Lessons in Diversity: How 40 different data sources were combined to create Version 2 of the Integrated Global Radiosonde Archive

Paper 3.6

Imke Durre and Russell S. Vose

NOAA National Centers for Environmental Information, Asheville, North Carolina, USA

Xungang Yin

ERT, Inc., Asheville, North Carolina, USA

Integrated Global Radiosonde Archive (IGRA)

- Observations: radiosonde and pilot balloon
- Coverage: global land, 1905-present
- Example applications: reanalysis input, climate assessments, satellite verification, air pollution modeling
- Created by merging data from 40 sources containing 11,500 station records

The Classic Duplicate Elimination Problem

Which source station records should be combined to form one IGRA station?

➢Input:
- Multiple, sometimes overlapping, time series for the same location

➢Desired output:
- One time series per location, containing as much data as possible
- No data duplication between distinct locations

➢Challenges:
- Imprecise and changing station locations
- Various names and station identifiers for the same location
- Differences in data precision

The IGRA Solution

Decision-making algorithm:

- Input: all ~11,500 source stations

- Steps:
  1. Identify matching pairs of source stations on the basis of data and metadata.
  2. Arrange paired stations into groups.
  3. Resolve conflicts.

- Output: final groups of source stations that constitute IGRA stations

Step 1: Find Pairs

- Compare data, station identifiers, station names, and station locations

Example 1: **Match**

Source: NCDC6301 WBAN= 24233 WMO= 72793 NAME= Seattle Tacoma AP	100% data match, IDs match 0.0 km apart	Source: NCAR-MIT WBAN= WMO= 72793 NAME= Seattle-Tacoma Intnl

Example 2: **Match**

Source: NCDC6301 WBAN= 24233 WMO= 72793 NAME= Seattle Tacoma AP	No overlap, IDs and names match 3.1 km apart	Source: CDMP-USM WBAN= 24233 WMO= NAME= Seattle-Tacoma Airport

