations. For this reason, a multi-step decision-making algorithm was designed and implemented that utilizes multiple pieces of information about the data and metadata.

In general, the process consists of the following steps:

A. Preparation of needed information, including inventories of the number of soundings available for each station and year in each data source; comparison of data between all possible pairs of source stations with overlapping records, and creation of a list of all source stations that should be considered during the station matching process.
B. Identification of matching pairs of stations on the basis of data similarity, station identifiers, distance, and station names.
C. Creation of "mingle groups" from the identified pairs, each of which provides the ingredients of the data record for one IGRA 2 station in terms of a list of one or more source station identifiers.

Throughout this process, each source station is identified by a 19-character string consisting of the eight-character IGRA data source code (see Table 3.2) followed by the 11-character station ID created during the standardization process. The combination of the data source code with the station ID ensures that each station in each data source receives a unique identifier.

3.3.2 Preparation

3.3.2.1 Data Comparison

In preparation for the calculation of data comparison statistics, the standardized source data were reformatted into a synoptic-sort format suitable for performing the subsequent data comparison. When reformatted in this way, the data are stored in up to eight files per day, one for every third hour between 0000 and 2100 UTC. Each file contains all available soundings that have at least two data levels and are between two hours before and three hours after the date and time specified in the file name. The observation hour, when available, is used to identify the time of the sounding; when the observation hour is missing, the release hour is used. The minimum of two data levels is imposed to facilitate more robust data comparison results than would be possible with only one level.

The data comparison then worked as follows. For each station pair, concurrent and nearly-concurrent soundings were compared on the basis of both standard pressure levels and non-pressure levels, and the sounding-by-sounding comparison results were aggregated into annual counts of the number of matches and the total number of soundings compared. A match between two soundings was declared when one of the following conditions was met:

A. 10 or more pressure-level values were compared and at least 75% of

them were within the similarity thresholds shown in table 3.3;
B. Ten or more values were compared at height-only wind levels, and at least 75% of them were within the defined similarity thresholds;
C. Fewer than 10 pressure-level values were compared, all of them were found to be similar, and no non-pressure-level values could be compared; or
D. Fewer than 10 values were compared at height-only wind levels, all of them were similar to each other, and no comparison was possible at pressure levels.

These conditions were established after an extensive evaluation of various thresholds and conditions. They are designed to maximize the possibility of identifying segments of data that originate from the same observation site to the extent possible while minimizing the chance of inadvertently matching up data from nearby, but distinct, sites.

The reason for comparing not only concurrent, but also close-in-time, soundings lies in the fact that there are slight variations in how observation time is recorded in different data sources (see Section 3.1.2). The comparison with non-concurrent soundings was accomplished as follows. Each sounding was first assigned a reference synoptic hour, chosen from one of the eight hours of 0000, 0300, 0600, 0900, 1200, 1500, 1800, and 2100 UTC. The reference synoptic hour was defined as the synoptic hour between the sounding's observation hour and three hours after the sounding's observation hour. The sounding was then compared to all soundings between five hours before and three hours after that synoptic hour. For example, a sounding with an observation time of 1000, 1100, or 1200 UTC was compared to soundings between 0700 and 1500 UTC. This ensured that soundings from the same synoptic hour were compared, even if their reported observation hours differed due to different reporting or processing practices.

Once the annual counts of the number of matching sounding pairs and the number of soundings had been computed for every possible station pair, the results for each station pair were summarized in terms of an average percent similarity. Specifically, the percent similarity was calculated by averaging the results from all years in which at least 28 soundings could be compared. The resulting percentages were organized into one file per station, with each record in a station's file showing the percent similarity with one other station.

3.3.2.2 Year-by-Year Inventories

In preparation for further processing, an inventory was created for each standardized data source that provides, in one file per data source, the total number of soundings available for each station and year. Any sounding was counted that appeared in the standardized source data. One inventory file is created for each data source, with each record providing the number of soundings for one source station ID and year.

3.3.2.3 Consolidated Source Station List

The next step in the process is to combine the individual lists of source stations that were created during the standardization process into one consolidated list containing the data source code+standardized source station ID, latitude, longitude, elevation, station name, call letters, and WMO number.

During this process, additional measures were taken to apply corrections to coordinates, names, or WMO IDs as needed and to weed out stations with invalid station IDs as well as stations with poor-quality metadata or data that cannot be corrected.

First, as the station records from all standardized source station lists were collected, the following types of stations were removed:

3082 stations where no year had more than 27 soundings;
2846 CHUAN stations from CDMP, Meteo-France (French West Africa only), and NCAR TD52/53 which exhibited various data or metadata problems;
134 NCEI DSI-6301 stations, ending before 1960, whose data exhibited a one-sounding lag relative to other data sources with realistic-looking diurnal cycles;
109 stations with other unresolvable data or metadata problems revealed by the station matching process;
42 moving ships from erac-hud (original station IDs < 1000) that have few soundings and many position errors;
38 stations that were known test sites or where the station ID, name, or coordinates were invalid; and
20 stations with elevation problems revealed during the analysis of monthly median elevations (see Section 3.4.3 below).

Next, some station names were edited to further standardize them in preparation for station name matching, and WMO numbers ending in 999 were set to blank. Specifically, the following changes were made to station names:

A. When a name ended in an asterisk, apostrophe, or ampersand, the last character was removed.
B. The names of Ocean Weather Ships were standardized to "OCEAN WEATHER SHIP" followed by the single letter identifying the ship.
C. Words identifying the regions Antarctica, Arizona, Utah, and West Indies were removed from names, and the corresponding state or country code was updated if needed.
D. "AIR TERMINAL" was abbreviated as "AIRTERM" and "PROVING GROUND" as "PRVGR".
E. The spelling of some names was corrected, including, for example, "GRIEFSWALD" to "GREIFSWALD" and "TALLAHASEE" to "TALLAHASSEE".

All of these changes were programmed rather than implemented manually, so that these issues can be addressed automatically when a new data source with the same types of problems is added to IGRA.

Some additional station-specific corrections to coordinates, elevation, station name, call letters, or WMO numbers were made on the basis of specially-designed input files. In general, the need for these corrections was identified during the initial station matching process, and corrections were made on the basis of metadata found in the station lists of other data sources, Internet searches, and NCEI's archive of station history information. Corrections were provided to the program applying them in one of four ways:

A. In the form of entire station list records in which any needed corrections have been made to one or more fields (file record-fixes.txt). The program replaces these records for the corresponding records found in the standardized source station lists.
B. In a file showing mismatches between the coordinates of source stations with the same WMO number (wmo-fixes.txt). This file shows an early version of matching statistics between pairs of source stations with the same WMO number and identifies the source station that appears to have the correct coordinates. The program uses these coordinates for all source stations with that WMO number.
C. In a file showing mismatches between the coordinates of source stations with the same call letters (icao-fixes.txt). This file shows an early version of matching statistics between pairs of source stations with the same call letters and identifies the source station that appears to have the correct coordinates. The program uses these coordinates for all source stations with those call letters.
D. In a file showing mismatches between the coordinates of source stations with the same WBAN number (wban-fixes.txt). This file shows an early version of matching statistics between pairs of source stations with the same WBAN number and identifies the source station that appears to have the correct coordinates. The program uses these coordinates for all source stations with that WBAN number.

3.3.3 Identification of Matching Station Pairs

3.3.3.1 Overview

As a first step in the actual matching process, each possible pair of source stations was first classified according to the degree to which the data and metadata of the two member stations match with each other. Four criteria were used: data similarity, shared station identifiers, distance, and station name similarity. These criteria are summarized in Table 3.4 and described in greater detail in Section 3.3.3.2.

Table 3.4: Matching criteria used to identify matching pairs of source stations.

Criterion	Condition for good match	Condition for mediocre match	Condition for mismatch	Condition for inconclusiveness

Percent data similarity	>=75%	>=50% and <75%	<50%	Insufficient overlap
Station ID	WMO IDs identical; or WMO IDs not compared, and WBAN, ICAO, or ship call signs identical	n/a	WMO IDs differ; or WMO IDs compared, and WBAN, ICAO, or ship call signs differ	No IDs of the same type available for both stations
Distance	<=10 km	>10 km and <=40 km	>40 km or one station is fixed and one is mobile	Both stations are mobile
Station name	Names similar and distance <=40 km or undefined	n/a	n/a	Distance >=40 km

Each station pair was classified on the basis of the four criteria into one of the seven categories listed in Table 3.5. For each source station, all pairings that were classified as anything other than "SEPARATE" were written to a station-specific output file for use in subsequent processing.

Table 3.5: Classification of station pairs.

Category	**Description**
MATCHED	Condition for a good match in data, station ID, or name is met, and none of the four criteria indicates a mismatch.
SOSOMATCH	Data or distance indicate a mediocre match, at least one other criterion indicates a mediocre or good match, and no criterion indicates a mismatch.
UNKNOWN	No criterion indicates A good match, and none indicates a mismatch; includes cases in which either distance or data, but not both, indicate a mediocre match, and other criteria are inconclusive.
NEAR	Distance indicates a good match, and data indicate a mediocre match; no other criterion indicates a match.
CONFLICT	Data indicate at least a mediocre match, and distance indicates a mismatch; or station ID indicates a match, and data or distance indicate a mismatch; or name indicates a match, and data indicate a mismatch.
SEPARATE	Pair does not fall into any of the above categories.

In addition to the classification of station pairs, each individual source station was assigned a flag that describes the quality of the pairings associated with that station. These flags, listed in Table 3.6, were particularly useful in identifying stations where the matching criteria yielded conflicting results.

Table 3.6: Flags describing the overall quality of a source station's pairings

Flag	Meaning	Condition
A	All pairings have conflicts	All pairings yield a conflict between matching criteria.
C	Conflict with at least one station	There is at least one pairing with a conflict between matching criteria, and the station is not already flagged with A, D, d, or c.
C	Conflict are minor data mismatches	All conflicts are due to data mismatches during short overlaps (i.e., total overlap of less than 20% of PO r) with ID- or name-matching stations.
D	Pervasive data mismatches	With at least half of POR overlapping with at least one station that has the same name or station ID, no good data match is found with any of those stations and an outright data mismatch is identified with at least one of them.
D	Less pervasive data mismatches	With at least half of POR overlapping with at least one station that has the same name or station ID, a Data mismatch is found in every year for which data are compared with those stations, but there's also at least one data match.
L	Location conflicts	All conflicts occur when distance exceeds 40 km, and data, IDs, or names match.
M	Matching	All pairings are based on only good matches.
M	Matching despite minor location discrepancies	All mediocre matches are caused by a minor coordinate mismatch when at least two other criteria match.
N	Near	Station has no matches or conflicts with any other station, but is within 10 km of another station.
S	Separate	Default; none of the other conditions are true
U	Undetermined	Station has no matches or conflicts with any other station and lies within 10-40 km of another station.

X	Also matching	All pairings are based on good and mediocre matches, and station is not classified as "m".

3.3.3.2 Rationale

The use of multiple matching criteria was necessary because no single criterion would be capable of identifying all of the matches. For example, out of the 27,620 pairings with a good station ID, name, or data match (Table 3.5), 84% could be identified by a station ID match alone, 53% by only a name match, and 35% by only a data match. Conversely, if the station ID match were not employed, 30% of the matches would not be detected; without the name match, 10% would not be found, and without the data match, 4% would not be identified. In addition, the data criterion helped to identify cases in which the station ID or name matched, yet the data did not. The distance criterion was useful not only for defining the radius within which station names can be matched, but also for the identification of cases in which the data, station IDs, or names match, yet the stations were far apart.

3.3.3.3 Matching Criteria

3.3.3.3.1 Data Match

Using the average percentages of similarity described in Section 3.3.2.1, the data of two stations were considered to match well if the corresponding percent similarity was at least 75%. A mediocre match was declared when the percent similarity was at least 50%, but less than 75%. A percent similarity less than 50% resulted in the declaration of a data mismatch. In the absence of sufficient overlapping data to compare, neither a data match nor a data mismatch was declared.

The above thresholds were determined on the basis of a systematic evaluation of samples of station pairs within different ranges of similarity percentages (Table 3.7), using the method of Durre et al. (2008). In choosing the thresholds, not only the false-positive rate was considered, but also the types of matches that would be allowed if a particular threshold were chosen.

Table 3.7: Evaluation of the data match criterion.

Data match rate	# evaluated	# belonging together	Total pairs	Estimated good pairs	Cumulative false-positive rate
>=90%	10	9.5	25880	24586.00	5.00%
>=80% & <90%	10	10	2366	2366.00	4.60%
>=70% & <80%	10	9.5	1709	1623.55	4.61%
>=60% & <70%	10	8	1192	953.60	5.19%
>=50% & <60%	10	7	1159	811.30	6.08%
>=40% & <50%	10	4	654	261.60	7.15%

>=30% & <40%	10	1	1716	171.6	11.25%
>=25% & <30%	10	0	1827	0.00	15.70%
>=20% & <25%	10	1	3281	328.1	21.82%

3.3.3.3.2 Station ID Match

Various identification systems are used across the different IGRA data sources (see Sections 2.1, 3.2, and 6). While in most sources, stations are identified by WMO numbers, others use WBAN numbers, standard ship call signs, consecutive numbering, or an unidentified numbering scheme. In some cases, the metadata provided along with the particular data source included cross-references to one or more standard station identification systems. For example, the station list for the Comprehensive Historical Upper Air Network (CHUAN), in which stations are numbered consecutively, contained the WMO numbers that corresponded to some of the stations. In addition, the U.S. Air Force station catalog that accompanied source usaf-ds3 as well as that CARDS station list accompanying many of the NCEI data sources, which are based on two sometimes differing six-digit numbering schemes, both contained ICAO call letters.

Therefore, when comparing the station identifiers of two source stations, four types of standard IDs were considered: WMO numbers, WBAN numbers, ICAO call letters, and ship call signs. A station ID comparison was possible only if at least one of these types of IDs was associated with both stations. A station ID match was declared when either of the following two conditions was true:

A. Both stations had a WMO ID, and those WMO IDs were identical to each other; or
B. One or both of the stations were not associated with a WMO ID, there was a match for one of the other types of IDs, and none of the remaining types of IDs differed from each other.

A mismatch was declared when one of the following two conditions was true:

A. Both stations had a WMO ID, and those WMO IDs were not identical to each other; or
B. One or both of the stations were not associated with a WMO ID, and the stations were associated with two different WBAN numbers, ICAO call letters, or ship call signs.

The comparison of call signs for fixed stations required some special considerations. Call letters for fixed stations were compared only if they consisted of three or four letters. ICAO call signs typically are four-letter strings. However, the national call signs of U.S. stations contain three letters whose ICAO equivalent is formed by appending a "K" to the front of those three-letter groups. Similarly, the ICAO call signs for Canadian stations begin with the letter C and contain the three-letter Canadian call sign in characters 2-4. Since some of the sources of station metadata displayed the national rather than the ICAO call letters for U.S. and Canadian stations, and these were easily converted to ICAO call signs as described above, the call signs of two fixed

stations were not only considered to match whenever they were identical, but also when one consisted of four letters and began with either C or K, and the other consisted of three letters that were identical to the characters 2-4 of the four-letter call sign.

All conditions that are incorporated into the station ID matching criterion were evaluated and refined using the strategies outlined in Durre et al. (2008). The results of these evaluations are summarized in Table 3.8.

Table 3.8: Evaluation results for station ID matching conditions.

Condition	Cases evaluated	Good matches	Notes
WMO IDs identical	10	10	
ICAO call signs identical	10	9.5	One match was questionable
Call signs matches letters 2-n of n-letter callsign beginning with K	10	10	
Call signs matches letters 2-n of n-letter callsign beginning with C	10	9	Bad match when call signs contained three and two letters
Call signs matches letters 2-n of n-letter callsign beginning with P	10	8	Bad matches when stations far apart
Ship callsigns identical	10	10	
IDs of one type match, IDs of another type differ	10	9	Bad match involves at least one ship
WBAN IDs identical	10	10	

The evaluation process is best illustrated by way of an example. When assessing the value of matching stations on the basis of their ICAO call letters, 10 cases were randomly chosen from the approximately 12,000 station pairs with matching call letters, and the metadata and data similarity percentage of each of those station pairs were examined in order to assess the fidelity of the match. In addition, station pairs where the WMO numbers, WBAN numbers, or data matched, yet the ICAO signs differed were inspected to determine the possible causes for the mismatches. It is through these inspections that the

ICAO callsign matching criterion was refined to incorporate three-letter call signs and the special conditions for ICAO signs beginning with "C" or "K". These refinements were, in turn, again evaluated by choosing 10 cases in which a call sign began with "K" and 10 in which the call sign began with "C". Both conditions were found to be appropriate. The same condition was also tested for call signs beginning with "P", but in that case, two of the 10 matches evaluated were found to be inappropriate, and all of the eight good matches were also identified by matches in WMO numbers. Therefore, no special condition was implemented for call signs beginning with "P".

