Practice in International Data Collection and Harmonization

Matt Menne
NOAA-National Centers for Environmental Information (NCEI)
Asheville, NC USA

June 30, 2015

Global In Situ Datasets from NCEI

- International Global Radiosonde Archive (IGRA)
- International Comprehensive Ocean-Atmosphere Dataset (ICOADS)
- Land Surface Station Data (hourly/daily/monthly)
  - –Global Historical Climatology Network (GHCN) -Monthly
  - –Global Historical Climatology Network (GHCN) - Daily
  - –Integrated Surface Dataset (Hourly)

Global In Situ Datasets from NCEI

- International Global Radiosonde Archive (IGRA)
- International Comprehensive Ocean-Atmosphere Dataset (ICOADS)
- Land Surface Station Data (hourly/daily/monthly)
  - –Global Historical Climatology Network (GHCN) -Monthly
  - –Global Historical Climatology Network (GHCN) - Daily
  - –Integrated Surface Dataset (Hourly)

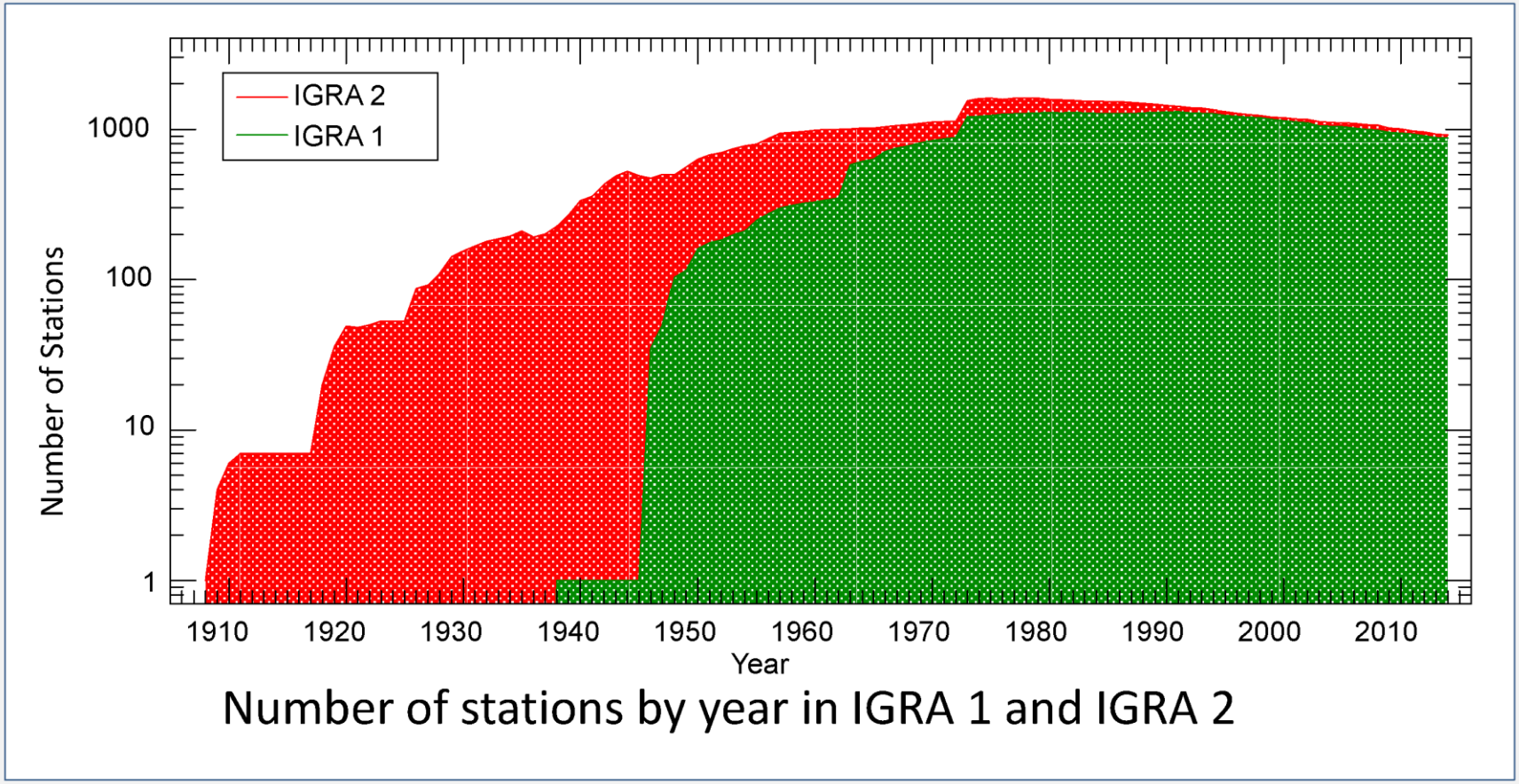
Some Common Themes in Global in situ Dataset Construction

- Data sets built from multiple source archives
  - –Requires reformatting native formats to a common format (not trivial!)
  - –Requires some mechanism for ongoing integration of newly available historical sources
- Near-real time updates
- Management of station histories & other metadata (e.g., aliases, location/instrument changes etc).
- A system for documenting, tracking and addressing errors