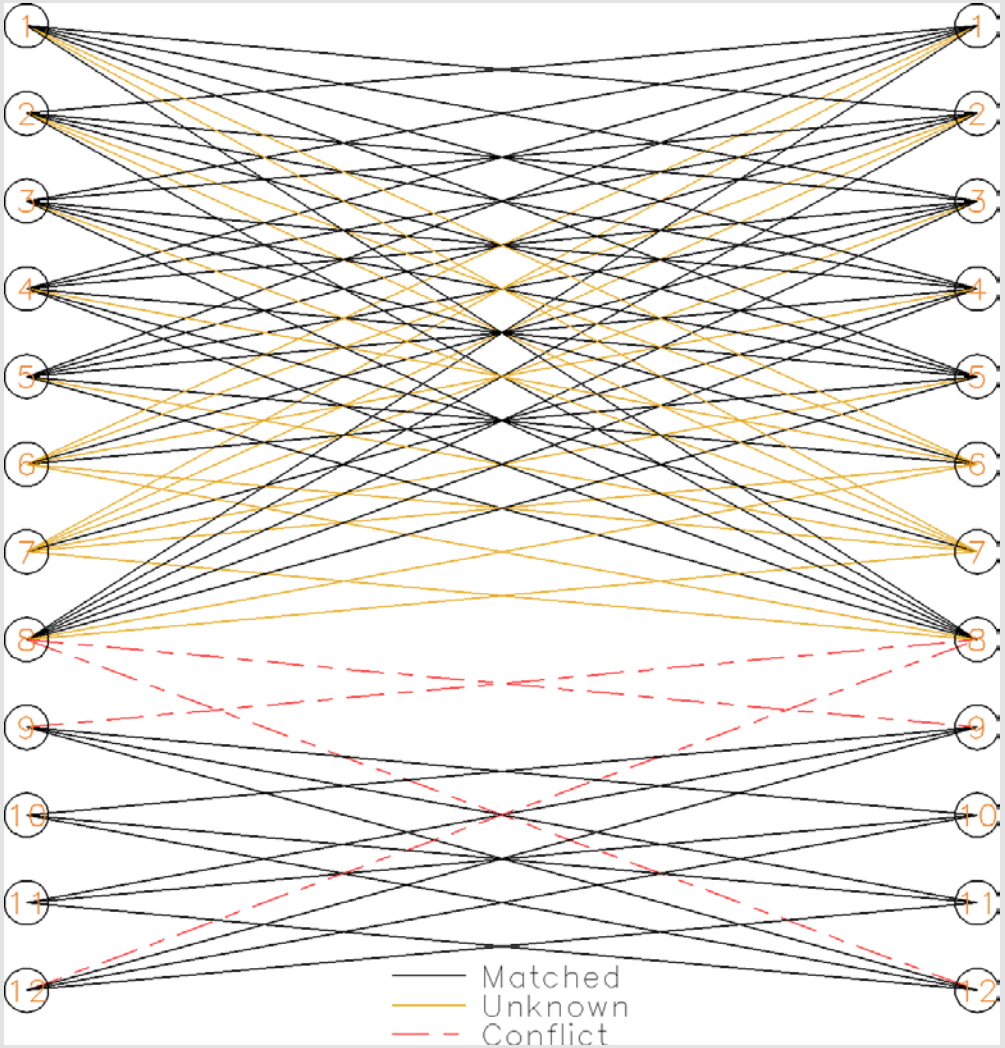
Example 3: **Conflict**

Source: NCAR-MIT WBAN= WMO= 72793 NAME= Seattle-Tacoma Intnl	100% data match 70.1 km apart	Source: NCDC6310 WBAN= WMO= 72792 NAME= OLYMPIA/MUNI (WASH)

Step 2: Identify Groups

- Classify connections between stations as:
  - ✓MATCH,
  - ✓SO-SO MATCH,
  - ✓CONFLICT,
  - ✓UNKNOWN, or
  - ✓SEPARATE.
- Form a group from each set of stations that are connected with matches or conflicts.

	Source	Station name
1.	usaf-ds3	SEATTLE-TACOMA INTL
2.	ncdc6309	SEATTLE/TACOMA INTL
3.	ncdc6310	SEATTLE-TACOMA INTNL
4.	ncdc6301	SEATTLE TACOMA AP
5.	chuan101	SEATTLE 3556
6.	chuan101	SEATTLE 3557
7.	cdmp-usm	SEATTLE-TACOMA AIRPORT
8.	ncar-mit	SEATTLE-TACOMA INTNL
9.	ncdc6301	OLYMPIA MUN I AP
10.	ncdc6326	OLYMPIA/MUNI (WASH)
11.	ncar-mit	OLYMPIA/MUNI (WASH)
12.	ncdc6310	OLYMPIA/MUNI (WASH)