3.3.3.3.3 Distance Match

The distance between two stations was determined by calculating the great circle distance between the geographical coordinates provided in the consolidated source station list. A distance in excess of 40 km was considered to be indicative of a mismatch, while a distance within 40 km, but greater than 10 km, was classified as a mediocre match. Although a distance of less than 10 km was considered to represent a good match, a good distance match alone proved to be insufficient grounds for combining two source stations. Rather, a distance less than 10 km was used as supporting information when the data criterion indicated a mediocre match, as can be seen from the list of pair classifications in Section 3.3.3.1. The primary reason for not combining stations on the basis of distance alone was the occasional coexistence of civilian and military stations in close proximity to each other. This was the case, for example, in the Honolulu area during the 1940s.

3.3.3.3.4 Name Match

The process of identifying whether two station names are the same is complicated by many factors which require consideration.

First, cities and towns with the same name exist in multiple locations. For example, South Carolina and Missouri are two states with the city named Columbia. For this reason, name matches were considered to be indicative of a station match only if the stations were located within 40 km of each other. Second, names may be spelled differently in different sources of station metadata. Reasons for spelling differences include typographical errors (e.g., Tallahasee rather than Tallahassee), the existence of multiple accepted spellings over time, different spellings in English versus in the language of the country where the city is located (e.g., Munich and Munchen), differences in punctuation or spacing (e.g., St. Louis and St Louis), and the use of different abbreviations for the same term (e.g., AP or ARPT for airport). Consequently, an algorithm was devised that identifies a match between station names in the following five ways:

A. Exact String match: Names contain the same number of characters, and there is a character-for-character match between the two strings.
B. Exact alphanumeric match: When spaces and punctuation signs are removed, the resulting sequences of letters and numbers are of the same length, and there is a character-for-character match between them (e.g.,

"Amundsen Scott" matches "Amundsen-Scott").

C. Exact word match: The two names have the same number of words, and after some standardization of common abbreviations and phrases, all of the words match exactly (e.g., "SAINT PAUL ISLAND" matches "ST PAUL ISLAND" and "SEATTLE-TACOMA AIRPORT" matches "SEATTLE TACOMA AP"). The characters " ", "/", "-", ".", ",", "(", ")", and "&" are assumed to be delimiters between words. Any resulting word that represents an abbreviation or foreign expression for airport, fort, international, island, mount, municipal, national, point, regional, saint, station, or upper-air is replaced with the spelled out English equivalent, and the word "MILITARY" is substituted for all abbreviations designating military bases. In addition, the phrase "INTERNATIONAL AIR" at the end of a name is replaced with "INTERNATIONAL AIRPORT", and when a name ends in the word INTERNATIONAL, MUNICIPAL, or REGIONAL, the word AIRPORT is added.
D. Approximate match: The names have the same number of words, each of those words contains at least five letters and no numbers, and, after standardization of abbreviations and phrases, all words either match exactly or differ by one letter (e.g., "SEWASTOPOL" matches "SEBASTOPOL", "MIRNYJ matches MIRNY", "CANTON ISLAND" matches "KANTON ISLAND", and "NEWBERGH STEWART AFB" MATCHES "NEWBURGH/STEWART AFB").
E. Consonant match: The names have the same number of words, all of those words contain at least five letters and no numbers, and, after standardization, all words match exactly, differ by one letter, or have the same sequence of consonants (e.g., "BENGALORE" MATCHES "BENGALURU" and "EKATERINO-NIKOL'SKOE" matches "YEKATERINO-NIKOLSK"). Words matching on the basis of consonants only each must have at least four consonants.

For names beginning with the word "SHIP", the word-by-word criteria could result in inappropriate matches. For this reason, only the exact and alphanumeric criteria were applied if either name started with the word "SHIP" followed by a space.

Approximately 90% of all name matches were obtained through exact string, alphanumeric, or word matches, while the other 10% came from the approximate and consonant matches. On the basis of systematic evaluations of samples of 10 cases for each of the above five matching conditions, the false-positive rate for each condition was near 0%.

All of the edits described above were made only for the purpose of name comparisons; they were not recorded in the IGRA station list.

3.3.3.4 Example

As an example, consider the seven source stations associated with Amundsen-Scott Air Force Base at the South Pole. Their metadata are shown in Table 3.9:

Table 3.9: Sample set of source stations that are matched with each other on the basis of the pair-finding results.

Source	Station ID	Latitude	Longitude	Station Name	POR
ncdc-gts	00M00089009	-90.0000	0.0000	AMUNDSEN-SCOTT	2000-2013
bas-data	AYM00089009	-90.0000	0.0000	AMUNDSEN SCOTT	1961-2010
usaf-ds3	AYA00890090	-90.0000	0.0000	AMUNDSEN-SCOTT	1981-2008
ncdc6301	00W00090001	-90.0000	0.0000	AMUNDSEN-SCOTT	1961-1990
ncdc6309	00R00890090	-90.0000	0.0000	AMUNDSEN-SCOTT	1971-1972
ncdc6326	00R00890090	-90.0000	0.0000	AMUNDSEN-SCOTT	1967-1970
ncdc6310	00R00890090	-90.0000	0.0000	AMUNDSEN-SCOTT	1961-1961

The pair-finding results for these source stations are shown in Table 3.10. Note that different combinations of criteria can lead to a match, and that there is a conflict in matching criteria between sources bas-data and ncdc6309.

Table 3.10: Pairings of the source stations whose metadata are shown in Table 3.9.

Source1	Source2	Overall match	Data similarity (%)	ID match	Name match	Distance (km)
ncdc-gts	bas-data	SOSOMATCH	69.5	WMO	Alphanumeric	0.000
ncdc-gts	usaf-ds3	N/a	Yes	Exact	0.000	
ncdc-gts	ncdc6301	MATCHED	N/a	N/a	Exact	0.000
ncdc-gts	ncdc6309	N/a	Yes	Exact	0.000	
ncdc-gts	ncdc6326	N/a	Yes	Exact	0.000	
ncdc-gts	ncdc6310	N/a	Yes	Exact	0.000	
bas-data	usaf-ds3	MATCHED	75.2	WMO	Alphanumeric	0.000
bas-data	ncdc6301	MATCHED	97.2	N/a	aAlphanumeric	0.000
bas-data	ncdc6309	CONFLICT	43.6	WMO	Alphanumeric	0.000
bas-data	ncdc6326	MATCHED	75.3	WMO	Alphanumeric	0.000
bas-data	ncdc6310	MATCHED	100.0	WMO	Alphanumeric	0.000

usaf-ds3	ncdc6301	85.7	N/a	Exact	0.000	
usaf-ds3	ncdc6309	N/a	Yes	Exact	0.000	
usaf-ds3	ncdc6326	N/a	Yes	Exact	0.000	
usaf-ds3	ncdc6310	N/a	Yes	Exact	0.000	
ncdc6301	ncdc6309	SOSOMATCH	65.3	N/a	Exact	0.000
ncdc6301	ncdc6326	MATCHED	80.7	Exact	0.000	
ncdc6301	ncdc6310	MATCHED	100.0	Exact	0.000	
ncdc6309	ncdc6326	N/a	Yes	Exact	0.000	
ncdc6309	ncdc6310	N/a	Yes	Exact	0.000	
ncdc6326	ncdc6310	N/a	Yes	Exact	0.000	

3.3.4 Identification of Mingle Groups

In this processing step, a decision-making algorithm was applied to the station pair classification results in order to identify groups of source stations that can be combined into IGRA stations. Following is a description of the steps involved:

A. Creation of first mingle group: For the first source station, all of the stations were read in for which the pair finding program found a match with the target station based on at least one criterion as well as the stations likewise paired with those stations and so on. The resulting collection of stations formed the initial version of the first mingle group.
B. Creation of other mingle groups: Proceeding through the list of all source stations, step #1 was repeated for each station that was not already part of a mingle group. At the end of this process, each source station either was included in a mingle group containing two or more members or was identified as a station that was to be added to the dataset by itself.
C. Elimination of conflicts and transitivity violations within mingle groups: As described in Section 3.3.3.1, a “conflict” exists between two stations when one or more matching criteria indicate that they represent the same observational record, while either the data or coordinate comparison suggest that they are distinct records (e.g., they have identical names and coordinates as well as overlapping records, but different data during that overlap). In the context of the members of a mingle group, a transitivity violation exists if station A matches with station B, and station B matches station C, but stations A and C do not match at all or have a conflict. Both of these situations compromise the integrity of a group. Therefore, each initially created mingle group needed to be checked for such conflicts and violations. This was accomplished by analyzing a comparison matrix in which each position represented one of all possible source station pairs, And each value (i,j) was an indicator of the quality of the match between stations i and j as determined from the pair finding results. The quality of the match was identified numerically as follows:

2 = Good match (pair classification MATCHED in Table 3.5)
1 = mediocre match (pair classification SOSOMATCH)
0 = Undetermined (including pair classifications UNKNOWN and NEAR)

-1 = Separate
-2 = Conflict between matching criteria

If a matrix contained a -2 (conflict), some or all of the offending source stations were eliminated in an iterative process that continued until either no conflict or no member stations remained in the group. In each iteration, a station was removed from the group if, in any iteration, it met one of the following conditions: (a) it had the single largest number of conflicts among the remaining members of the group; (b) it had the single largest number of transitivity violations among members with the largest number of conflicts in the group; (c) it had the fewest number of soundings among stations tied for the largest number of conflicts and transitivity violations; or (d) it had the fewest number of good matches (indicated by 2s in its matrix row and column) among members with the largest numbers of conflicts and transitivity violations and the fewest number of observations. In each iteration, these conditions were checked in the order listed. Only condition (d) can result in the removal of more than one station during a single iteration. If transitivity violations remained after all conflicts had been removed from a group, an analogous iterative process was performed to remove these violations, this time treating each transitivity violation as a conflict. After all conflicts and violations had been removed,

Once all conflicts had been removed, the matrix was checked to verify that all of the group's remaining members were interconnected. If not, the group was split into two or more groups accordingly. For example, if a group contained five members, and two of its members matched with each other but not with any of the other group members, then those two stations were split off into a separate new two-member mingle group.

For example, the mingle group for the city of Karonga, Malawi, consisted of station records from three sources (cdmp-adp, cdmp-awc, usaf-ds3), all of which matched with each other because their WMO station IDs (67587), coordinates (9.95°S, 33.883°E), and names (KARONGA) were identical, and a data comparison was not possible among any of the three possible station pairs and therefore did not provide any additional information on the quality of the matches. Thus, there were no transitivity violations in this example.

3.4 Data Integration

Each of the mingle groups created in Section 3.3 listed the "ingredients" for one IGRA station record by identifying the source station IDs whose data were to be merged. In the data integration step, this merging of the data from multiple sources was performed, and each merged record was assigned an IGRA station ID. Since the data records to be merged often overlap, a simple combination of the source station records would have resulted in the presence of multiple soundings for the same observation time. The detection of such sounding duplication was complicated by intersource differences in how the observation time was reported. A somewhat elaborate data integration algorithm therefore was required in order to create an integrated data record in which there was only one sounding record for each unique observation. Following is a description of the steps taken by this algorithm.

3.4.1 Setup

First, the data from each source specified in the mingle group were read in, and only those soundings were stored that meet certain data requirements. A sounding was retained if

A. It contained either at least two mandatory-pressure levels with either temperature or wind between the surface and 100 hPa or at least two height-only wind levels, and
B. It was from a year in which at least 28 soundings from the same source met requirement (a).

Next, release times that had been flagged during the data source standardization process were set to missing. In the data through 2013, 68.4 million source soundings were stored, and the release time was set to missing in 364 of them.

For each sounding, the time difference between the reported observation date and time and the reference time of 0000 UTC on January 1 of the first year with data was then computed. The observation hour was used to determine the time difference whenever it was available. If no observation hour was available, the release time was used. The time difference was stored in minutes since the reference time.

Using these time differences, the soundings were sorted in chronological order. Multiple soundings that had the same time difference to the reference time were sorted by release time, with soundings without a release appearing after all non-missing release times. All subsequent steps assume that the soundings are in chronological order.

3.4.2 Assignment of IGRA Station ID

Next, the 11-character IGRA station ID (see Section 2.2.1) was determined, and the associated metadata were collected. The station name, coordinates, and elevation were taken from the station list entry of the source station with the most recent period of record. The FIPS country code that forms the first two characters of the IGRA station ID was determined from one of four sources: one of several hardwired “if” statements which override any other source of information; the country code provided in one of the sources of station metadata (see Section 2.2.4); the country code retrieved from www.google.com using the coordinates provided in the source station's metadata; or the country code returned by ws.geonames.org on the basis of the same coordinates.

Whenever possible, the IGRA station ID was based on one of the standard station IDs - ship call sign, WMO number, WBAN number, and ICAO callsign, in that order of preference. If a ship callsign was available for any of the source stations that contributed to a particular IGRA station, then the most recent of these call signs formed the basis of the IGRA station ID; if no ship call sign was

found, the most recent WMO number available from the source stations was used, and so on. If none of the four standard station IDs was associated with the IGRA station record, then a custom station ID was constructed from the source station ID. This custom ID has a network code of X and is followed by the characters UA and a 6-character string that identifies the type of custom ID used and contains the original source station ID.

For example, the most recent WMO number for Key West, Florida, is 72201. Therefore, the corresponding IGRA station ID is USM00072201. On the other hand, chuan101 station 3671 was not combined with any other source station, and, therefore, its IGRA station ID is based on the chuan101 ID and takes the form USXUAC03671.

3.4.3 Check for Similar Consecutive Soundings

In this step, soundings that were between 21 and 24 hours apart and whose data matched with a similarity percentage of at least 90% at pressure levels and at non-pressure levels were removed in an iterative process that was designed to minimize collateral damage. Each sounding was compared to all soundings that preceded it by between 21 and 24 hours. A data similarity percentage was computed separately for pressure levels and for non-pressure levels. For each of the two groups of levels, the similarity percentage was calculated only if at least two levels with a total of at least 10 values were available. If the calculated percentage or percentages equaled or exceeded 90%, the soundings were considered to match.

Moving through the entire record at a particular station, a count of the number of matches found was tracked for each sounding. If there were soundings with more than one match, they were removed, and all remaining soundings were re-compared in the same manner. If, then, the maximum number of matches throughout the entire record was 1, all soundings with at least one match were removed. The process ended for that station when there were no more matches throughout the entire record or no more soundings remained.

For example, if a 0000 UTC sounding on October 2, 2015, were identical to the 3 October 2015 0000 UTC soundings from two other sources, then the 2 October sounding would be removed because it matched with two soundings, while the 3 October soundings each matched with only one. If, on the other hand, the 2 October sounding matched only with one 3 October sounding, both of these soundings would be removed.

The timeframe of 21-24 hours was chosen after careful and systematic evaluation of various time frames because duplication is rare for soundings that are more than 24 hours apart, and data duplication among soundings that are significantly closer in time can represent true meteorological similarity rather than a data duplication problem.

In the data ending in 2013, Approximately 70,000 soundings were removed with this step. The approach proved to be particularly useful for identifying

cases in which the data of one source were systematically shifted by one sounding or by 24 hours relative to the data in another source, as was the case between the NCAR DS 353.4 and U.S. Air Force DS3 datasets, eventually leading to the exclusion of the DS 353.4 dataset from IGRA (see Section 3.2).