Sources used for IGRA 2

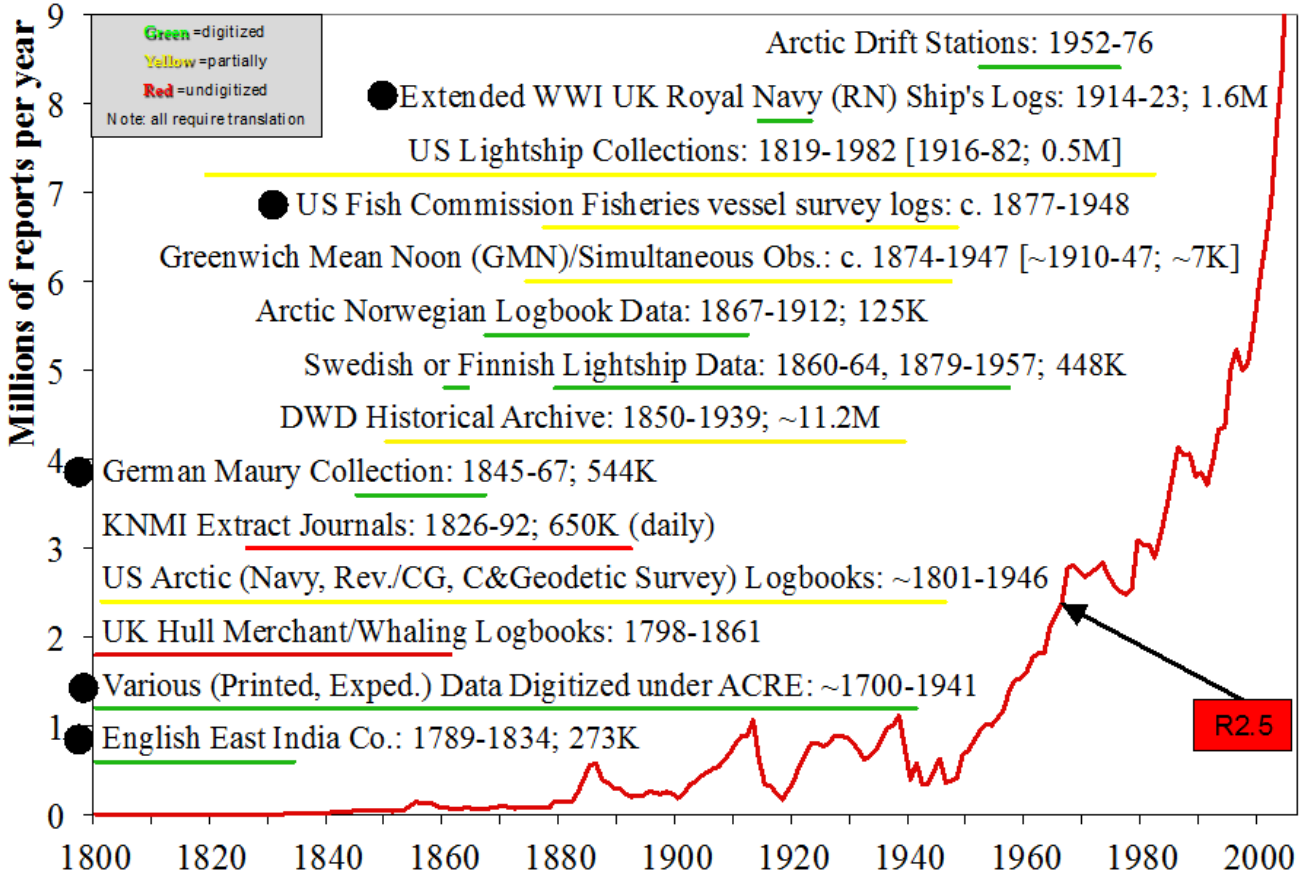
Origin	# Datasets	Spatial Coverage	# Stations	Period of Record	%Contribution
U.S. Air Force	1	Global land+ships	1963	1973–2008	46
NCDC archive, non-NWS	10	Global land+ships	1601	1938–2010	21
NCDC/NCEP GTS	1	Global land+ships	1203	2000–present	14
US NWS	2	US, US territories, US military sites	272	1946–present	9
Climate Database Modernization Program (CDMP)	14	U.S., Africa	441	1918–2002	6
NCAR	2	Global land	624	1949–1966	3
Data digitized for ERA-CLIM reanalysis	1	Global land+ships	252	1909–1972	1
Historical Arctic Radiosonde Archive	1	Arctic	66	1948–1996	<1
Meteo France	2	West Africa	50	1940–1965	<1
Other	3	Global land	241	1905–2010	<1

Upper Air Stations in IGRA



Number of stations by year in IGRA 1 and IGRA 2

Historical data awaiting blending into ICOADS



The time periods of selected candidate historical data sources to be blended into ICOADS are spanned by horizontal colored lines: green candidates are fully digitized but require format translation, yellow are partially digitized, and red are in the planning stages for digitization. Each dataset name is appended with the date range and approximate number of reports if known. The solid red curve is the number of reports (millions per year) in the current version of ICOADS (R2.5). Black dots mark sources definitely planned for inclusion (fully or partially) in the next release (version 3.0 with a target date of late 2015).