Step 3: Resolve Conflicts

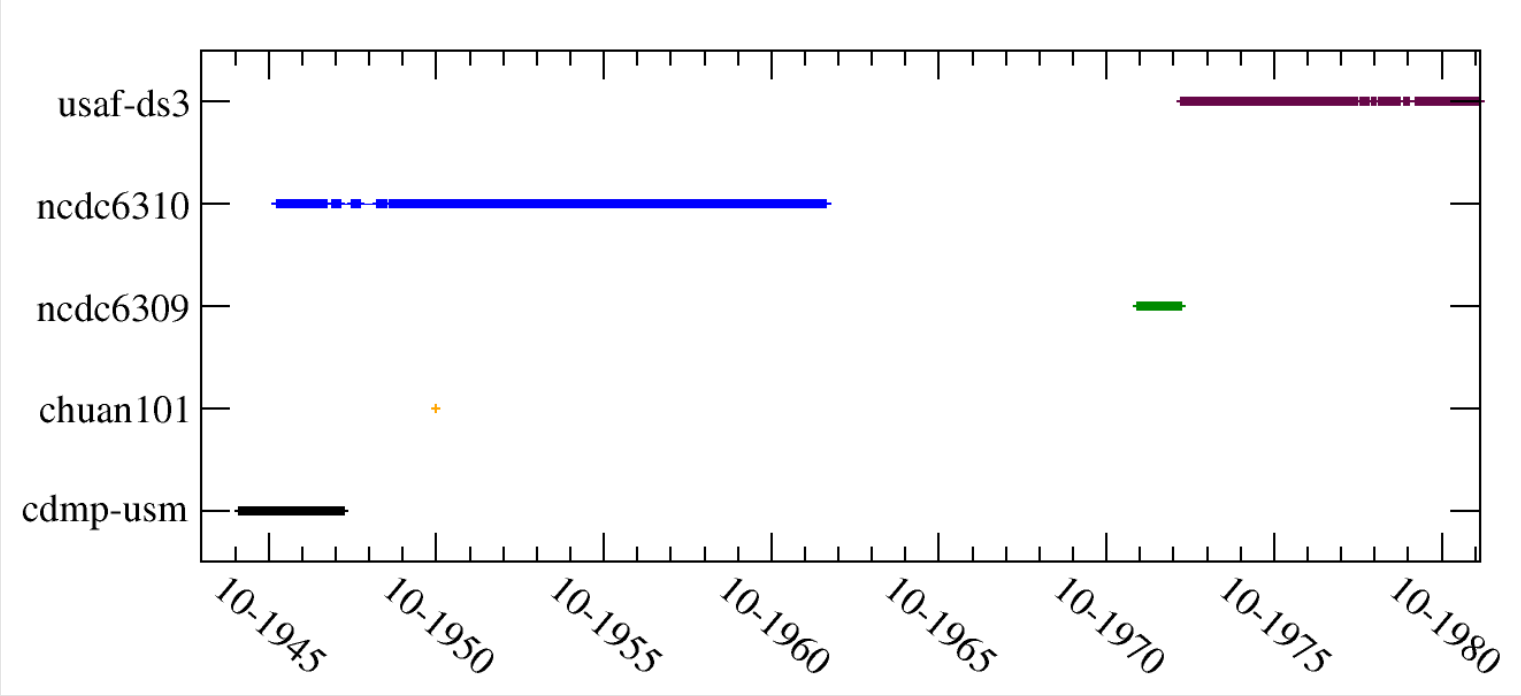
	1	2	3	4	5	6	7	8	9	10	11	12
1		2	2	2	2	0	0	2	-1	-1	-1	-1
2	2		2	2	2	0	0	2	-1	-1	-1	-1
3	2	2		2	2	1	0	2	-1	-1	-1	-1
4	2	2	2		2	0	2	2	-1	-1	-1	-1
5	2	2	2	2		1	0	2	-1	-1	-1	-1
6	0	0	1	0	1		0	0	-1	-1	-1	-1
7	0	0	0	2	0	0		0	-1	-1	-1	-1
8	2	2	2	2	2	0	0		-2	-1	-1	-2
9	-1	-1	-1	-1	-1	-1	-1	-2		2	2	2
10	-1	-1	-1	-1	-1	-1	-1	-1	2		2	2
11	-1	-1	-1	-1	-1	-1	-1	-1	2	2		2
12	-1	-1	-1	-1	-1	-1	-1	-2	2	2	2	

- Organize pairwise comparison results into one matrix per group.
- Eliminate certain source stations from groups with conflicts.
- Split groups into subgroups if needed.