3.4.4 Treatment of High-Frequency Observations

Although radiosonde and pilot balloon observations are typically not taken more than four times a day, there are circumstances under which observations are made more frequently. During the first half of the 20th century, observations were often irregularly spaced throughout the day and were not necessarily made at standard times. More recently, frequent observations have been associated with specific weather events or projects. A combination of observations that are closely spaced in time and the different ways in which observation time is reported in the various data sources complicates the task of retaining only one sounding record for every observation in the integrated IGRA dataset. Therefore, the integration algorithm was designed to address clusters of high-frequency soundings before proceeding with the actual integration of the data sources.

Specifically, within clusters of high-frequency soundings, only those soundings with one common data source were retained. A high-frequency cluster was defined as a sequence of soundings within which all pairs of consecutive observations are at least 60 minutes and less than 300 minutes apart and that is separated from the previous and subsequent soundings by at least 24 hours. If all soundings in such a cluster originated from the same source, they were assumed to represent true high-frequency observations and were retained unchanged. If more than one data source was represented within a cluster, then the close-in-time observation times often were the result of differences in how observation time was reported in the different sources, although true high-frequency observations could also be present within the same cluster. Therefore, the data source with the largest number of soundings in the cluster was chosen when multiple data sources were represented in the cluster. If two or more sources had the same number of soundings in the cluster, the source whose soundings were retained was chosen on the basis of a hierarchy of sounding characteristics. For the data ending in 2013, there were 283,020 high-frequency clusters with two or more sources from which only one source was chosen and 114,092 that contained soundings from only one source.

For example, at Amundsen-Scott (IGRA ID AYM00089009), there was a 0000 UTC observation from bas-data each day between July 6 and 10, 1998, each of which was preceded by a 2200 UTC observation from usaf-ds3. Since bas-data always contained fewer levels and less data than usaf-ds3 within the stretch of 10 soundings, only the usaf-ds3 soundings were retained. If all of the soundings had come from the same data source, they would all be retained.

3.4.5 Elimination of Duplicates of the Same Observation

The primary part of the integration algorithm consisted of an iterative process in which one sounding was chosen from each of three types of clusters of

soundings. The iterations continued on a station's record until no more modifications were made in the previous iteration. Following are the three types of clusters:

A. Clusters of two or more consecutive soundings in which the time spacing between consecutive soundings is 180 minutes or less, and each sounding (1) has the same release time as another sounding, (2) has a release time that is within 30 minutes of the nominal hour of another sounding, or (3) has a nominal hour that is within 30 minutes of the release time of another sounding.

   Examples of year-month-day-nominal hour-release time sequences that meet these criteria are shown below:
   Example 1: 2013 03 20 14 1332; 2013 03 20 15 1332
   Example 2: 1992 01 06 17 1844; 1992 01 06 19 9999; 1992 01 06 19 9999
   Example 3: 2001 01 02 21 2040; 2001 01 03 00 2040
   Example 4: 2001 03 22 21 9999; 2001 03 23 002040

B. Clusters of two or more consecutive soundings with the same time difference to the reference time. Given the preceding check for clusters of soundings with similar release times, these clusters were generally characterized by soundings with the same observation hour and no release time.

C. Clusters of two or more consecutive soundings spaced at 60 minutes or less, spanning no more than 60 minutes total, and separated from the sounding before and the sounding after the cluster by at least 300 minutes. For example, the three soundings 1928 11 20 99 1900, 1928 11 20 99 1930, and 1928 11 20 99 2000 form a cluster based on this criterion, as do the soundings 1988 02 02 23 9999 and 1988 02 03 00 9999.

Whenever a cluster of the above types was identified, one of the soundings in the cluster was chosen for retention in the integrated dataset based on the following six criteria, in order of importance:

1. Source hierarchy, consisting of six priority levels (Table 3.11). The priority level of a particular source was determined subjectively on the basis of several data completeness and quality characteristics relative to the other sources. Datasets containing observed relative humidity were ranked higher than those without; digitized data were ranked higher than GTS data, and unprocessed GTS data were ranked above datasets to which additional processing had been applied by the provider; the archive version of the NWS data (ncdc6301) was ranked higher than the real-time NWS data (ncdc-nws); the lowest priority sources were those characterized by some type of processing uncertainty, such as frequently poor matching results with other sources or the lack of documentation about the origin of their humidity or wind data.
2. In the event of a tie based on source hierarchy, a score indicating the

number and types of variables available: 7 = temperature, wind, and humidity; 6 = temperature and wind; 5 = temperature and humidity; 4 = temperature only; 3 = humidity and wind; 2 = wind only.
3. If there is still a tie, the number of mandatory pressure levels.
4. If there is still a tie, the total number of levels.
5. If still tied, the presence or absence of a surface level.
6. Finally, if still tied, a score indicating the nominal hour/release time information available: 3 = nominal hour and release time available; 2 = nominal hour is one of the synoptic hours 00/06/12/18; 1 = non-nominal synoptic hour without release time.

Table 3.11: Data source hierarchy used during data integration. Data source codes are defined in Table 3.2.

Rank	Description	Sources
1	Good-quality datasets with observed RH or sources with good data for the U.S. during the 1950s	cdmp-amr, iorgc-id, mfwa-ptu, ncar-ccd, ncar-mit, ncdc6310
2	NCEI-processed NWS data; considered official by NWS; observed RH; 1950s data of poorer quality than highest-priority datasets	ncdc6301
3	Raw NWS data; observed RH; better resolution and precision than GTS data	ncdc-nws
4	Other single-source or digitized datasets of generally good quality	cdmp-mgr, cdmp-us3, cdmp-usm, cdmp-zdm, erac-hud, mfwa-wnd, ncdc6322, ncdc6326, ncdc6355, ncdc-gts
5	Uncertainty about how humidity or height were obtained and processed by provider; multi-source archive; low vertical resolution; or some poor data comparisons	bas-data, cdmp-adp, cdmp-us2, ncdc6210, ncdc6309, ncdc6314, ncdc6315, ncdc6316, ncdc6319, ncdc6323, ncdc6324, nsi-hara, usaf-ds3
6	Frequent poor data comparison and station matching results with other sources	chuan101

3.4.6 Output

Three output files were created for each station:

1. XXXXXXXXXXX-sourcedata.txt contains the source soundings considered for integration.
2. XXXXXXXXXXX-data.txt contains the retained soundings.

3. XXXXXXXXXXX-srcs.txt lists, for each sounding in the integrated output, the source station ID, date, and time of the pressure levels and of the non-pressure levels.

In addition, a detailed log of the decisions made by the program was kept in a log file that contains all messages for all stations processed.

3.5 Manual Verification of Station Matching and Integration Results

At various points throughout the station matching and integration process, the results were inspected manually in a variety of ways in order to check if the algorithms were working as intended and to identify errors in station metadata that could be corrected. To some extent, these tasks were accomplished as part of the threshold selection evaluations that were performed during the design of the various algorithms. However, some additional steps were taken to ensure that the stations were matched as cleanly as possible and that as few source stations as possible were removed from consideration. These steps should be repeated whenever IGRA is rebuilt from scratch and new data are incorporated.

3.5.1 Identification and Resolution of Common Coordinate Errors

As part of the pair-finding algorithm (see Section 3.3), the coordinates of each pair of stations were checked for specific types of errors that were found to be common during the initial inspection of pairs with conflicts in matching criteria. If the automated algorithm determined that one of these types of errors was likely to be present at one of the stations in a pair, these findings were noted via error codes in the log file generated by the pair-finding process. If the names of the two stations matched according to the criteria specified in Section 3.3.3.4, yet the station IDs and data did not indicate a match, and the distance between the two stations exceeded 40 km, a record identified with the matching code of "LOCERROR?" was written to the associated log file.

The specific error codes for a given pair of station 1 and station 2 were recorded separately for latitude and longitude in fields 16 and 17 (fields 18 and 19 if using gawk) of the find-pairs.log log file, regardless of whether the record is labeled as "LOCERROR?". The codes can take the following values:

0 = None of the common errors is suspected.

1 = The latitude or longitude of station 2 appears to have been converted incorrectly from positive degrees and minutes to negative decimal degrees. Correct conversion involves changing the minutes to a decimal fraction, adding that decimal fraction to the degrees portion of the coordinate, and then multiplying the result by -1. An error that was detected, particularly among CHUAN stations, suggested that the degrees portion of the coordinates had been multiplied by -1, and then the decimal equivalent of the minutes was added to that negative number, resulting in a coordinate that was to the north or east of the actual location. The error thus introduced increased with the number of minutes. For example, the longitude of an early station at Honolulu was listed by the CDMP as 157° 57', which converts to a decimal longitude of

-157.95°. In v1.01 of CHUAN, however, there existed a station with the same name, "HONOLULU", and latitude, but with a longitude of -156.05°, i.e., -157+(57/60) rather than -1*(157+(57/60). This error was detected for all CHUAN stations whose data had originated from the CDMP and which were located in the Western or Southern Hemispheres.

2 = Latitude or longitude of station 1 appears to have been incorrectly converted from positive degrees and minutes to negative decimal degrees.

4 = Latitude or longitude may be off by 1 degree.

5 = Possible errors 1 and 4.

6 = Possible errors 2 and 4.

10 = Latitude or longitude may have the incorrect sign. This error was particularly likely for latitudes near the Equator and longitudes near the Prime Meridian or Dateline. For example, in one data source, the longitude of Mys Shmidta was given as 179.48 rather than the correct -179.48.

11 = Possible errors 1 and 10.

12 = Possible errors 2 and 10.

14 = Possible errors 4 and 10.

15 = Possible errors 1, 4, and 10.

16 = Possible errors 2, 4, and 10.

Particularly when a pair has been labeled as "LOCERROR?", Leaving such errors uncorrected and unaddressed could result in two source stations that should represent the same location to be added as separate IGRA stations or in one or both stations being withheld from IGRA. Consequently, all records with a nonzero error code for either latitude or longitude were then inspected manually to determine the likely cause and appropriate solution to the problem. In most cases, it was easy to correct the offending coordinate on the basis of the pair finding results and independent confirmation of the correct location of the city or town referenced in the station name. These corrections were implemented via the record-fixes.txt file used in the process of creating the consolidated list of source stations (see Section 3.3.2 above). In the case of the CHUAN stations that had originated from the CDMP, the entire set of stations were withheld from IGRA since the original CDMP data were already being included in IGRA.

3.5.2 Investigation of Various Types of Conflicts Between Matching Criteria

Stations whose pairings included conflicting matching criteria were examined to determine whether their metadata required any correction, whether the station should be removed from the dataset, or whether the elimination of the conflict could be left to the algorithm. If reliable location information could be obtained, either from other data sources or from the Internet, then such errors were corrected by creating a correction record in one of the files designed for this purpose, and the consolidated source station list was re-created (see Section 3.3.2.4). Stations whose metadata could not be corrected and whose pairs were systematically affected by conflicts were removed from the consolidated list of source stations by means of if statements in the program creating that list.

Three types of cases in which the various matching criteria delivered different

signals were investigated, using the records in the find-pairs.log log file:

A. Cases in which the station IDs or data matched, but the distance exceeded 40 km: reasons for these situations included errors in station coordinates, incorrect assignment of station IDs, and the reuse of a station identifier for a completely different location. Solutions included corrections to the coordinates or removal of one or both stations from consideration for IGRA.
B. Cases in which one type of station ID matched, and another type of station ID differed (e.g., same WMO ID, different WBAN ID): reasons included changes in one type of station ID over time, while the other state unchanged (requiring no intervention) and incorrect assignment of a station ID (which could sometimes be corrected by removing or correcting the offending station ID without withholding the data from IGRA).
C. Cases in which stations were within 10 km of each other, and the PORs overlapped, but the data did not match or could not be compared: reasons included the existence of civilian and military stations with different observing practices (i.e., different measurement precisions, observation times, vertical resolutions, or variables observed) in close proximity to each other (often necessitating the retention of both records as separate stations or the quarantining of one), the presence of thermodynamic observations in one data source and pilot balloon observations for the same location in the other (requiring that the two records be merged), an erroneous timeshift of the data in one record relative to the other (requiring the removal of all or part of one or both records from consideration), and inexplicable data problems (requiring the quarantining of one or both source stations).

3.5.3 Inspection of Complex Mingle Groups

The output from the algorithm that created the mingle groups from the pair-finding results was also carefully inspected. Included in the inspections were samples of complex mingle groups, including groups with more than 15 members, collections of stations which were initially assigned to one group and later split into two groups by the automated algorithm, and groups from which stations were removed due to conflicts or transitivity violations. In some cases, these examinations resulted in corrections to the metadata of some additional stations or in the removal of certain stations from consideration. When such actions were required, they were again implemented at the stage of creating the consolidated source stations list, and the processes of creating a list, running the pair-finding algorithm, and creating the mingle groups was repeated.

3.5.4 Inspection of Quarantined Stations

The stations that were quarantined by the automatic pair-finding and mingle group creation algorithms were also examined in various ways to determine whether their removal resulted in the decimation of any long or otherwise prominent data records and whether such decimation was caused by a

rectifiable issue in the data or metadata. Approximately 40% of the 389 stations quarantined during the station matching process had 2 years or less of data, while another 30% had 2-6 years of data. The inspection of individual cases therefore focused on source stations whose removal resulted in a loss of more than 10 years of data that were not covered by any other source station for the same location. This inspection resulted in a few additional corrections to metadata, again necessitating the rerunning of all components of the station matching and integration process.

3.5.5 Check of Country Codes Assigned During Data Integration

If a country code had been supplied as part of the station's metadata, this country code was compared to the country code obtained from the Internet on the basis of the station's coordinates. Cases in which these two independently-determined country codes differed were investigated, and appropriate action was taken within the integration program to ensure that the appropriate country code would be used in the station ID. In a few cases, these investigations also resulted in additional coordinate corrections. In others, the difference in country codes could be traced to political changes over time (Namibia vs. South Africa, for example) or continuing political disputes (e.g., China vs. Taiwan).

In addition, the complete set of IGRA station IDs generated by the integration process was examined for any cases in which no country code had been assigned, i.e., the IGRA station ID began with XX. Such situations resulted in an additional if statement in the integration program to set the correct country code or in a coordinate correction implemented at the stage of creating the consolidated station list.

3.6 Quality Assurance

3.6.1 Overview

The IGRA 2 QA system is based largely on the QA procedures in the IGRA 1 system (Durre et al. 2006; Durre et al. 2008). Like the IGRA 1 system, it consists of a deliberate sequence of specialized algorithms, each of which makes a binary decision on the quality of a value, level, or sounding; either the data item passes the check and remains available, or it is identified as erroneous and thus set to missing.

The procedures in the IGRA 2 QA system can be grouped into eight categories: fundamental “sanity” checks, checks on the plausibility and temporal consistency of surface elevation, internal consistency checks, checks for the repetition of values, checks for gross position errors in ship tracks, climatologically-based checks, checks on the vertical and temporal consistency of temperature, and data completeness checks (Table 3.12).

The following ten processing steps are required to apply the full sequence of checks:

1. Generation of time series of monthly median elevations for each station (medelev.f95). These time series are required for various elevation-related checks.
2. Application of basic QA checks to the integrated dataset (qa1.f95), including absolute limits checks, hypsometric pressure-height consistency check, checks for multiple surface levels, insertion of monthly median elevations into surface levels, height sequence and elapsed time sequence checks, checks for below-surface levels, vertical runs checks, and some checks for certain unique problems.
3. Application of a semi-automatic process for identifying gross position errors within the tracks of observing ships.
4. Application of temporal runs checks to surface pressure and standard-level geopotential height and temperature and removal of soundings flagged by the ship track checks (qa2.f95).
5. Calculation of gross full-year climatologies for each station and surface/standard pressure level for use in gross climatological checks (climstn.f95).
6. Application of gross climatological checks to surface temperature, to pressure at or above the station's main surface pressure, as well as to geopotential height and temperature at all other pressure levels (qa3.f95).
7. Computation of day- and time-specific climatologies for each station and surface/standard pressure level for use in various subsequent checks (climwin.f95).
8. Utilization of the day- and time-specific climatologies for the application of outlier and vertical consistency checks to temperature as well as of outlier checks to pressure at or above the mean surface pressure and to geopotential height (qa4.f95).
9. Application of temporal consistency checks to temperature at the surface and standard pressure levels (qa5.f95).
10. Removal of small, isolated groups of soundings from station records and other cleanup (qa6.f95).