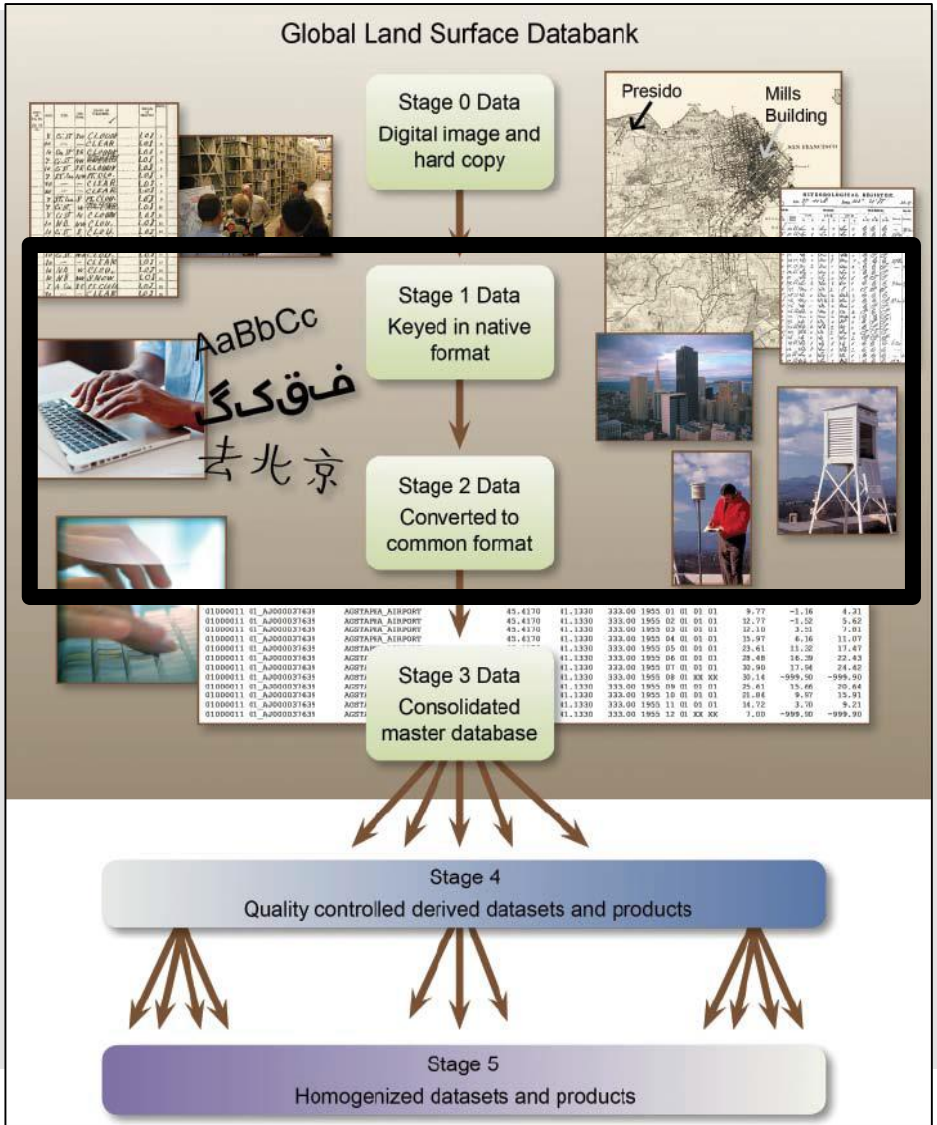
Land Surface Station Data

Datasets covering the three time resolutions were developed and have evolved independently

- Monthly Data
  - –Informal exchanges between colleagues
  - –National Archives
  - –CLIMAT message exchange over the GTS
- Hourly Data
  - –GTS data exchange
  - –Data rescue
  - –Mesonet data hubs
- Daily Data
  - –Informal exchanges between colleagues
  - –National Archives
  - –Web services
  - –No GTS data, but new daily CLIMAT message on the horizon

International Surface Temperature Initiative Databank Design

Conversion from Stage 1 → Stage 2 (***output data***)



ISTI Databank Sources

Name	Source	# Stns
Antarctica	SCAR Reader Project	44
Antarctica (AWS)	Antarctic Meteorological Research Center	136
Antarctica (Palmer Station)	Antarctic Meteorological Research Center	1
Antarctica (South Pole Station)	Antarctic Meteorological Research Center	1
Arctic	IARC/Univ of Alaska Fairbanks	133
Argentina	National Institute of Agricultural Technology (INTA)	35
Australia	Australia Bureau of Meteorology	103
Brazil	INPE, Nat. Institute for Space Research	495
Brazil-Inmet	INMET	289
Canada	Environment Canada	338
Canada	Environment Canada	6045
Central Asia	NSIDC	234
Channel Islands	States of Jersey Met	2
Colonial Era Archives	*Griffith*	*1021*
CRUTEM3	UKMO	5113
CRUTEM4	UKMO	6190
East Africa	Univ. of Alabama Huntsville	263
Ecuador	Inst. Nacional De Met E Hidrologia	1
Europe / N. Africa	European Climate Assessment (Daily, Non-Blended)	10269
Europe / N. Africa	European Climate Assessment (Daily, Blended)	2905
Europe / N. Africa	European Climate Assessment (Monthly)	4278
Germany	DWD- Germany	106
GHCN-Daily	*NCDC*	*30633*
GHCN-M v2	*NCDC*	*13500*
GHCN-M v2 Source	*NCDC*	*26241*
Giessen	University of Giessen	44
Global Summary of the Day	*NCDC*	23863
Greater Alpine Region	Histalp / ZAMG	138
Greenland	*NCAR*	*8*