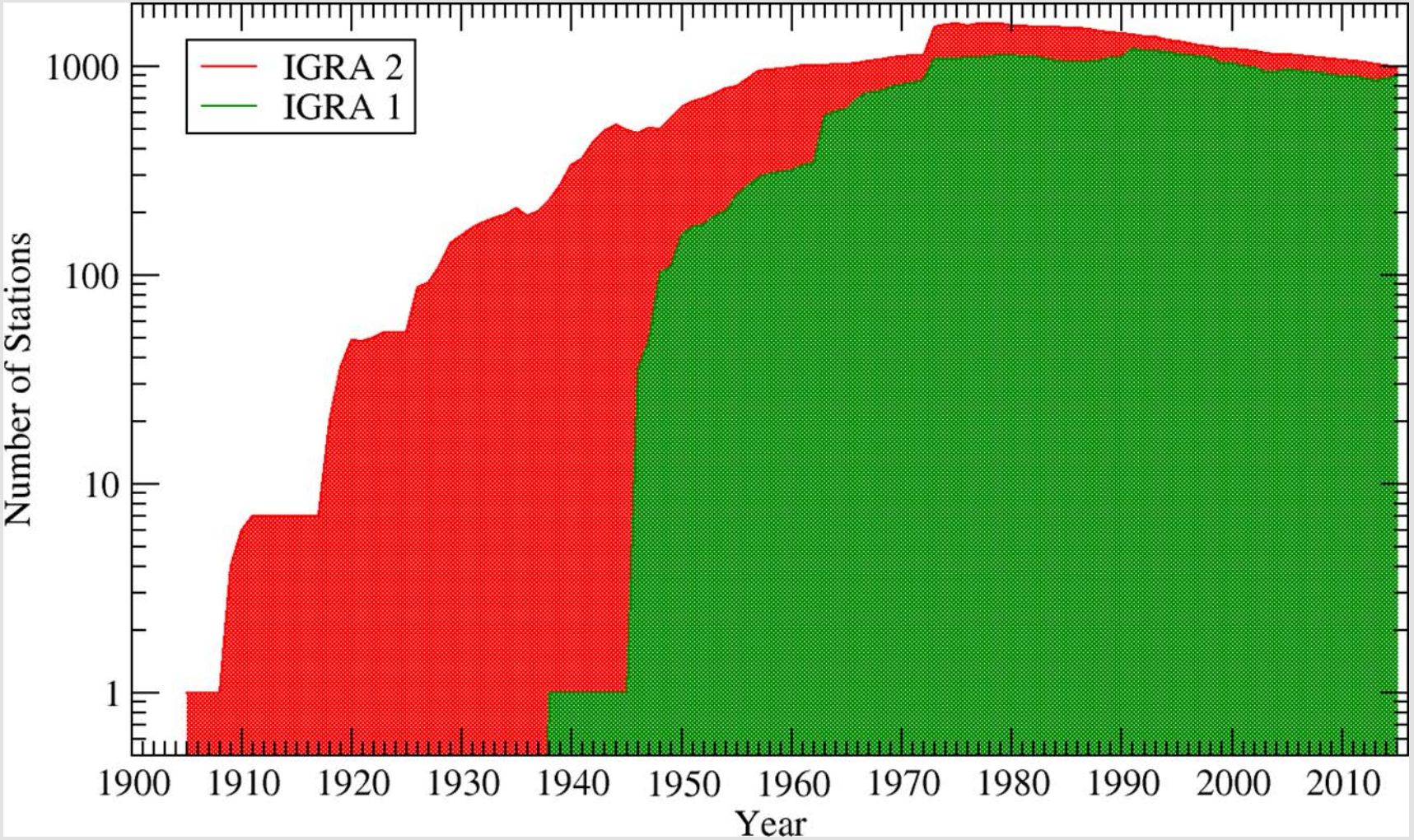
Outcome

- ✓ One multi-source time series per location
- ✓ More data per location than from a single source
- ✓ Clean separation of data for obviously distinct locations