Table 3.12: IGRA quality assurance procedures and their impact.

Category	**QA Procedure**	**Items Checked**	**Items Deleted**
Basic Plausibility Checks	Invalid Release Time Check	Release time	Release Time
	Invalid Data Value Check	p, z, T, RH, d, ws, wd	Individual values
	Unrealistic Wind Profile Checks	WS	WS profiles
	Invalid Level Type Indicator Check	LT	Level
Station Elevation Checks	Paucity/Inconsistency/Spike Checks	Surface height	Surface height

Internal Consistency Checks	Hypsometric Check	p, z	Individual levels
	Vapor Pressure/Pressure Consistency Check	RH or d; T, p	RA or d; T
	RH-Dew point Depression Consistency Checked	RH, d, T, p	RH, d, T
	Height Sequence Check	z	Levels
	Release time sequence check	Release time	Level
	Multiple Surface Levels Check	Level type indicator	Levels
	Below-surface Level Check	p, z	Level
	Obs Hour/Release Time Check	Obs Hour – Release Time	Release time
	Zero-speed Wind check	ws, wd	ws, wd
	Level Type Checks	LT, p, ET	nothing; LT is edited
Ship Position Checks	Excessive Speed Check	Coordinates	Soundings
	Speed Spike Check	Coordinates	Soundings
	Speed Consistency Checks	Coordinates	Soundings
Checks for Repetition of Values	Temporal runs check (generic)	p, z, T	Levels or values
	Temporal runs check (by hour)	p, z, T	Levels or values
	Vertical runs check	T	values
	Joint vertical runs check	T, d, ws, wd	values
	Frequent erroneous values check	z, T	values
	Fixed Geopotential Height Check	z (Russian GTS only)	values
Climatological Checks	Tier 1	p, z, T	levels or values
	Tier 2	p, z, T	levels or values
Additional Checks on Temperature	Crazy Profile Check	T	T-soundings

	Generic vertical outlier check	T	values
	Vertical sore-thumb check	T	values
	Temporal sore-thumb check	T	values
Data Completeness Checks	Lone dew point depression check	d. T	values
	Lone wind value check	ws, wd	values
	Incomplete level check	p, z, d, T, ws, wd	levels
	Isolated sounding check	date and time	sounding

When a value is removed by a QA procedure, the removal is recorded in two log files. One is a station-specific log file that lists all values removed by a particular QA program from that station's record. The other provides explanatory messages for all values removed by a specific QA program at all stations. In both types of files, a three-letter code identifies the cause of the removal. These codes, along with the programs that set them and the number of values removed, are listed in Table 3.13: This table is intended to be used as a reference when the cause of a particular removal needs to be traced or when verifying the QA system's output during a future reprocessing of the dataset. Files listing the values that were identified as erroneous can be provided to users of the dataset upon request.

Table 3.13: Error codes written to QA log files. Listed for each code are the meaning, responsible program, and frequency of occurrence. Frequencies of occurrence are expressed in the total number of soundings removed; for all other codes, frequencies represent the total number of values removed. Counts of 0 are shown for QA checks but did not find any errors since the relevant problems had already been eliminated during the standardization and integration of data sources.

Code	Meaning	Program	Frequency	
BSP	Below-surface pressure	qa1	763016	
BSZ	Below-surface height	qa1	1177527	
DVI	Data value invalid	qa1	584562	
EEP	Excessive water vapor pressure-pressure ratio	qa1	97700	
ELR	Elevation replaced	qa1	2515145	
ELU	Elevation unknown	qa1	234830	
ETD	Elapsed time duplicate	qa1	510922	
ETS	Elapsed time out of sequence	qa1	410954	
FEV	Frequent erroneous value	qa1	145977	

GHF	Geopotential height fixed	qa1	88566	
GHS	Geopotential height out of sequence	qa1	14219503	
GPF	Pressure failed gross (tier-1) climatological check	qa3	109244	
GVF	Value other than pressure failed gross (tier-1) climatological check	qa3	1705867	
HLT	Height (non-pressure) level with thermodynamic data	qa1	0	
IS1	Isolated soundings, type 1	qa6	2614 soundings	
IS2	Isolated soundings, type 2	qa6	2361 soundings	
IS3	Isolated soundings, type 3	qa6	1211 soundings	
IS4	Isolated soundings, type 4	qa6	10711 soundings	
LNU	Level Nolan and usable (after first basic checks)	qa1	779647	
LTE	Level type indicator edited	qa1	12430447	
LTI	Level type indicator invalid	qa1	0	
PCF	Pressure failed tier-2 climatological check	qa4	203595	
PEH	Pressure-elevation pair fail hypsometric check	qa1	358892	
PZH	Pressure-geopotential height pair failed hypsometric check	qa1	6258677	
PZO	Level with only Pressure and geopotential height	qa1	0	
RTI	Release time invalid	qa1	0	
RTO	Run in time for temperature or geopotential height	qa2	12256	
RTP	Run in time for pressure	qa2	50205	
RV1	Run in the vertical for one element	qa1	31633	
RVJ	Around in the vertical for joint elements	qa1	83006	
SLM	Multiple service levels	qa1	2267126	
TCF	Temperature failed tier-2	qa4	417157	

	climatological check			
TPE	Track position error (for mobile stations)	qa2	162488	
TST	Temperature sore thumb	qa4	39358	
TT1	Temporal temperature outlier, type 1	qa5	20803	
TT2	Temporal temperature outlier, type 2	qa5	4064	
TZO	Temperature z-score profile outlier	qa4	265011	
TZS	Crazy temperature z-score sounding	qa4	99750	
UVI	Humidity values inconsistent	qa1	14694	
UVL	Lone humidity value (no valid temperature)	qa1-qa5	1319695	
WPI	Wind profile invalid	qa1	13932	
WSC	Surface wind levels combined	qa1	4730583	
WSD	Winds at surface differed	qa1	97709	
WSI	Height of surface wind level inconsistent with other levels	qa1	12654	
WVI	Wind value invalid	qa1	276326	
WVL	Lone wind value	qa1	1941334	
ZCF	Geopotential height failed tier-2 Climatological check	qa4	1199970	
DVI	for relative humidity and wind direction - invalid data value			
HLT	non-pressure levels with thermodynamic data			
LTI	Invalid level type indicator			
PZO	Levels with only pressure and height and no other data			
RTI	invalid release times			

Although many of the checks are the same as those in the IGRA 1 system, a number of enhancements were implemented in order to modernize and streamline the code to accommodate characteristics of the IGRA 2 data that were not present in IGRA 1 and to improve the quality of the final wind and humidity data. These modifications are described in the following subsections.

3.6.2 Handling of Surface Levels

3.6.2.1 Definition of the Surface Level

As discussed in Durre et al. (2006), a sounding typically contains a set of surface observations at or near the site where the balloon was launched. These observations, when provided, are identified in one way or another as the so-called surface level. In IGRA 1, which did not contain any soundings without pressure levels, all surface levels were required to contain a surface pressure. In Version 2, pilot balloon observations without any pressure levels are now included. Therefore, the criteria for a surface level were modified: The surface level in a sounding is now permitted to be a surface pressure level, a surface level with only a height, or a surface level without either pressure or height.

3.6.2.2 New Approach for Calculating Monthly Median Elevations

When geopotential height is provided at a surface level, it was likely not transmitted by the observer, but was inserted at a data processing center on the basis of station metadata available at the time of the insertion. As shown in Durre et al. (2006), the time series of surface heights can be plagued by spurious spikes and shifts that are caused by incorrect metadata, by processing problems, or by the integration of multiple sources reporting different elevations for the same station and time. Consequently, it was necessary to develop procedures for the removal of gross errors and unrealistic temporal variations in surface level heights. These procedures involved the computation of time series of median elevations for each station, the removal of elevations causing unrealistic spikes and jumps in the resulting series of surface heights, and the replacement of each sounding's surface height with the corresponding monthly median elevation. The resulting surface heights were used in subsequent QA procedures for determining the validity of surface levels and for identifying levels whose height is below the surface.

For IGRA 1, this approach was implemented in a semiautomatic fashion which required extensive manual inspection and, therefore, was not reproducible. For IGRA 2, many aspects of the IGRA 1 approach were integrated into a new algorithm that can be applied fully automatically. Nevertheless, the algorithm flags cases in which a manual correction of the metadata, or the quarantining of a station, might improve the overall quality of its output.

The IGRA 2 algorithm for generating and cleaning up the time series of monthly median elevations at a particular station consisted of the following nine steps:

1. Pre-computation checks and requirements: First, the station's soundings were subjected to the height sequence and hypsometric checks (Durre et al. 2006). Only soundings that did not have height sequence or hypsometric check violations at any pressure level were considered. In an early version of the algorithm, the coordinates in the sounding header record were also required to be within 40 km of the coordinates provided in the station list, but this requirement was dropped because there were many cases in which the header coordinates were incorrect, yet the

surface height or elevation was correct.

2. Computation: The median elevation was calculated for each year and month as long as an elevation was available from either the surface level or the sounding header of at least five soundings. Compared to IGRA 1, the minimum number of individual heights required was increased from three to five in order to improve the robustness of the resulting medians.
3. Preparation for analysis: For each time series of monthly median elevations, the unique elevations were identified, and some statistics were calculated for each of these unique elevations. The statistics included the number of year-months in which the elevation occurred, the length of the longest streak of consecutive occurrences of the elevation (ignoring missing values), the first and last years of occurrence, and the average of the monthly median surface pressures during the year-months in which the elevation occurred.
4. Paucity check: When a station had more than one unique median elevation, and the longest consecutive streak of any of those elevations was shorter than four months, all occurrences of that elevation were flagged with a flag of "K", and the elevation was excluded from consideration in all subsequent steps. The threshold of four months was chosen based on an evaluation of five cases each for thresholds of 1, 2, 3, and 4. At thresholds of 1 and 2, none of the elevations evaluated was considered to be of any use. At a threshold of 3, the elevation evaluated was judged to be incorrect in three of the five cases; in the other two cases, the elevation seemed plausible, but the few occurrences diminished its credibility, and its removal (which would later result in the use of an elevation from an adjacent month) was not considered to be disadvantageous. At a threshold of 4, however, there were clear cases in which the elevation might indeed be correct, and its removal might be considered a loss of information.
5. Determination of reference elevations: In order to be able to perform additional checks on the median elevations, it was necessary to know which, if any, of the station elevations were reasonable. To assist in that determination, two types of reference elevations were obtained:
    (a) Gridded elevation model: The coordinates provided in the IGRA 2 station list were used to obtain the location's elevation from the Google Web service or, if Google did not provide a useful elevation, from the geonames Web service. In order to account for effects of complex terrain and coordinate inaccuracies, reference elevations corresponding to the four points that were located 40 km to the north, east, south, and west of the station list coordinates were also obtained. From the resulting up to five reference elevations, the maximum and minimum were determined after excluding missing values and values less than the lowest elevation on earth. If the station is a ship, buoy, or ice island, the minimum and maximum reference elevations were set to zero.
    (b) As an additional reference, the elevation above sea level that corresponded to the average surface pressure at the station was calculated by averaging the median surface pressure values from all year-months in which there were at least five surface pressure observations and then estimating the corresponding elevation

using the hypsometric equation and an average layer temperature of 273.15 K. This pressure-based reference elevation could not be calculated for stations without surface pressure levels.

6. Determination of anchor elevation: The two types of reference elevations were then used to check whether the most recent monthly median elevation could be used as an "anchor" against which the remaining elevations in the time series were to be checked. The latest monthly median elevation was used as the anchor elevation if it met one of the following three conditions:
    (a) It fell within, or reasonably close to, the range of grid reference elevations;
    (b) it was relatively close to the pressure-related reference elevation; or
    (c) in the case of a ship, buoy, or ice island, it lay between 0 and 10 m.

    If none of these conditions are met, the elevation provided in the station list was used as the anchor. "Relatively close" meant that the anchor elevation was no more than 150 m below the smallest relevant grid elevation and no more than 150 m above the largest grid elevation. This seemingly liberal range was necessary because both the grid-based and pressure-based reference elevations can be inaccurate in the presence of complex terrain and/or imprecise station locations.
7. Inconsistency check: Each of the unique elevations not flagged by the paucity check was compared to the anchor elevation. If it was found to be inconsistent with the anchor elevation, all occurrences of that elevation were flagged with "I". Inconsistency occurred when one of the following three conditions was true:
    (a) the elevation and the anchor elevation differed by more than 250 m;
    (b) the difference to the anchor elevation was between 125 and 250 m, yet the corresponding difference in surface pressure did not support the elevation difference; or
    (c) the elevation was between 125 and 250 m higher than the anchor elevation and the corresponding surface pressure difference could not be calculated.

    The threshold of 250 m was obtained by evaluating multiple cases within several ranges of elevation differences. For elevation differences greater than 500 m, no false positives were found, but three cases were identified in which the anchor elevation was incorrect due to erroneous surface pressure levels. For differences between 300 and 500 m, still no obviously good elevation was found to be flagged. However, there were two or three cases in which it was difficult to determine whether the elevation should be flagged, yet flagging seems slightly preferable to not doing so. The same was true for elevation differences between 250 and 300 m. Lowering the difference threshold to 125 m, however, would have flagged some clearly valid cases, particularly when the elevation being checked was lower than the anchor elevation. The validity of those cases was generally corroborated by a comparable change in surface pressure. Hence, conditions (b) and (c) were added for elevations between 125 and

250 m. For differences below 125 m, cases started to appear where it was difficult to distinguish changes in surface pressure that were caused by changes in elevation from those caused by other, climatological or non-climatological, factors.

8. Last resort: If all elevations at the station were flagged by the inconsistency check, they were replaced by the anchor elevation in order to produce a suitable set of elevations. An evaluation of the affected cases suggested that this was a reasonable approach. Without it, there would have been no elevation information to use when checking both for the reasonableness of surface levels and for the presence of below-surface levels.
9. Spike check: The time series of monthly median elevations was separated into segments of constant elevation. The spike check then compared each segment to the segments before and after. A segment consisting of only one or two months (not counting intervening months with missing elevation) was flagged if the segment elevation was at least 25 m higher or lower than the elevations in the previous and subsequent segments. The same was done for segments consisting of 3, 4, 5, and 6 identical elevations that were at least 50, 75, 100, and 125 m higher or lower than adjacent segments, respectively. Elevations that failed this check were flagged with the letter "J".

Following the application of this algorithm to the integrated dataset, the following three types of situations were manually investigated in order to determine, for each affected station, whether it should be removed from the dataset or whether its metadata could be corrected: stations for which no median elevation file was written to the output file; cases in which the pressure-based reference elevation was radically different from the gridded elevation, and stations where all median elevations were flagged.

Once all feasible corrections had been made and the algorithm had been reapplied, elevation time series were produced for 3053 of the 3067 stations with integrated IGRA 2 data. These included 168 stations where no monthly median elevations could be calculated directly, and where monthly elevations were therefore assumed to be equal to either sea level (73 mobile stations or fixed ships) or the elevation shown in the station list (95 cases). Of all stations and months in which an elevation was either calculated or estimated, 1.2% were flagged by the paucity, inconsistency, and spike checks combined, a flag rate that was comparable to the 1% flag rate which resulted from the manual identification of erroneous elevations in IGRA 1 (Durre et al. 2006). The 14 stations for which no monthly median elevation could be calculated or estimated contained a very small number of soundings and were excluded from IGRA 2.

3.6.2.3 Use of the Monthly Median Elevation

The process for inserting the monthly median elevations into the sounding surface levels was modified to incorporate the new monthly median elevation time series as well as the presence of surface levels without the pressure. As in IGRA 1, the median elevation for a particular station, year, and month, when

available, was inserted into the surface pressure level of every sounding in the month. In addition, the median elevation was now also inserted into surface levels without pressure as long as that level was the first non-pressure level in the sounding, and the monthly median elevation was lower than the height of any other non-pressure level in the sounding. If neither of these criteria were met, the surface non-pressure level was removed.