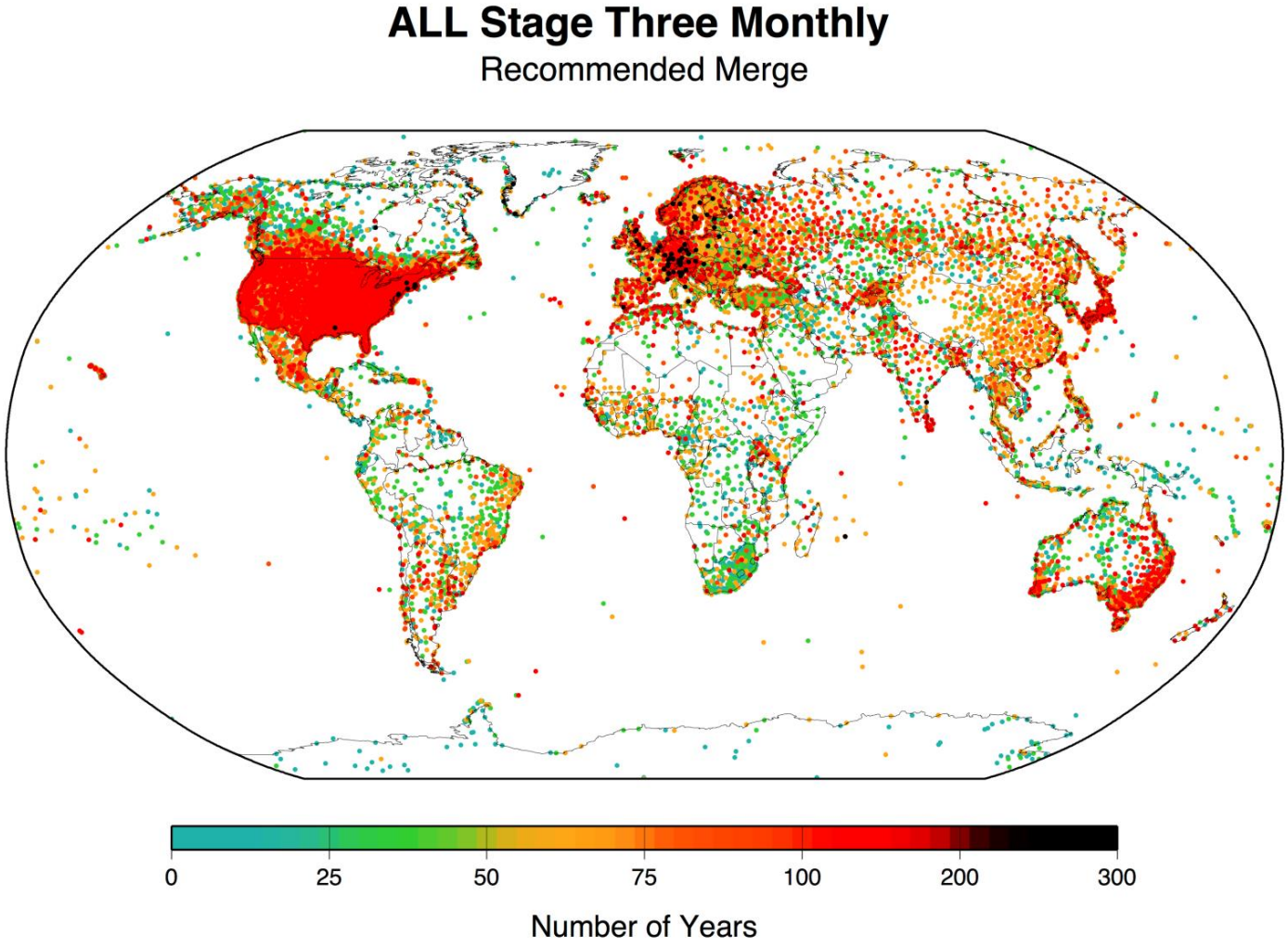
Name	Source	# Stns
HadISD	UKMO	5865
India	India Meteorological Department	53
Japan	JMA	157
Max/Min Stations from R. Vose	*NCDC*	*36158*
Mexico	*CDMP*	*95*
Mon. Clim Data of World (MCDW)	*NCDC*	*2876*
MCDW (Completed, unpublished)	*NCDC*	*2392*
Mon. Surf. Station Clim. (WMSSC)	*NCAR*	*4752*
Norway	Norwegian Meteorological Institute	906
Pitcairn Island	Met Service of New Zealand	3
Polar	ISPD	2
Preliminary CLIMAT	*NCDC*	*2883*
Russia	Roshydromet	517
Southeast Asia	Southeast Asia Climate Assessment (Non-Blended)	577
Southeast Asia	Southeast Asia Climate Assessment (Blended)	206
Spain	Univ. Rovira I Virgili	22
Sweden	GCOS Surface Network	13
Switzerland	ISPD	3
Switzerland	Digihom/MetoSwiss/IAC-ETH	3
Sydney	ISPD	1
Tunisia/Morocco	ISPD	13
Uganda	Univ. of Alabama Huntsville	32
UK CLIMAT	UKMO	240
Uk Met Office Historical	UKMO	37
Uruguay	Universidad de la Republica, Montevideo, Uruguay	11
Uruguay	Inst. Nacional de Invest Agropecuaria	5
US Forts	*CDMP*	*217*
Vietnam	*CDMP*	*32*
World Weather Records	*WMO*	*3036*

58 sources, ~180,000 total stations. 15 sources from NCDC (~125,000 stations)

ISTI Databank

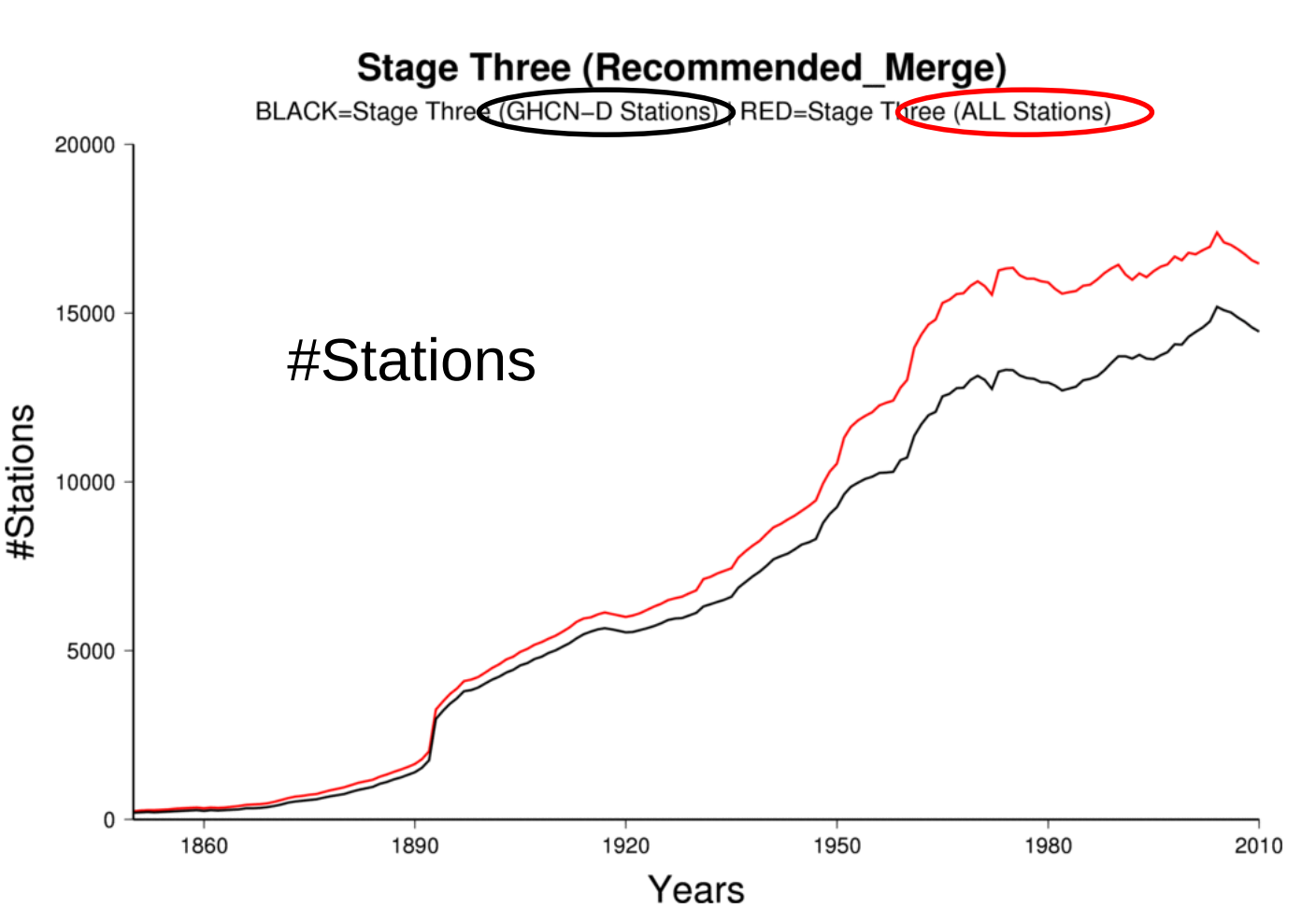
- Stage 3 Monthly Databank released on June 30, 2014 with ~32,000 stations
- A new build of the monthly databank (v. 1.1.0) planned for release in the coming weeks (approximately 35,000 stations)