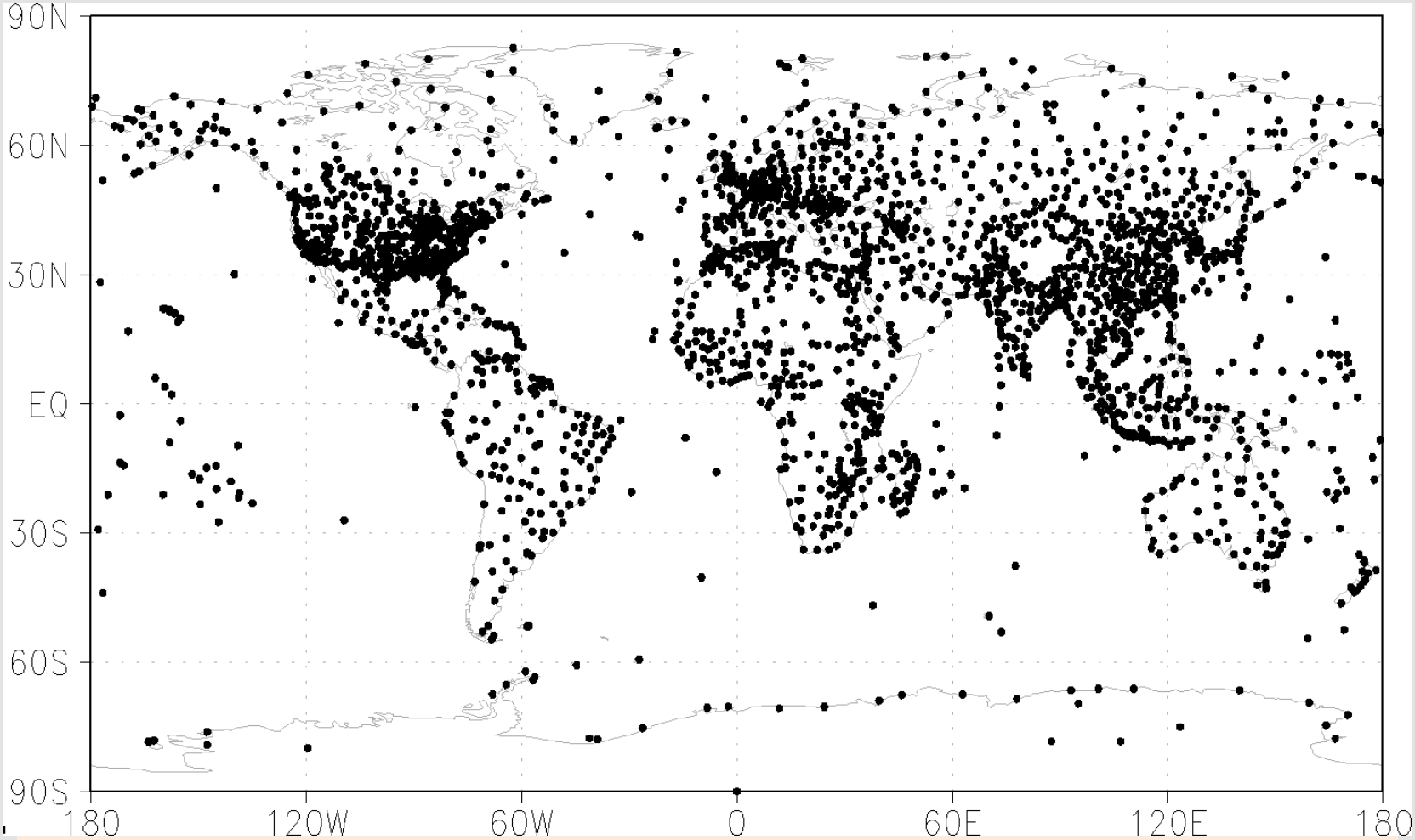
Time series indicating change data sources over time at Seattle Airport



Number of Stations by Year in IGRA 1 and IGRA 2



IGRA 2 Station Map



https://www.ncdc.noaa.gov/data-access/weather-balloon/integrated-global-radiosonde-archive/