All in all, the median elevation was inserted into approximately 2.4 million, or 4.5%, of the soundings when IGRA 2 was first created. In approximately 480,000 of these cases, the surface height was missing originally, and the median elevation was inserted. In another 189,000 cases, or 0.4% of all soundings, the original surface height was replaced by a monthly median elevation that was more than 10 m different.

When the median elevation was missing or flagged for a particular station/year/month, no missing surface level heights were filled in. Rather, any surface height that may be available during the month was checked against the last unflagged median elevation from a previous month; if the surface height differed from that proxy elevation, it was removed from the sounding. If monthly median elevations were missing or flagged at the beginning of a record, the first available median elevation was used to check the surface heights during that part of the record. This approach for handling months without a median elevation represented a change from the IGRA 1 system, which removed all surface heights in such months. The modified IGRA 2 approach led to the retention of 50% more surface heights in these situations compared to IGRA 1.

A modification was also made to the check for below-surface levels (Durre et al. 2006). Considering the uncertainty in monthly median elevations, the threshold used for identifying below-surface heights in soundings without a surface level was increased from 10 m, used for IGRA 1, to 50 m. This reduced the flag rate of this check by approximately 40%.

3.6.2.4 Check for Multiple Surface Levels

Given the expanded definition of a surface level, the check for the presence of multiple surface levels in the same sounding (Durre et al. 2006) was modified to both accommodate the presence of surface levels in pilot balloon observations and limit collateral damage. In IGRA 1, whenever there were at least two surface levels of any kind in a sounding, all such levels were removed. In IGRA 2, the process worked as follows:

1. Pressure levels and non-pressure levels were checked separately.
2. If there were two pressure levels that were identified as a surface level, and one of them was a mandatory pressure level while the other was not, only the mandatory pressure level was removed. If there were multiple surface pressure levels, and they did not include exactly one level at a standard pressure, all pressure levels identified as surface levels were removed. This happened most frequently when there was a 1000-hPa level that was erroneously identified as the surface level in addition to the seemingly actual surface level at a nearby pressure.

3. When there were multiple levels that had neither height nor a pressure and were identified as a surface level, then one was retained if the wind data at all of those levels were the same; otherwise, all such levels were removed.
4. If there were still multiple non-pressure surface levels after step #3 above, all such levels were removed.
5. If, after all of the above steps, there remained exactly one surface pressure level and one surface non-pressure level, the surface non-pressure level was removed. If the wind data at that non-pressure level differed from those at the pressure level by more than 10° or more than 2 m/s, the wind data at the surface pressure level were also removed. If, on the other hand, that non-pressure level contained wind data that were not available at the surface pressure level, those wind data were inserted into the surface pressure level.

Multiple surface levels were detected in 0.83% of the soundings in the IGRA 2 data ending in December 2013. In two thirds of those cases, it was possible to retain one surface level. In the other one third, all surface levels had to be removed. Most of those latter cases involved surface levels without either a height or a pressure that were frequently included in the wind level groups of soundings transmitted via the GPS.

3.6.2.5 Checks on Surface Pressure Time Series

As part of the development of IGRA 1, surface levels were removed from specific periods of time at both Chinese and Russian stations, following the visual inspection of all surface pressure time series. This removal of manually identified surface pressure and temperature segments at Chinese and Russian stations was eliminated since the configuration of underlying data sources was different in IGRA 2 compared to IGRA 1. Instead, particular attention was paid to the time series of surface pressure during the post-processing validation exercises. As described in the Verification and Validation Report, no systematic problems were identified in the IGRA 2 series, although shifts at some individual stations remained.

3.6.4 Other Modifications to Previously Existing Checks

3.6.4.1 Check for Excessive Differences between Observation Hour and Release Time

The check for excessive differences between observation hour and release time was changed to remove the release time rather than remove the entire sounding when the difference exceeded four hours. This modification was implemented because it had been discovered that at some stations, release time appeared to have been systematically entered in local time rather than in UTC. It was therefore desirable to retain the data at those stations, despite the apparent disagreement between release time and observation hour. In the IGRA 1 system, the South Pole station had been exempted from this check because the frequent excessive observation hour-release time differences seemed inconsequential at a location with no 24-hour day/night cycle. In the IGRA 2 version of the check, this exemption was no longer necessary since the check no longer removed any observations.

A total of approximately 22,500 soundings were flagged by this check, 6241 of them at one station, AGM00060390. The stations with the next two highest number of soundings affected had flags in approximately 1600 and 1000 soundings, respectively. At all of those three stations, the observation hour appeared to be reasonable, but there was a problem with the release time.

3.6.4.2 Basic Plausibility Checks

Basic plausibility checks determine whether the data values in each sounding fall within certain gross plausibility limits (Table 3.14). The upper limit for pressure was raised from the value of 1078.5 hPa used in IGRA 1 to 1090 hPa, a value slightly above the newly accepted world record highest sea level pressure of 1089.1 hPa (see http://wmo.asu.edu/highest-sea-lvl-air-pressure-above-700m). The upper limit for geopotential height was lowered from 70,000 m to 60,000 m, closer to the highest reported height achieved by a weather balloon (Blackmore, personal communication). As discussed in Section 3.6.5 below, the upper limit for temperature was lowered from 70°C to 60°C. Each of these three checks removed fewer than 0.01% of the available data values for the respective variable. Finally, the checks for out of range RH and wind direction values, which were not just implausible, but physically impossible or unmeasurable, were also performed as part of the integration program and therefore did not remove any additional values within the QA system.

Table 3.14: Plausibility limits used in basic validity checks.

Item (Field)	Valid Range
Elapsed time	0 - 240 min
Pressure	1090 – 0.1 hPa
Geopotential Height	-1000 to 60,000 m
Temperature	-120 to 60 °C
RH	0 - 100%
Dew point Depression	0 to 70 °C
Wind Speed	0 – 150 m/s
Wind Direction	0 - 360°

3.6.4.3 Specialized Checks for Frequent Erroneous Values

The IGRA 1 system contained several checks for data problems that were specific to certain regions and years (Durre et al. 2006). For IGRA 2, the sequencing and components of these checks were changed in two ways. First, the check for temperatures between seven and 8°C at the 1000 hPa level in 1969 in China and the check for temperatures of -88.9°C were moved from a later program to immediately after the plausibility check, so that the affected values were removed before the surface-level and runs checks. Second, the

1990-1992 Brazil and 1991 top-pressure level checks were removed entirely. These procedures were designed to address specific data problems in IGRA 1 that were no longer present in IGRA 2. However, the relevant subroutines have been retained for use should data sources containing these problems be incorporated into IGRA in the future.

3.6.4.4 Miscellaneous Other Modifications

The following additional adjustments were made to the QA system:

- Release times in which the minute was 99 were no longer considered invalid because such release times were found in recently digitized data from the early part of the 20th century.
- A provision was added to exclude levels from further QC checks if no temperature or wind data remained after the initial absolute limits, specific-problem, and inconsistency checks.
- The check for vertical runs in temperature was moved from a later QA program to just after the check for joint vertical runs in two elements in QA1. In addition, for this check, mandatory pressure levels were now defined as all standard pressure levels between 1000 and 1 hPa. In IGRA 1, they were defined as standard pressure levels between 1000 and 10 hPa.

3.6.5 New Checks on Humidity

3.6.5.1 Background

In the pre-QA dataset through the end of 2013, there were 467 million pressure levels with dew point depression and 222 million levels with RH, including 122 million levels with both dew point depression and RH. Dew point depression was the only humidity variable transmitted over the GTS, while RH was the primary humidity variable in data prior to 1970. Within this large sample of humidity observations, various basic QA procedures were tested to assess the quality of the humidity data and determine which procedures should be implemented in the IGRA QA system.

3.6.5.2 Checks for Excessively High Vapor Pressure

First, vapor pressure was computed from either RH or dew point depression, whichever was available, and cases in which the vapor pressure was greater than the atmospheric pressure at the same level were identified. The rationale for this approach was that, by definition, for water, vapor pressure exceeding atmospheric pressure means that the boiling point has been surpassed; also, since vapor pressure is part of the atmospheric pressure, it can never exceed the atmospheric pressure. This check flagged 7501 cases based on dew point depression and 1075 cases based on RH. An inspection of the cases flagged showed that almost all of the cases seemed to be the result of erroneously high temperatures, even at the highest levels in the atmosphere (lowest pressures). Only at the very highest levels were there some cases where the temperature

might actually be possible, but in those cases, either the RH was unrealistically high or the pressure was unbelievably low (< 1 hPa) considering the sensitivity of humidity sensors. Even when the check was modified to flag cases in which the vapor pressure exceeded 10% of the atmospheric pressure, all of the cases inspected were related either to excessively high temperatures or excessive levels of humidity. For example, for temperatures less than -10°C, the RH was almost always greater than 70%, values that were unrealistic for levels between 10 and 25 hPa where these cases were found.

Another approach that was tested was to check whether vapor pressure exceeded the vapor pressure that was consistent with the highest reported dew point temperature in the world. Although no official dew point temperature record was available, the unofficial record was around 35°C, observed in the Persian Gulf, which corresponds to a vapor pressure of 56 hPa. Setting the threshold to 60 hPa yielded approximately 10,000 flags, which were concentrated in hot locations in the summer. Without an official world record in dew point depression, it was not possible to determine which of the flagged values were valid and which were not. This basic plausibility check on vapor pressure was therefore not implemented.

Two additional findings are worth noting. First, a significant portion of the cases flagged with any of these approaches involved temperatures above the world record surface air temperature of 56.7°C since the upper limit of the fundamental sanity check for temperature had been set to 70°C for IGRA 1 (Durre et al. 2006) in order to leave sufficient tolerance for higher temperatures at higher levels in the atmosphere. However, all temperatures greater than 60°C that were flagged by the excessive water vapor check seemed to be unrealistic, regardless of their altitude in the atmosphere. Hence the decision to lower the threshold for the fundamental sanity check on temperature to 60°C (Section 3.6.4). Second, in the process of examining the frequency distribution of RH at 10 hPa in the NWS data, it was discovered that when dew point depression was missing, RH was often set to 999, or 99.9%, yielding a frequency of that value that was 1000 times larger than the occurrence of 98 or 100% RH, For a total of 32,000 cases in that data source. Consequently, the program for reformatting the ncdc-nws data was modified to set RH values of 999 to missing when dew point depression was missing, and IGRA 2 was re-created with the modified ncdc-nws data.

3.6.5.3 Inconsistencies Between RH and Dew point Depression

Whenever both humidity values were available at the same pressure level, the dew point depression was used to derive a RH (also using temperature and pressure), and the thus calculated RH was compared to the reported RH. The message was written to the log file when the difference between the calculated and reported RH values exceeded 10%. That threshold was the same as that used when determining whether two RH values from different data sources can be considered to be the same (Table 3.3).

When this check was first applied, differences greater than 10% were found at 443,000 levels, or nearly 0.4% of the total number of levels at which both

humidity variables were available. More than 260,000 of these cases were found to originate from the Meteo-France French West African radiosonde data (IGRA 2 source code mfwa-ptu). A re-examination of the original mfwa-ptu data and documentation led to the determination that dew point temperature rather than dew point depression had inadvertently written into the IGRA 2-reformatted data files for that source. This issue was promptly fixed.

Another 154,000 of the derived-reported RH differences exceeding 10% involved situations in which the dew point depression was 30°C and the RH was 20% at U.S.-operated stations between 1993 and 1995. It is well documented that until around the year 1996, it was a practice at NWS stations to set the dew point depression to 30°C and the RH to 20% whenever the observed RH dropped below 20%. Since GTS data only contain dew point depression, and NWS-originating non-GTS data only contain RH prior to 1993, 1993 was the first year in which both humidity variables were available at U.S. stations. Although the RH and dew point depression values provided in these cases were not accurate, they do provide an indication that conditions were very dry, and removing them from the dataset would create a bias towards more moist conditions. It was therefore decided to raise the threshold of the check from 10% to 20%, implying that only those cases would be flagged in which the reported RH differed from the RH derived from dew point depression by more than 20%. Following correction of the mfwa-ptu data, this variant of the check identified 4898 cases, approximately 60% of which originated from NCEI's DSI-6316 dataset for Argentina.

3.6.5.4 Vertical Runs

The IGRA QA system already contained several checks for vertical runs, the repetition of the same value across several levels in a sounding (Durre et al. 2006). One checked for vertical runs in temperature across mandatory and significant levels separately. The other looked for joint vertical runs in two elements and was applied to the pairs of temperature/dew point depression and wind direction/speed. In a joint vertical run, the same value is repeated across two variables and multiple levels.

With the inclusion of RH in IGRA 2, the opportunity presented itself for two additional joint vertical runs checks. The RH-dew point depression check flagged a total of 43 values, mostly at stratospheric levels where comparable values may occur by chance given the associated temperature. No temperature-RH joint runs were found at all. Consequently, a test file was created to verify that the check was able to identify such runs whenever present. Based on these findings, neither check for joint vertical runs involving RH appeared to be worth including in the IGRA 2 system.

During the development of IGRA 1, it had already been determined that a check for vertical runs in dew point depression would likely lead to the overflagging of deep saturated atmospheric layers. For this reason, the single-element vertical runs check was also not applied to RH.

3.6.5.5 New Humidity Checks Implemented

Based on the above tests, it was decided to implement two checks on humidity. The first checked for vapor pressure exceeding 10% of the atmospheric pressure. When such a case was identified by the procedure, both the humidity variable and temperature from which the vapor pressure was derived were removed from the data file. By extension, if the other humidity variable was also reported at the same level, it was also removed since temperature was no longer available at that level, and since humidity values were allowed to remain in IGRA only when a valid temperature was present at the same level (Durre et al. 2006). This check flagged a total of 51,150 cases for the full period of record through 2013, 5856 of which involved RH. Secondly, when the difference between the reported RH and the RH computed from dew point depression exceeded 20%, both humidity values as well as the temperature at the same level were removed. As stated above, this procedure removes humidity and temperature at an additional 4898 levels.

3.6.6 New Checks on Wind Observations

3.6.6.1 Background

In IGRA 1, the only two checks on wind observations were a check for values exceeding 150 m/s, which were considered to be highly implausible, and an inconsistency check for zero wind speed and nonzero wind direction. Several users of IGRA 1 had noticed that wind speed values during certain months were approximately an order of magnitude larger than expected at a significant number of stations. Particularly affected were the months of January and February during 1975-1978 and 1981, suggesting that the problem may have been the result of the widespread processing error in one of the IGRA 1 data sources, NCEI's global-coverage, 1973-1999 DSI-6302. Since this data source was replaced by the U.S. Air Force data (usaf-ds3), it was not clear a priori whether the same issues would be found in IGRA 2. Consequently, some modest attempts were made to enhance the quality checks on wind values.

3.6.6.2 Checks for Systematically Excessive Wind Speeds During Individual Months

Several techniques were used in order to determine whether systematic shifts in wind speed were present in IGRA 2. The first involved the application of the Standard Normalized Homogeneity Test (SNHT) to each time series of monthly-mean wind speed.

In order to determine whether the SNHT would detect the types of shifts that were known to exist in IGRA 1, the test was first applied to IGRA 1. Specifically, the maximum wind speed at the expected altitude of the jetstream (500-100 hPa, 6-18 km) was determined in each sounding, and a monthly average maximum wind speed was calculated for any station/year/month for which there were at least five such maximum wind speeds. When the SNHT was applied to the resulting time series for each station, the known periods of excessive wind speeds were detected in IGRA 1. When the same approach was applied to IGRA 2, no similarly widespread or systematic issues were detected,

neither in the years 1975-1981 nor in any other year. Although some isolated suspicious-looking features were detected, it was decided not to permanently flag any values with this test because it also identified many changepoints that looked reasonable.

Second, wind speeds in excess of 130 m/s were examined for the presence of any spatial or temporal patterns. Values 2001 and 2567 (in m/s times 10), as well as four-digit values beginning with the digit 9, appeared substantially more frequently than adjacent values. These values may have had some meaning to somebody at some point, but that meaning was no longer documented. These excessive frequently-occurring values were excluded from the next check described below.