ISTI Monthly Data Set (Version 1.0.1)



ISTI Global Databank
GHCN-Daily vs all other Databank sources

- First step in reconciling monthly and daily datasets
- Similar effort occurring for monthly precipitation



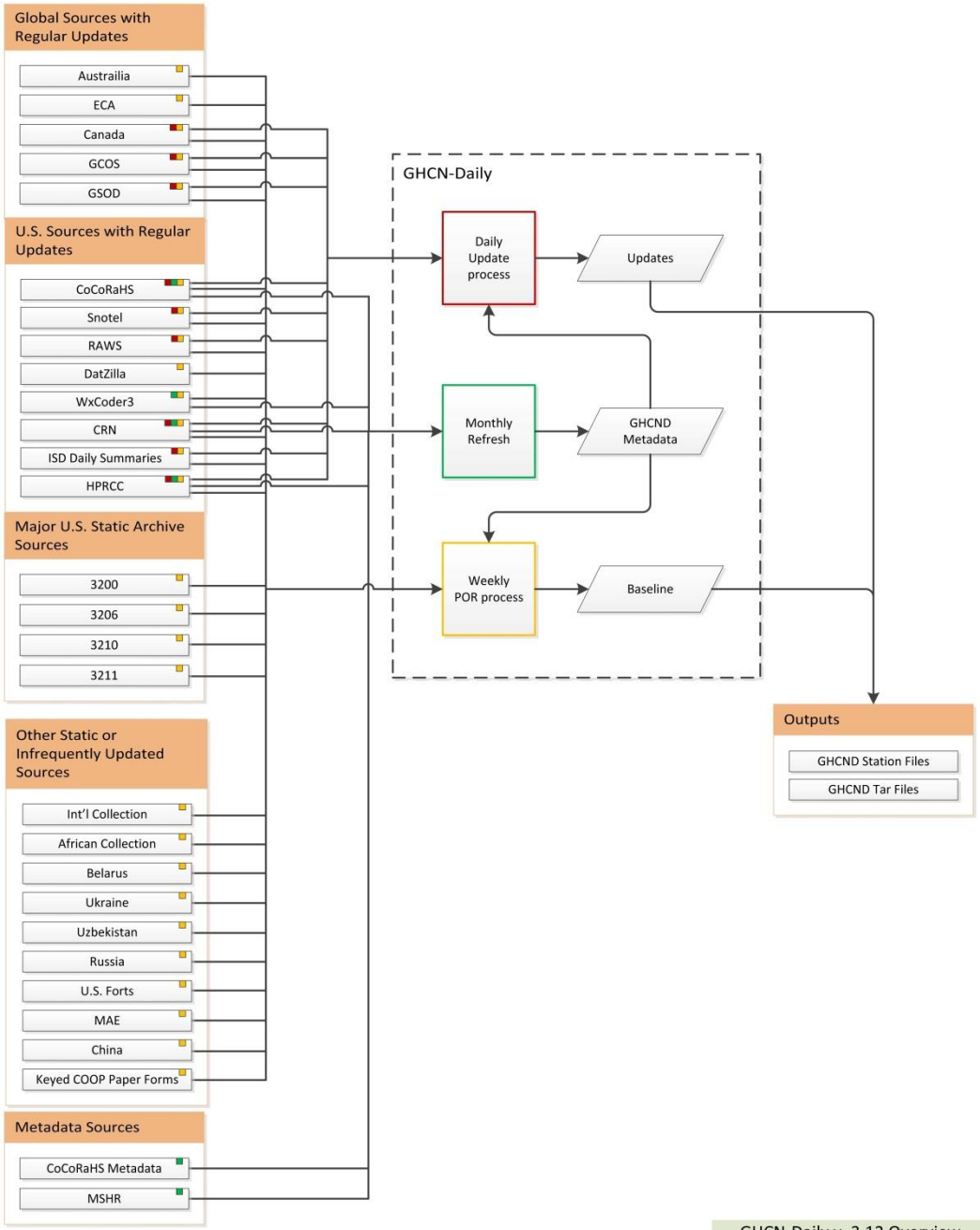
GHCN-Daily

- Daily in situ dataset derived from multiple about 30 sources
- Comprehensive daily dataset for the USA (multi-variable) with good coverage over many other parts of the world (precipitation, temperature, snow depth)
- Integrates latest U.S. daily source archives and real-time updates for many U.S. Networks as well as Canada. GTS updates for non-U.S. sites in some cases. Monthly updates for Australia, ECA&D sites.
- >30,000 temperature stations
- >90,000 precipitation stations
- >40,000 snowfall or snow depth stations
- Serves as foundation for new monthly temperature and precipitation datasets + other monthly summaries
- Includes an error tracking system known as “Datzilla”
- Candidate datasets waiting on deck for integration as time permits