Finally, a test was explored that looked for tropospheric portions of profiles in which at least three wind speeds exceeded 130 m/s. To restrict the analysis to the tropospheric portions of profiles, only the winds at pressures greater than 100 hPa or with altitudes below the roughly approximated maximum tropopause were considered. For this purpose, the height of the tropopause was assumed to be at 18000 m for latitude less than 30°, set to 10000 m for latitude greater than 70°, and assumed to decrease by 200 m for each degree poleward between latitudes of 30° and 70°. This check identified approximately 1200 profiles spread over 428 stations, with no indication of any significant clustering of dates at individual stations or across all stations. Upon analyzing these profiles, some possibilities for additional wind profile checks were identified.

3.6.6.3 Implementation of Whole-Profile Checks

Based on the findings described in the previous subsection, checks were explored that were designed to flag wind profiles in which the maximum wind speed was small as well as profiles in which the minimum wind speed was rather large. In order to assess the feasibility of either of these checks, the minimum and maximum wind speed for all pressure and all non-pressure wind profiles in which there were at least two levels with a wind speed were written out and analyzed.

During the initial test, 2889 profiles were found in which the wind speed was always zero. Of these profiles, one had eight levels with wind speed, 170 had four levels, 376 had three levels, and the rest had two levels. Approximately 1/3 of the cases were in India, and approximately 1/2 were from years between 1998 and 2006, when there were between 100 and 200 cases per year. These patterns suggested that most of these cases were not credible.

For the 3551 additional profiles with a maximum wind speed of 0.5 m/s, the frequency distribution of cases through time was more aligned with the frequency distribution of observations/stations through time than in the case of zero maximum wind speed. Approximately 1/3 of the cases were found in profiles that did not reach higher than 700 hPa and therefore might be more likely to have low wind speeds than higher-reaching profiles. It was therefore less obvious than in the case of zero maximum wind speed that these profiles

should be removed.

For profiles with high minimum wind speed, five randomly-chosen profiles were examined in each of several threshold categories. Within the 128 profiles with minimum wind speed greater than 100 m/s, most were partial or incomplete profiles with only two wind levels. The same was true for the additional 178 profiles with the minimum wind speed between 70 and 100 m/s. For thresholds between 60 and 70 m/s, there were an additional 237 profiles. Among 10 cases evaluated in this category, two looked potentially possible. Therefore, it was decided to set the threshold for this check at 60 m/s.

On the basis of these findings, it was decided to implement the versions of the checks that removed profiles with a minimum wind speed of at least 60 m/s as well as profiles with a maximum wind speed of 0 m/s. The checks were applied to pressure and non-pressure levels separately and required that at least two levels with wind speed be present in order for a pressure or non-pressure profile to be checked. In the final implementation, the checks flagged a total of 3081 profiles, 2570 due to excessive minimum wind speed and 611 due to insufficient maximum speed.

3.6.6.4 Other Enhancements Considered

The following types of checks were tested for inclusion in the IGRA 2 QA system, but were not pursued for a variety of reasons.

- Lowering the threshold for the absolute limits check on wind speed from 150 m/s to as low as 120 m/s: Several of the 22,370 pressure-level and non-pressure-level profiles with a maximum wind speed of at least 120 m/s, but less than 150 m/s, were examined for reasonableness. One of these profiles was smooth, with a peak wind speed at 279 hPa and therefore did not seem implausible. It was therefore decided not to lower the threshold to 120 m/s. An evaluation of other thresholds between 120 and 150 m/s would require some independent information and would likely result in little gain relative to the 150 m/s threshold and was therefore not pursued. The threshold of 150 m/s was retained.
- Applying the temperature vertical runs check (Durre et al. 2006) to wind speed: This check operates on standard and other pressure levels separately and looks for repetitions of the same value across at least four consecutive levels. Applying this procedure to wind speed yielded approximately 60,000 cases with such runs, 46,000 of which had exactly four levels in the run. Many of the cases were soundings in which wind speed was given to the nearest whole meter per second, making a run more likely than in higher-precision measurements. Many of the soundings were from the tropics, the Arctic, or Continental locations where conditions could be relatively calm under certain circumstances. In many cases, wind speeds at levels not flagged by this check were consistent with the speeds in the run and/or were increasing above the run. Given all of these findings, the check would do more harm than good by flagging too many realistic cases.

3.6.7 New Checks on Elapsed Time

For the newly added variable of elapsed time, three basic checks were implemented. First, lower and upper limits of 0 and 240 minutes for elapsed time were added to the basic plausibility checks. The upper threshold of 240 minutes was chosen since soundings can last for more than three hours (Blackmore, personal communication), yet the frequency distribution of elapsed time longer than 240 minutes looks unreasonable. This check flagged 4305 out of the 282.6 million elapsed times in IGRA 2. A second procedure looked for cases in which the same elapsed time appeared at two levels whose geopotential heights differed by 10 m or less. If one of the two levels was a mandatory pressure level, surface level, tropopause, or freezing level, the other level was removed. Finally, the logic of the height sequence check (Durre et al. 2006) was applied to elapsed time, thereby removing levels whose elapsed time was out of order with respect to pressure at pressure levels or with respect to height at non-pressure levels.

Out of the 282.6 million elapsed times in the integrated IGRA 2 dataset, the basic check removed 4305 values, less than 0.0002%. The check for identical elapsed times and the out-of-sequence check removed 0.02% and 0.03% of the values, respectively.

3.6.8 Handling of Mobile Stations

New to IGRA in Version 2 are 95 mobile stations, including ships and ice islands. These stations pose special QA challenges since, by definition, their positions change from sounding to sounding. This means that their data are not suitable for any check that is based on a station's climatology or time series and that, like all data, their coordinates are prone to random data processing problems. Consequently, conditions were inserted into the QA programs that exclude mobile stations from checks for climatological outliers, climatology-based vertical temperature inconsistencies, temporal runs, and temporal inconsistencies. All single-sounding checks, however, apply.

A new series of checks was devised with the intent of removing gross errors from the positions of mobile stations. These checks were designed to be performed in a semiautomatic fashion after the single-sounding checks of QA1 had been applied. The output from the position checks was a list of soundings whose positions were considered to be erroneous. That list was then provided to the QA2 program, which removed the identified soundings from the data.

For the position checks, a ship's speed and direction were computed from the sequence of positions. A velocity vector for a current time was based on its current position and the next future position. Then, the following steps were taken.

1. Land based location check. A check that the ship’s location was reasonably near a coast or open water. Using a 0.25 degree land water mask from http://ldas.gsfc.nasa.gov/gldas/GLDASvegetation.php, the mask value for the nearest grid point was checked. If zero (water), the point passed. If one (land),

the points in the 5x5 grid points box centered on the nearest point were also checked for water. Not counting the outer corners of this square, there were 21 possible grid points. If 19 or more were land, the position was determined to be on land and flagged as such.

This technique flagged a large number of positions near certain port cities that actually seemed like reasonable locations. Consequently, ship coordinates within 50 km of these eight "ports" were allowed as not being on land. The ports are: Paramaribo, Turbo, Aalborg, Hamburg, Antwerp, Rotterdam, Baltimore, Kangerlussuaq, and Staten Island. After this revision, 243 positions were flagged.
2. Spike check. One typical position error involved one position jumping away from the rest of the series, resulting in two large consecutive speed values. When two consecutive speeds were greater than 30 knots, the position of the second speed was flagged. A total of 197 values were flagged with this check.
3. Stop and go check. While stationary ships were acceptable, it seemed unnatural that a ship would move along at a speed greater than 5 kts for a period of 6 or 12 hours, stop, move again, and then stop again all at equal time intervals (typically 6 or 12 hours). When this condition was detected, the position at the end of the first stop, a repeat of its previous coordinate, was flagged, leading to a total of 123 flags.
4. Rapid acceleration check. Similar to the stop and go check, this test identified cases where a very small velocity (less than 0.1 kts) was followed by a speed greater than 30 kts. Such a situation was usually the result of repeated coordinates. In looking at the neighboring speeds, often 15-20 kts, it was judged that the point at the beginning of the large speed was in error and thus should be flagged. The coordinate sets were passed through this test twice with 120 observations flagged in the first run and 2 in the second. All in all, 122 values were flagged with this check.
5. Manual excess speed check. Conditions in which the speed exceeded 30 kts were visually examined, and a determination was made typically to flag the current or next point in the series for conditions that were similar to, but not captured by the previous tests. For example, if a spike were made up of two consecutive points instead of one, the spike test would not have identified it. Frequently through this manual examination, some larger underlying cause of a problem led to the manual editing of the ship coordinate file to flag blocks of observations. This check was also run twice, flagging 152 in the first run and 8 in the second. In addition, 92 values were manually flagged for other inconsistencies.

At each step, the previous checks starting with the spike check were performed again. This resulted in the flagging of 5 more spikes and 1 more rapid acceleration case.

The results of these checks were written into station-by-station files. Each file contained relevant portions of all sounding headers extracted from the station's output data file from QA1, the program performing basic Q&A checks. A numeric flag was appended to each header, indicating the result of the semiautomatic ship position checks. The flag codes used include 0 when no problem is detected, 1 through 5 for each of the five checks described above,

and 8 for additional manual flags as described in check #5 above.

The QA2 program then removed all soundings with a flag other than 0 or 1 as long as it still had the same coordinates in the data file as it had at the time of the ship track QC. With this process, a total of 780 soundings were removed from the data ending in 2013.

3.7 Compositing

Once all data integration and quality assurance checks had been completed, a compositing algorithm was applied in order to combine the records of any stations that were located close to each other, whose data records did not overlap, and that did not get combined as part of the other integration steps. Such stations were kept separate during the integration process if their names and station IDs did not match, and a data comparison was not possible due to the lack of record overlap.

To provide the longest records representative of a location in IGRA , stations within 5 km of each other that did not have observations at the same time were composited. Additionally, Canadian stations for which the first two digits of the WMO number changed from 72 or 74 to 71 in 1977 were composited. For example, CAM00072877 and CAM00071877 form a single record for Calgary, Alberta.

The 5 km threshold was chosen after intensive evaluation of various distance thresholds between 0 and 10 km. It was found that a threshold of 10 km would result in the combination of stations that should remain separate, while a threshold of 1 km would leave too many stations separate.

After the initial run of the compositing procedure, station pairs that appeared to be candidates for compositing, but whose periods of record overlapped, were investigated manually to determine why they did not get matched and whether they should be combined, whether they should stay separate, or whether one or both should be removed. These generally were cases in which different observation schedules were followed at the two stations. Based on these investigations, a white list and two black lists were created for use during subsequent executions of the compositing algorithm. One pair of stations that were to be composited with each other was added to a white list. Ten pairs of stations that were to be kept separate from each other were placed on a pairs blacklist, and six single stations that were not to be composited with any station were placed on a black list for single stations.

3.8 Results for IGRA 2.0

Once all corrections and quarantining actions had been taken, the entire sequence of station matching, integration, QA, and compositing steps was re-executed on the data for the full POR through December 31, 2013, to ensure that all decisions were being incorporated. The final consolidated source station list contained 11,415 of the 17,687 source stations whose data had been standardized. Several hundred metadata edits were applied to those remaining

station records. During the station matching process, 389 stations with conflicting matching criteria were removed. The remaining 11,182 source stations were grouped into 3074 mingle groups which, together, contained a total of 68.4 million soundings.

During the data integration process (Section 3.5), 1.9 million soundings were removed due to insufficient data. Another 22.9 million were eliminated during the process of selecting soundings from various types of sounding clusters. When all was said and done, seven of the mingle groups did not have sufficient data to be retained as IGRA stations. The remaining 3067 stations contained a total of 43.7 million soundings.

Among the 3067 mingle groups for which data were integrated, 1260 consisted of only one source station, 360 of exactly two source stations, and the remaining 1447 of more than two source stations. The IGRA station with the largest number of source stations was RSM00027612 with 22 source stations.

Table 3.15 lists, for each element, the total number of values that were present prior to the application of the entire quality assurance system as well as the percentage of values removed by the QA procedures. The removal of a value was caused either by its failure to pass one of the element-specific QA tests that applied to it or by a test that triggers removal of certain other elements at the same level, of the entire level, or of the entire sounding (Table 3.12). Therefore, the last column in Table 3.15 shows the percentage of the removed values that can be attributed to "direct" checks specific to the particular element.

Table 3.15: QA flag rates by element

Element	# values before QA	% removed by QA	% of removals by direct checks
Temperature	785005150	0.57%	30.6%
Dew point depression	468618624	0.54%	2.0%
Relative humidity	219013829	0.53%	0.9%
Wind speed	1.005E+09	0.79%	7.6%
Wind direction	1.003E+09	0.57%	3.8%
Pressure	1.049E+09	0.42%	32.6%
Geopotential height	973182221	0.77%	94.1%
Elapsed time	282780869	0.06%	78.1%

Geopotential height and wind speed experienced the highest removal rates, 0.77% and 0.79%, respectively. However, in the case of geopotential height, more than three fourths of the removals were caused by element-specific checks, while most of the wind speed values were removed when a procedure involving other elements, particularly geopotential height, deleted an entire level from a sounding. The element with the lowest removal rate was elapsed time,

with only 0.06% of values deleted. The removal rates for all other elements were around 0.5%. In general, these flag rates were consistent with those of other robust, automated QA systems (e.g., Durre et al. 2010).

In the final application of the compositing algorithm to the full dataset through the year 2013, 130 compositing operations took place, involving 124 pairs and three triads of stations. As a result, the 2877 stations that have survived the QA process were reduced to 2747 stations, 130 of which represented composites.

The post-compositing dataset constitutes the inaugural static POR version of IGRA 2 to which data are then added on a daily basis with the IGRA 2 update system described in Section 3.9. The updated period-of-record version of IGRA 2 forms the basis for the derived products described in Section 4.

3.9 Updates

3.9.1 Overview

Two mechanisms exist for updating and maintaining the static POR version whose creation was described in Sections 3.2-3.8. Data from days after the end of the POR version are added once a day by the IGRA 2 update system. In addition, periodically, the static POR version is re-created and extended to the end of the last calendar year.

3.9.2 Daily Updates

During a daily update, newly arrived data from sources ncdc-gts and ncdc-nws (Table 3.2), the two sources for which daily updates are readily available at NCEI, are ingested and processed together with all of the data from these two sources that have been collected since the end of the last static POR version. Typically, the newly incorporated data primarily include observations from the previous day, but some late-arriving observations from earlier days as well as some measurements from the current day are also included. In IGRAv2.2, the ncdc-gts data are now supplemented with BUFR data as described in Section 1.4.4. Data from the ncdc-nws and augmented ncdc-gts are reformatted and standardized as described in Section 3.2, then integrated using the mingle list that was created when the static POR version was last reprocessed. Next, all basic and climatological QA procedures are employed, using the climatologies created during the last reprocessing cycle. Checks for temporal runs, temporal outliers, errors in ship positions, and isolated soundings are not applied during daily updates because they work best when the entire time series are available. The quality-assured data extending back to January 1 of the year after the static POR version, referred to as the year-to-date files, are then appended to the static POR files to create the updated POR files. Both the year-to-date files and the updated POR files are uploaded to NCEI's public FTP site.

The entire process runs fully automatically as a cron job, starting in the late afternoon and typically finishing in the early morning hours of the next day. Various system messages are sent to those monitoring the process.

3.9.3 Rebuilding the POR Version

The purpose of periodically rebuilding the static POR version is fourfold: to incorporate newly opened stations; to apply the full suite of QA procedures to the recent data that had previously only been processed by the update system; to update the climatologies used by the QA checks in the update system; and to reduce the amount of data processed by the update system. Ideally, the reprocessing will take place during the first quarter of each calendar year. The same reprocessing steps are necessary when new data sources or updated versions of existing data sources are to be incorporated or when the processing software has been modified. Although a fully automated process exists for re-creating the static POR version from the reformatted data sources, the tasks of collecting and reformatting data sources and of verifying the output from the rebuilding system require human attention. The required steps are described in the supplement of the IGRA Operations Document.

4. Derived Products

4.1 Monthly Means

4.1.1 Available Files

Monthly means of geopotential height, temperature, vapor pressure, and zonal and meridional wind components are provided at the surface and standard pressure levels for the nominal times of 0000 and 1200 UTC. Accompanying each monthly mean is the number of values used to compute the mean. All means and numbers of values used for one variable and time are stored in a single file. For example, file ghgt_00z-mly.txt contains the monthly means of geopotential height for 0000 UTC at all stations.

For each variable and time, two time periods are provided: the full POR and the last available month only. Both sets of files are fully re-created on the sixth day of each month to include data through the end of the previous month. In general, the means for all but the latest one or two months are expected to remain the same from month to month. However, after a full reprocessing of the sounding data, monthly means for earlier months may also be modified, and means for newly added stations may appear.

4.1.2 Processing

A minimum of 10 values is required for the calculation of the mean for any particular station, year, month, variable, and nominal time. In the computation, soundings from up to two hours before and two hours after the nominal time are taken into account; if more than one value is found within that window for a particular date and time, the one closest in time to the nominal hour is used; If two values are equally close, the earlier value is used. Given that sufficient pressure-level data around the nominal times are not available for all stations and months for which observations are present in the sounding data, monthly means are not available for all IGRA stations and are not always provided for a station's full POR. In addition, monthly means are not computed for mobile stations because such a mean would represent an average of conditions at many different locations during the month.

For geopotential height and temperature, monthly means are computed directly from the variable's values provided in the sounding data files. The monthly mean of vapor pressure is calculated by first converting the individual dew point depression or RH values to vapor pressure, using the procedure described in Durre and Yin (2008), and then averaging the resulting vapor pressure values. The monthly means of the zonal and meridional wind components (u and v) represent averages of individual u and v obtained from the observed wind direction (D) and wind speed (V) using the following equations:

$$u = - V * \sin(D*\pi/180.0) \qquad (4.1a)$$
$$v = - V * \cos(D*\pi/180.0) \qquad (4.1b)$$

The resulting u component is positive when wind is blowing out of the West (at

270°), while v is positive when wind is coming from the south (at 180°).

4.1.3 Updates

The monthly means are recalculated once a month on the sixth day of the month. This recalculation takes place as part of the regular IGRA update system. For simplicity, all monthly means are recalculated, and monthly mean files for both the full POR and the latest month are uploaded and archived.

4.2 RATPAC

4.2.1 General Information

The temporal inhomogeneity that results from changes in instruments and measurement practices over time renders many radiosonde time series unsuitable for the study of long-term climate variations. The radiosonde-based temperature products known as RATPAC consist of time series that have been processed in such a way as to reduce the influence of these inhomogeneities. Produced through a collaboration among NOAA scientists at ARL, GFDL, and NCEI, the time series are based on observations from 85 stations located around the world. Data are available on 13 atmospheric pressure levels and as means over certain atmospheric layers. Where available, data begin in 1958 and extend through the present. Beginning with IGRA 2, the RATPAC time series are produced monthly as part of the IGRA update system.

The RATPAC data can be found at https://www.ncei.noaa.gov/pub/data/ratpac/. See the RATPAC readme.txt file for the names of the various RATPAC files and a description of the format of each of those files.

4.2.2 Available Products

The directory structure and formats of the various files are described in the readme.txt file available in the IGRA FTP directory (see Section 4.1). All RATPAC time series are recreated monthly on the sixth day of the month as part of the IGRA 2 update system in order to incorporate data from the latest complete month. Two distinct sets of products, RATPAC-A and RATPAC-B, are available (Free et al. 2005). Both are based on the Lanzante/Klein/Seidel (LKS) bias-adjusted temperature data (Lanzante et al. 2003a,b) and on IGRA, but they were derived using different approaches. The RATPAC-A time series are recommended for analyses of interannual and longer-term changes in global, hemispheric, and tropical seasonal means, since they contain more robust large-scale averages than RATPAC-B. For individual station data, monthly data, or regional means on smaller scales, use of RATPAC-B is recommended, with careful attention to the potential of inhomogeneities influencing analyses after 1997. For analyses that require data not included in RATPAC and that are less sensitive to long-term biases, users are directed to the IGRA sounding data and monthly means.

RATPAC-A contains adjusted global, hemispheric, tropical, and extratropical mean temperature anomalies. From 1958 through 1995, the bases of the data are spatial averages of LKS adjusted 87-station temperature data. After 1995, they are based on the IGRA station data, combined using a first difference method (Free et al. 2004). RATPAC-A time series are available for seven large regions: the Globe, the Northern and Southern Hemispheres, the tropics (30°N-30°S), 20°N-20°S, and the northern and southern extratropics. Two types of regional time series are provided for these regions in two separate files: annual mean temperature anomalies for 13 pressure levels from the surface to 30 hPa (surface, 850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, and 30 hPa) as well as seasonal mean temperature anomalies for the three atmospheric layers of 850-300 hPa, 300-100 hPa, and 100-50 hPa.

RATPAC-B contains data for individual stations, for large-scale arithmetic averages for the seven regions used for RATPAC-A, and for the additional zonal bands of 30°-60° and 60°-90° in each hemisphere. The station data consist of adjusted data produced by LKS for the period 1958–1997 and unadjusted data from IGRA 2 after 1997. The regional mean time series in RATPAC-B are based on arithmetic averaging of these station data, rather than on the first difference method used to create RATPAC-A. For individual stations, monthly mean temperature anomalies are organized into three files, one for 0000 UTC, one for 1200 UTC, and one for the averages of 0000 UTC and 1200 UTC monthly means. On a regional basis, temperature time series are provided in the form of annual mean anomalies, which are stored together in one file. Both the station and regional RATPAC-B series are provided at the same 13 pressure levels used in RATPAC-A.

4.2.3 RATPAC Stations

RATPAC uses data from 85 of the 87 LKS stations. Because of problems with the data from India, RATPAC does not include the LKS station data for Bombay or Calcutta (Free et al. 2005). In addition, one of the LKS stations, Pechora, Russia, underwent a change in WMO number from 23418 to 23415 in the year 2000, and therefore is now identified by WMO# 23415 in RATPAC. The RATPAC stations are listed in the file ratpac-stations.txt.

Not all 85 stations contribute to RATPAC in every month and at every level. In particular, four of the possible 85 stations (Preobrazheniya, Chetyrekhstolbov, Mould Bay, and Ashabad) have no data after 1997, the last year of LKS data. Several others have closed since then.

4.2.4 Methodology

4.2.4.1 LKS Component

The LKS time series consist of Monthly means of temperature at 16 atmospheric pressure levels between the surface and 10 hPa for 87 carefully selected stations. The data were taken from CARDS data (Eskridge et al. 1995)

and were adjusted using a multifactor expert analysis by a team of three climate scientists. The team visually examined time series of temperatures at multiple levels, night-day temperature differences, temperatures predicted from regression relationships, and temperatures at other nearby stations. They also considered metadata, statistical change points, the Southern Oscillation Index, and the dates of major volcanic eruptions. Using these indicators, they identified artificial change points and remedied them by either adjusting the time series at each affected level or, if adjustment was not feasible, by deleting data.

The 10 and 20 hPa levels, available in the LKS data, were not used in the RATPAC products because of the scarcity of data at those levels. In addition, the 1000-hPa level was not included because LKS found the 1000 hPa data to be more erratic and less reliable than other levels in the troposphere, probably due to problems arising from days when the surface pressure was less than 1000 hPa (Lanzante et al. 2003b).

4.2.4.2 RATPAC-A Methodology

For RATPAC-A, the first difference ("FD") procedure was used to update the LKS data (Free et al. 2004). Since this method introduces a random error that increases with the number of time gaps in the data and with decreasing number of stations, reliable time series can be produced only for large regions.

The pre-1996 portions of the RATPAC-A time series represent area means of the adjusted LKS station data, without use of FD. Although the LKS dataset runs through 1997, IGRA data was substituted for 1996 and 1997 because the short record left after the adjustments makes LKS adjustments in 1996 and 1997 less reliable than those at earlier times.

The FD method was applied to the IGRA monthly means starting in 1996. In this method, the difference in temperature between one time step and the next (the "first difference") was taken, then large-scale means of the FD series were computed, and finally large-scale temperature series were reconstructed from the FD series (see Appendix of Free et al. (2005) for details). Prior to the application of the FD method, portions of the station time series around the times of known changes in instruments or procedures were eliminated from the IGRA data in an attempt to reduce the effect of inhomogeneities due to such changes (see Free et al. 2005 for details).

Finally, an endpoint outlier trimming procedure was used to reduce the random errors introduced by the FD procedure. As described in Peterson et al. (1998) and Free et al. (2004), this procedure removes data exceeding a prescribed multiple of the STD of the original time series if the data fall at the end of a data segment (immediately before or after a gap). Here, the multiple, or trim factor, chosen was 1.0 STD. This choice was based on sensitivity tests with reanalysis data as well as with the LKS data (Free et al. 2005).

4.2.4.3 RATPAC-B Methodology

Because the FD method used for RATPAC-A does not allow production of individual station time series, and to provide alternative large-scale mean time series for comparison with the FD time series, a set of updated station time series were also created by appending monthly mean station data from IGRA for 1998-present to the corresponding adjusted LKS station time series for 1958-1997 without any adjustment for inhomogeneities after 1997. For consistency, only those observation times from the IGRA data were used that were present in the LKS adjusted station data. The 00 UTC and 12 UTC observations were combined where both were available, and the updated station time series were also available for the two observation times separately.

To minimize the discontinuity at 1997/1998, a factor equal to the difference between the means of the IGRA and LKS data for 1996-1997 was added to the IGRA monthly means. The effect was to shift the IGRA data so that the means of the two datasets for the last two years of the LKS time series were equal. If both time series were present for fewer than 9 months in those two years, the time period 1990-1997 was used instead.

At one or more levels at ~14 stations, LKS found a discontinuity, but deleted data after the discontinuity rather than adjusting it because adjustment was not feasible. In creating RATPAC-B, the IGRA data for those levels at those stations were not appended after 1997 to avoid reintroducing known inhomogeneities. (These IGRA data were used in the creation of RATPAC-A, however, since the FD procedure was expected to deal with the discontinuity.)

4.2.4.4 Spatial Averaging

In an effort to obtain spatially unbiased large-scale means, the uneven longitudinal distribution of stations was compensated for by creating regional means before averaging data into zonal bands. Each 30-degree zonal band was divided into three longitudinal regions of 120 degrees each: 30°W to 90°E, 90°E to 150°W and 150°W to 30°W. Hemispheric (0-90°), tropical (30°S-30°N), and extratropical (30-90°) means were calculated from these zonal means, area-weighted using the cosine of the latitude of the midpoint of the zone. The global mean was the average of the hemispheric means.

4.2.5 Updates

The RATPAC time series are recalculated once a month on the sixth day of the month. This recalculation takes place immediately after the computation of the monthly means as part of the regular IGRA update system. After each recalculation, the recreated time series are uploaded to the public FTP site and archived.

4.3 Sounding-Derived Parameters

4.3.1 Overview

To facilitate studies of variations in the vertical structure of the atmosphere, the collection of IGRA data and products includes a set of relevant parameters derived from the IGRA soundings. The derived quantities are available at the subset of IGRA stations at which at least 100 soundings contain temperature, pressure, and height at the surface and temperature at least one other pressure level.

4.3.2 Organization of the Files

The derived parameters are organized into station files. Each file is identified by the 11-character IGRA station ID in its filename and contains all soundings for which derived parameters are available during the POR of the station. Like the IGRA observations, the derived parameters are updated once a day.

Within each file, soundings are sorted chronologically from earliest to latest. Like in the data files, each derived-parameter sounding consists of a header record, followed by a number of data records. The header record contains the station identifier, date, time, number of levels, and a variety of stability indices and other indicators that apply to the entire sounding. The data records correspond to pressure levels at which temperature is available and contain several observed and derived variables. Levels and soundings without any temperature are not included. Unlike the data soundings, derived-parameter soundings always contain a surface pressure level, and that level is the first level after the header record.

4.3.3 Variables

The derived parameters provided at each level include potential temperature, virtual temperature, virtual potential temperature, RH, actual and saturation vapor pressure, zonal and meridional wind components, the refractivity index, as well as vertical gradients of all of these variables (Durre and Yin 2008, 2011). In addition to the reported geopotential height and RH, hydrostatically-estimated geopotential heights and RH values derived from temperature, dew point depression, and pressure are included to fill in gaps in the original reports. For the sounding as a whole, surface-to-500-hPa PW, six stability indices, and the pressure and height of six special levels are available (Durre and Yin 2011).

4.3.4 Methodology

4.3.4.1 General

The methods for calculating the various parameters were chosen based on techniques that had been vetted in the peer-reviewed literature and take into account each method's sensitivity to the varying vertical resolution of the

sounding data and applicability to a wide range of atmospheric conditions. The computation of potential temperature, zonal, and meridional wind components, vertical gradients, and geopotential height were documented in Durre and Yin (2008). Actual and saturation vapor pressure, derived RH, and PW were calculated using the approach described in Durre et al. (2009), except that the 1000-hPa level was not included when calculating PW due to the often spurious availability and accuracy of data at that level. The calculations of the other parameters are described below.

4.3.4.2 Lifting Condensation Level, Level of Free Convection, and Equilibrium Level

When considering convective instability, three points along the trajectory of a rising air parcel are of particular importance:

Lifting condensation level (LCL): the level where a parcel lifted dry adiabatically from the surface first reaches saturation.

Level of free convection (LFC): the level at which the temperature of the environment decreases faster than the moist adiabatic lapse rate of a saturated air parcel at the same level.

Equilibrium level (ELL): the level at which an air parcel, rising or descending adiabatically, attains the same density as its environment.

For the IGRA-derived product, the heights and pressures of each of these levels are calculated as follows and included in every sounding in which they can be calculated. First, the LCL temperature (T_LCL) is calculated using the method given in Bolton (1980). Then, using the adiabatic lapse rate, the height of the LCL (Hlcl) was derived, and the corresponding pressure (p_LCL) was computed as the inverse of potential temperature.

$$p\_LCL = p\_s * (T\_LCL/T\_s)^{3.498} \quad (4.2)$$

Where p_s and T_s are the pressure and temperature at the surface.

Next, the pseudoadiabatic lapse rate is calculated based on T_LCL following Holton (2004). Moving from the LCL upward, the LFC then is the level at which the Air Parcel temperature becomes warmer than the temperature of the environment (T_env). Similarly, the EL is the level above the LFC at which the parcel first becomes colder than the environment.

When the LCL, LFC, or EL lies very close to a sounding level, consistency between the height and pressure of the computed and reported levels is ensured. This is done in order to avoid cases in which, for example, the computed height is above the height of the reported level, yet the computed pressure is closer to the surface than the reported level. In such cases, the pressure of the computed level is instead estimated by means of simple linear interpolation, using the height of the computed level and the heights and pressures of the reported levels above and below it. Simple linear interpolation is used in lieu of the hydrostatic equation because this situation only arises

when the computed level is extremely close to the reported level, e.g., within less than 1 m.

4.3.4.3 Convective Inhibition and Convective Available Potential Energy

Convective inhibition (CIN) is the energy needed to lift an air parcel vertically and pseudo-adiabatically from its originating level to its LFC, while convective available potential energy (CAPE) is equal to the amount of energy available to a parcel as it freely rises between the LFC and EL. CIN is a numerical measure of the strength of "capping". Values near zero imply that the capping is easily overcome by an air parcel, often resulting in Fair Weather Cumulus. For higher values of CIN, the capping can only be overcome by a forced mechanism which can lead to strong thunderstorms. CIN values exceeding 200 J/kg typically imply that the capping cannot be overcome by any forcing mechanism, and no thunderstorms can form. CAPE, on the other hand, is considered to be an indicator of the potential strength of updrafts within a thunderstorm (Bluestein 1993). Therefore, higher values indicate greater potential for severe weather. Values often exceed 1000 joules per kilogram (J/kg) in thunderstorm environments and can exceed 5000 J/kg in extreme cases.

On a thermodynamic diagram of a sounding, CIN represents the cumulative effect of atmospheric layers that are warmer than the parcel moving vertically along the adiabat between the LCL and LFC. CAPE is represented by the area enclosed between the environmental temperature profile and the moist adiabatic path of a rising air parcel over the layer within which the latter was warmer than the former, i.e., between the LFC and EL.