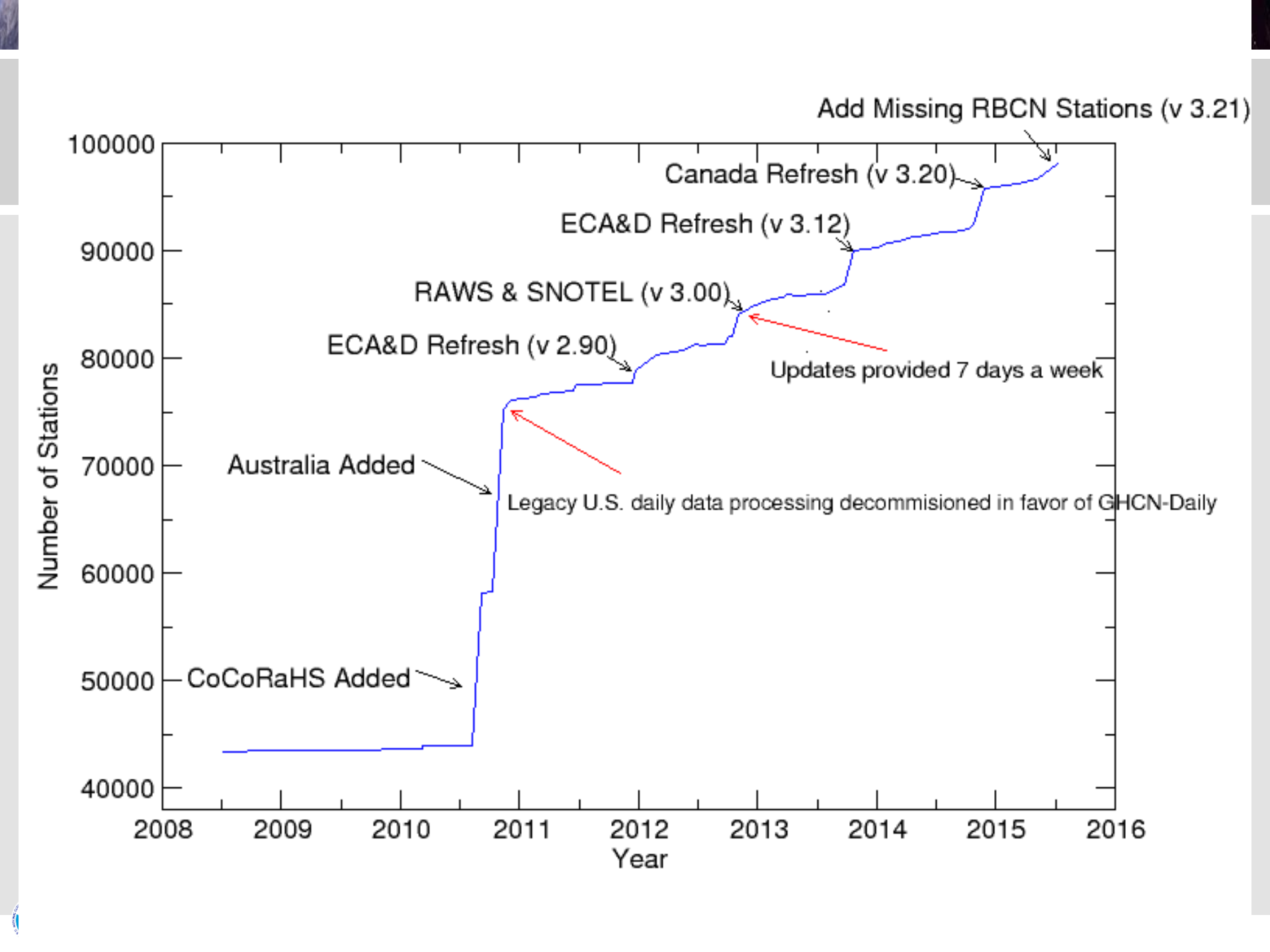
Four facets to GHCN-Daily Processing

1. Daily updates (automated) – Updates Data values
   - Updates values that are new since yesterday's update
2. Weekly reprocessing (automated) – Updates Data Values
   - Reintegrates source databases and reruns quality checks on all values. Helps to ensure that GHCN-Daily is synchronized with its external constituent sources, but does not add new stations
3. Monthly refreshes for select U.S. networks (automated, but requires approval/manual intervention to deploy refreshed list) – Updates Membership in GHCN-Daily
   - Removes and reintegrates active data sources for Coop, CoCoRaHS and CRN. Adds stations that are new since last monthly refresh
4. Periodic adding of new sources or refreshing of large existing data sources (semi-automated) – Updates Membership in GHCN-Daily
   - Removes and reintegrates large data sources to incorporate station additions since last refresh