To find CAPE, a CAPE value is first calculated for each layer between the LFC and EL, and then the CAPE values for the individual layers are summed to obtain the total CAPE. The lowest relevant layer is bounded by the LFC at the bottom and the next highest reported level in the sounding. The top layer is bounded by the EL at the top and the reported level immediately below the EL. The boundaries of the layers in between the top and bottom layers are defined by each pair of consecutive reported levels between the LFC and EL. Each layer's CAPE is calculated as follows: First, the Air Parcel's temperatures at the top and bottom of the layer are computed based on the pseudo-adiabatic lapse rate. Then, the average layer temperatures of the parcel (T_pa) and the environment (T_en) are calculated by averaging the respective temperatures at the top and bottom of the layer. The layer's CAPE is then computed according to the following formula:

$$CAPE = gH(T_{pa}-T_{ev})/T_{en} \quad (4.3)$$

Where g = 9.80665 m/s^2 is the gravitational acceleration and H is the thickness of the layer.

CIN is determined in a similar manner, using the layers between the LCL and LFC.

4.3.4.4 Other Indices of Convective Instability

Convective stability indices are designed to indicate the potential of thunderstorms. A variety of indices are being used by weather forecasters, each having its own strengths and weaknesses. Four common indices whose variations and trends have previously been analyzed are included in the IGRA-derived sounding parameters: the Lifted Index, the Showalter Index, the K Index, and the Total Totals Index (Derubertis 2006). Results were verified by comparing them to those available from the Storm Prediction Center at http://www.spc.noaa.gov/exper/soundings/ for some randomly selected soundings.

4.3.4.4.1 Lifted Index and Showalter Index

The LI and SWI are defined as the difference between a sounding's 500 hPa temperature, representing the temperature of the environment, and the temperature of a parcel lifted adiabatically to 500 hPa. In the case of the LI, the parcel is assumed to be listed from at or near the surface (Galway 1956); for the SWI, it is lifted from 850 hPa (Showalter 1947). Unlike the LI, the SWI does not take into account diurnal heating or moisture below 850 hPa and therefore must be used with caution. However, the SWI is useful in situations in which a shallow cool air mass below 850 hPa conceals greater convective potential aloft.

For both indices, the temperature at the LCL is estimated using the method of Bolton (1980), and the pseudo-adiabatic lapse rate is computed following Holton (2004). The latent heat of condensation that is required in the computation of pseudo-adiabatic lapse rate is estimated using a formula provided by Henderson-Sellers (1984). For both indices, negative index values indicate instability, i.e., that the lifted parcel is warmer than the environment; the more negative the values, the more likely the occurrence of strong thunderstorms.

4.3.4.4.2 K Index

The K Index measures thunderstorm potential in terms of vertical temperature lapse rate, moisture content of the lower atmosphere, and the vertical extent of the moist layer. Unlike the LI and SWI, the K Index is determined arithmetically from measurements of temperature and dew point temperature (George 1960):

$$K = T850-T500 + Td850-(T700-Td700), \qquad (4.4)$$

where T850 is the temperature at the 850 mb level, T500 is the temperature at the 500 mb level, Td850 is the dew point temperature (°C) at the 850 mb level, T700 is the temperature (°C) at the 700 mb level, and Td700 is the dew point temperature (°C) at the 700 mb level.

In the above formula, the temperature difference between the 850 and 500-hPa levels is used to parameterize the vertical temperature lapse rate, the 850-hPa dew point serves as an indicator of the moisture content of the lower atmosphere, and the vertical extent of the moist layer is represented by the 700 hPa dew point depression.

Higher values of K indicate a greater probability of thunderstorms. It is generally assumed that no thunderstorms will occur when K is less than 15°C and that thunderstorms are a near certainty when K exceeds 40°C.

4.3.4.4.3 Total Totals Index

The Total Totals Index (TT) is an indicator of low-level moisture and 850-to-500-hPa lapse rate (Miller 1972). It represents the arithmetic sum of two other indices: the Vertical Totals Index (temperature at 850 hPa minus temperature at 500 hPa) and the Cross Totals Index (dew point at 850 hPa minus temperature at 500 hPa).

$$TT = T850 + Td850 - 2T500$$

Typically, values of less than 50 or greater than 55 are considered weak and strong indicators, respectively, of potential severe storm development.

4.3.4.5 Virtual Temperature and Virtual Potential Temperature

Virtual temperature and virtual potential temperature are provided at every level at which both temperature and either dew point depression or relative humidity are available. The virtual temperature (T_v) at a particular level is calculated using the following formula:

$$T_v = t/(1-0.378*e/p), \quad (4.5)$$

Where T is temperature, e is the partial pressure of water vapor, and p is the atmospheric pressure.

To calculate virtual potential temperature, the potential temperature is substituted for temperature in the above formula.

4.3.4.6 Mixing Height

The mixing height is the height to which a parcel of air, or a column of smoke, rises, mixes, or disperses. Here, the parcel method is employed to calculate this height, using virtual potential temperature in order to account for the buoyancy effects of water vapor. According to this method, the mixing height is the lowest level at which the virtual potential temperature is greater than or equal to the virtual potential temperature at the surface (Morris et al. 1990). In the rare case in which this results in a mixing height that is more than 5000 m above the surface, the mixing height is set to missing.

4.3.4.7 Freezing Level

The freezing level is the lowest level at which temperature reaches 0°C. It can either be a reported level, or it is interpolated between the lowest pair of levels at which one temperature is above freezing and the other is below.

4.3.4.8 Inversion Top

The level of the warmest temperature in the sounding is used as a proxy for the top of a near-surface temperature inversion layer, when present. If the warmest temperature is found at the surface, No surface-based inversion is assumed to be present. Otherwise, the height and pressure of the level at which the warmest temperature occurs are taken to be the top of the inversion. Heights above 5000 m are considered to be unrealistic and are set to the missing value code.

To ensure the presence of significant thermodynamic levels in the sounding, the inversion top is determined only in soundings that have at least 10 levels above the surface and below a height of 5000 m above the surface. Since significant thermodynamic levels are defined as levels at which the slope of the temperature profile changes, one such level should be reported at the point when the temperature reaches its maximum. As a result, the warm-point technique for identifying the top of an inversion is less sensitive to vertical resolution than corresponding indicators based on changes in lapse rate and, therefore, is more robust when used in soundings whose vertical resolution varies greatly.

4.3.4.9 Refractive Index

The Atmospheric Refractive Index (N) is defined as the ratio of the speed of light in a vacuum to the speed of light in air (Smith and Weintraub 1953). It is a function of temperature, pressure, and water vapor content and therefore varies with height in the atmosphere. Because satellite-based GPS observations of the refractive index are used to infer temperature and water vapor, values of N computed from radiosonde data are useful in the validation of the GPS observations.

N is calculated at each level using the following formula (Bean and Dutton 1968):

$$N = 77.6*P/T + 373256*e/(T^2), \qquad (4.6)$$

Where T is temperature, p is pressure, and e is the partial pressure of water vapor.

4.3.5 Updates

The IGRA-derived sounding parameters are recalculated for the entire POR each day. This recalculation takes place as part of the IGRA update system after the IGRA sounding data have been updated. Files are uploaded to public FTP and archived on a daily basis.

5. References

Air Force Weather Agency/U.S. Air Force/U.S. Department of Defense, 1980: U.S. AFGWC Station (Surface and Upper Air) Library. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. [Available online at http://rda.ucar.edu/datasets/ds900.0/]

Bean, B. R., and E. J. Dutton, 1968: Radar Meteorology. Dover Publications, Inc., New York.

Bluestein, H. B., 1993: Synoptic-Dynamic Meteorology in Midlatitudes. Vol. 2, Oxford University Press, 594 pp.

Bolton, D., 1980: The Computation of Equivalent Potential Temperature. Monthly Weather Review, 108, 1046-1053.

Dept. of Earth/Atmospheric/and Planetary Sciences/Massachusetts Institute of Technology and Soil Environmental Atmospheric Sciences/School of Natural Resources/University of Missouri-Columbia, 1980: M.I.T. Radiosondes, daily 1958May-1963Apr. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. [Available online at http://rda.ucar.edu/datasets/ds398.0/.] Accessed 17 Apr 2009.

Derubertis, D., 2006: RecCouple daysent trends in four common stability indices derived from U.S. radiosonde observations. Journal of Climate, 19, 309-323.

Dupigny-Giroux, L. A., T. F. Ross, J. D. Elms, R. Truesdell, and S. R. Doty, 2007: NOAA's Climate Database Modernization Program: Rescuing, archiving, and digitizing history. Bull. Amer. Meteor., Soc., 88, 1015-1017.

Durre, I., and X. Yin, 2008: Enhanced radiosonde data for studies of vertical structure. Bulletin of the American Meteorological Society, 89, 1257-1262, .

Durre, I., and X. Yin, 2011: Enhancements of the dataset of sounding parameters derived from the Integrated Global Radiosonde Archive. 23rd Conference on Climate Variability and Change, Seattle, WA, 25 January 2011. [Available online at https://ams.confex.com/ams/91Annual/webprogram/Paper179437.html.]

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. Journal of Climate, 19, 53-68, DOI:10.1175/JCLI3594.1.

Durre, I., R. S. Vose, and D. B. Wuertz, 2008: Robust automated quality assurance of radiosonde temperatures. Journal of Applied Meteorology and Climatology, 47, 2081-2095, DOI:10.1175/2008JAMC1809.1.

Durre, I., C. N. Williams, Jr., X. Yin, and R. S. Vose, 2009: Radiosonde-based

trends in precipitable water over the Northern Hemisphere: An update. Journal of Geophysical Research, 114, D05112, DOI:10.1029/2008JD010989.

Durre, I., X. Yin,, R. S. Vose, S. Applequist, and J. Arnfield, 2018: Enhancing the Data Coverage in the Integrated Global Radiosonde Archive, *J. Atmospheric Oceanic Technol.*, 35(9), 1753-1770, DOI:10.1175/JTECH-D-17-0223.1.

Durre, I., X. Yin, S. Applequist, J. Arnfield, and R. S. Vose, 2022a: Integrated Global Radiosonde Archive (IGRA) version 2. NOAA/National Centers for Environmental Information, DOI:10.7289/V5X63K0Q. [Last accessed 20 December 2022]

Durre, I., X. Yin, R. S. Vose, and S. Applequist, 2022b: Radiosonde Atmospheric Temperature Products for Assessing Climate, Version 2. NOAA National Centers for Environmental Information, DOI:10.7289/V5SF2T7J. [Last20 Dec 2022].

Elliott, W. P., and D. J. Gaffen, 1991: On the utility of radiosonde humidity archives for climate studies. Bull. Amer. Meteorol. Soc., 72, 1507-1520.

Elliott, W. P., and D. J. Gaffen, 1993: Effects of conversion algorithms on reported upper-air dew point depressions. Bull. Amer. Meteorol. Soc., 74, 1323-1325.

Elliott, W. P., R. J. Ross, and B. Schwartz, 1998: Effects on climate records of changes in National Weather Service humidity processing procedures. J. Climate, 11, 2424-2436.

Elliott, W. P., R. J. Ross, and W. Blackmore, 2002: Recent changes in NWS upper-air observations with emphasis on changes from VIZ to Vaisala radiosondes, Bull. Amer. Meteorol. Soc., 83, 1003-1017.

Eskridge, R. E., O.A. Alduchov, I.V. Chernykh, P. Zhai, A.C. Polansky, and S.R. Doty, 1995: A comprehensive aerological reference dataset (CARDS): Rough and systematic errors. Bull. Am. Meteor. Soc., 76, 1759-1775.

Free, M., J. K. Angell, I. Durre, J. R. Lanzante, T. C. Peterson, and D. J. Seidel, 2004: Using first differences to reduce inhomogeneity in radiosonde temperature datasets. Journal of Climate, 21, 4171-4179.

Free, M., D. J. Seidel, J. K. Angell, I. Durre, and T. C. Peterson, 2005: Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new dataset of large-area anomaly time series. Journal of Geophysical Research, doi: 10.1029/2005JD006169.

Gaffen, D. J., 1996: A digitized metadataset of global upper-air station histories. NOAA Technical Memorandum ERL ARL-211, National Oceanic and Atmospheric Administration, Silver Spring, Md, 37 pp.

Galway, J. G., 1956: The lifted index as a predictor of latent instability. Bull.

Amer. Meteor. Soc., 37, 528-529.

Garand, L., C. Grassotti, J. Hall, and G.L. Klein, 1992: On differences in radiosonde humidity-reporting practices and their implications for numerical weather prediction and remote sensing. Bull. Amer. Meteor. Soc., 73, 1417-1423.

George, J. J., 1960: Weather Forecasting for Aeronautics. Academic Press, 673 pp.

Henderson-Sellers, B., 1984: A new formula for latent heat of vaporization of water as a function of temperature. Q. J. Roy. Meteorol. Soc., 110, 1186- 1190.

Holton, J. R., 2004: An introduction to dynamic meteorology, 4th edition. Academic Press, 535 pp.

Kahl, J. D., M. C. Serreze, S. Shiotani, S. M. Skony, and R. C. Schnell, 1992: In situ meteorological sounding archives for arctic studies. Bull. Amer. Meteor. Soc., 73, 1824-1830.

Lanzante, J. R., S. A. Klein, and D. J. Seidel, 2003a: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. J. Climate, 16, 224-240.

Lanzante, J. R., S. A. Klein, and D. J. Seidel, 2003b: Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. J. Climate, 16, 241-262.

Miller, R. C., 1972: Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Center. Tech. Rept. 200(R), Headquarters, Air Weather Service, USAF, 190 pp.

Morris, R. E., T. C. Myers, E. L. Carr, M. C. Causley, and S. G. Douglas, 1990: User's Guide for the Urban Airshed Model, Volume II: User's Manual for the UAM (CB-IV) Modeling System. EPA-450/4-90/007b, U.S. Environmental Protection Agency, 504pp.

NCAR, 1971: Global Time Series Radiosonde Observations, daily 1948-con. Subset: C-Cards. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. [Available online at http://rda.ucar.edu/datasets/ds390.0/.] Accessed 17 Apr 2009.

NCEP, 1980: NCEP ADP Operational Global Upper Air Observations, December 1972 - February 2007. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO. [Available online at http://rda.ucar.edu/datasets/ds353.4/.] Accessed 24 Feb 2010.

Okamoto, N., M. D. Yamanaka, S. Ogino, H. Hashiguchi, N. Nishi, T.

Sribimawati, A. Numaguti, 2003: Seasonal variations of tropospheric wind over Indonesia: Comparison between collected operational rawinsonde data and NCEP Reanalysis for 1992-99. J. Meteorol. Soc. Japan, 81, 829-850. 829

Peterson, T.C., T.R. Karl, P.F. Jamason, R. Knight, and D.R. Easterling, 1998: First difference method: Maximizing station density for the calculation of long-term global temperature change. J. Geophys. Res., 103, 25,967-25,974.

Showalter, A. K., 1947: A stability index for forecasting thunderstorms. Bull. Amer. Meteor. Soc., 34, 250-252.

Smith, E. K., and S. Weintraub, 1953: The constants in the equation for atmospheric refractive index at radio frequencies. Proc. Inst. Radio Engrs., 41, 1035-1037.

Stickler, A., A. N. Grant, T. Ewen, T. F. Ross, R. S. Vose, J. Comeaux, P. Bessemoulin, K. Jylhä, W. K. Adam, P. Jeannet, A. Nagurny, A. M. Sterin, R. Allan, G. P. Compo, T. Griesser, and S. Brönnimann, 2010: The Comprehensive Historical Upper-Air Network. Bull. Amer. Meteor. Soc., 91, 741–751, doi:10.1175/2009BAMS2852.1.

Stickler, A., S. Brönnimann, S. Jourdain, E. Roucaute, A. M. Sterin, D. Nikolaev, M. A. Valente, R. Wartenburger, H. Hersbach, L. Ramella Pralungo, D. P. Dee, 2014: Description of the ERA-CLIM historical upper-air data. Earth System Science Data, 6, 29-48, doi:10.5194/essd-6-29-2014.

Wade, C. G., 1994: An evaluation of problems affecting the measurement of low RH on the United States radiosonde. J. Atmos. Oceanic Technol., 11, 687-700.