GHCN-Daily
Level 0 Diagram



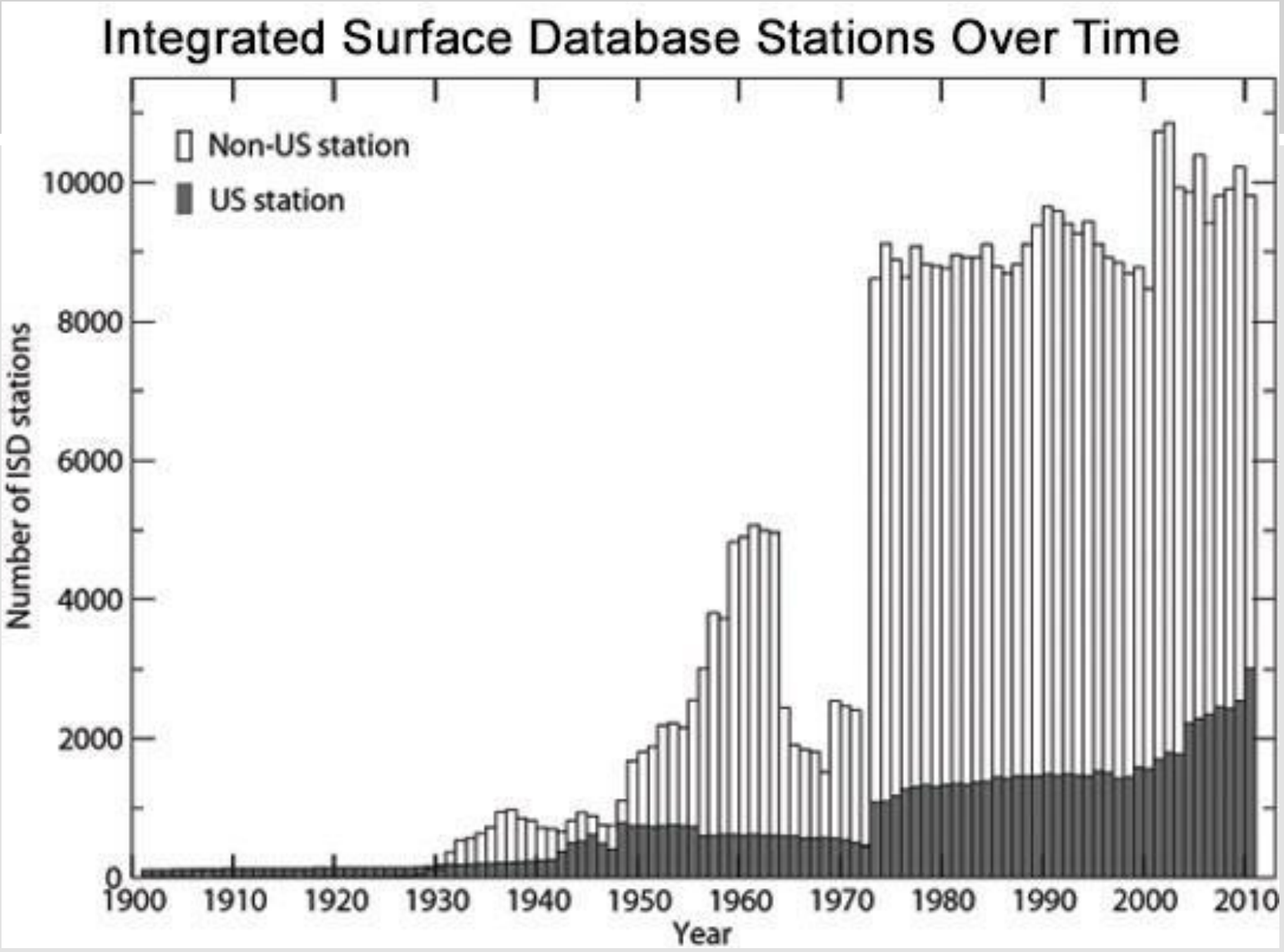
GHCN-Daily v. 3.12 Overview
Level 0 Flow Diagram
Version 0.3
Modified: May 28, 2014


Add Missing RBCN Stations (v 3.21)
Canada Refresh (v 3.20)
ECA&D Refresh (v 3.12)
RAWS & SNOTEL (v 3.00)
ECA&D Refresh (v 2.90)
Updates provided 7 days a week
Australia Added
Legacy U.S. daily data processing decommisioned in favor of GHCN-Daily
CoCoRaHS Added
Number of Stations
Year
100000
90000
80000
70000
60000
50000
40000
2008
2009
2010
2011
2012
2013
2014
2015
2016

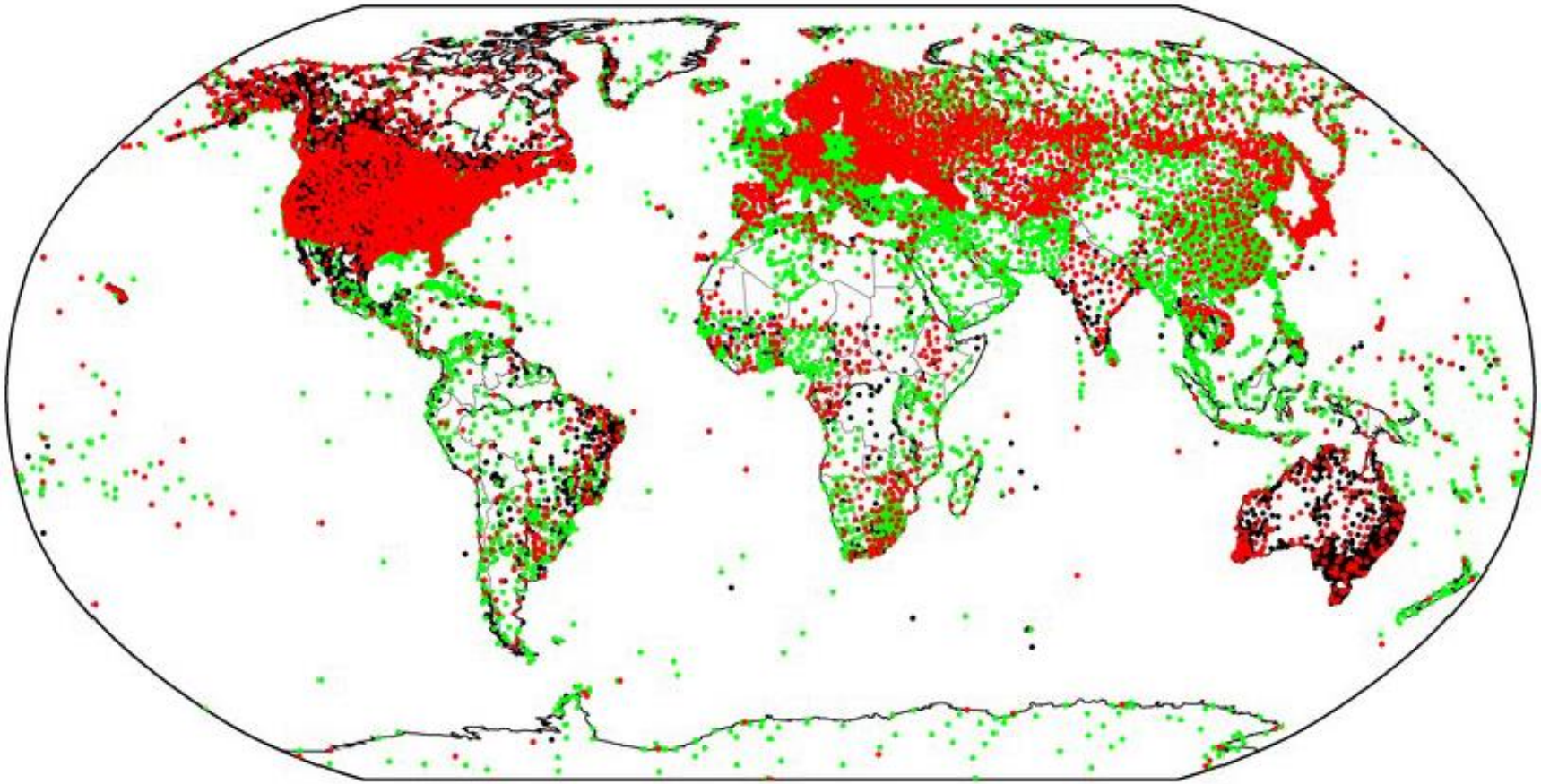
Integrated Surface Dataset
(Mostly Hourly: Metar/Synop)

- Contains global hourly and synoptic observations compiled from ~100 sources
- Developed as a joint activity within Asheville's Federal Climate Complex. Data feed and new additions largely managed by the 14th Weather Squadron of the U.S. Air Force. NOAA/NCDC's ISD is the public facing version of the Air Force database
- Comprises over 20,000 stations worldwide, with data as far back as 1901, though big increases in volume occur in the 1940's and again in the early 1970's
- Currently over 11,000 stations "active" and updated daily in the database.
- NCEI planning a major re-engineering effort of ISD processing to align it with daily processing principles, complete vertical integration across time intervals (hourly/daily/monthly) and improve metadata management

Integrated Surface Database Stations Over Time



GHCN–Daily Merged with ISD–Lite



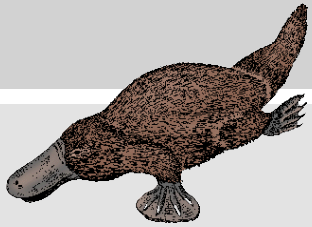
- Existing GHCN-Daily Temperature Station without an ISD-Lite data match
- Existing GHCN-Daily Temperature Station with an ISD-Lite data match
- Potential new GHCN-Station from ISD-Lite

HOMR's *In Situ* Station History

Identifiers	Consolidation of IDs over time (ICAO, WBAN, FAA, WMO, COOP, GHCN-Daily…)
Names	Stations can have many aliases
Locations	Latitude/longitude, elevations, topography, obstructions, relocations
Elements	Observation times, reporting methods
Equipment	Types, modifications and siting

Station history management is similar to building a dataset or product – acquire, QA, integrate, manage, provide access.

Our Metadata Sources

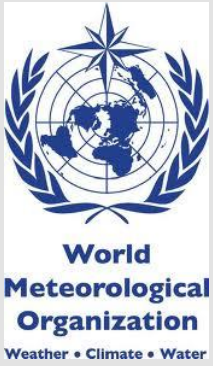


DATZILLA

US Climate Reference Network

USRCRN

GHCN-Daily



Series of automated and manual quality control checks before going to NCDC Archive

NOAA's National Climatic Data Center, Product Development Branch

Example of Pub. 9 Volume A History (since 1990s)

SOURCE	SOURCE_ID	WMO_ID	NAME_PRINCIPAL	WMO_COUNTRY_CODE	WMO_COUNTRY_NAME	FIPS_COUNTRY_CODE	FIPS_COUNTRY_NAME	LAT_DEC	LON_DEC	ELEVATION_GROUND_M	BEGIN_DATE	END_DATE
WMOPUB9VOLA	14862	71956	GORE BAY CLIMATE, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	188	11/22/2010	2/16/2015
WMOPUB9VOLA	14862	71956	GORE BAY CLIMATE, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	189	2/16/2015	12/31/9999
WMOPUB9VOLA	4666	71733	GORE BAY, ONT.	4020	CANADA	CA	CANADA	45.88	-82.57	193	2/3/1999	3/26/2001
WMOPUB9VOLA	4666	71733	GORE BAY, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	193	3/26/2001	9/16/2002
WMOPUB9VOLA	4666	71733	GORE BAY A, ON	4020	CANADA	CA	CANADA	45.88	-82.57	193	9/16/2002	9/29/2004
WMOPUB9VOLA	4666	71733	GORE BAY A, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	194	9/29/2004	12/18/2006
WMOPUB9VOLA	4666	71733	GORE BAY AWOS, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	194	12/18/2006	11/22/2010
WMOPUB9VOLA	4666	71733	GORE BAY AWOS, ONT	4020	CANADA	CA	CANADA	45.88	-82.57	193	11/22/2010	5/7/2012

Summary

- Vertical integration of monthly and daily datasets essentially completed
- Better integration also planned for subdaily dataset
- Ultimate goal is a land surface station dataset sorted by the time resolution (hourly/synoptic; daily; monthly) with common identifiers, source and QC flags etc. perhaps under a broader international umbrella (ICOADS for Land?)

Some Potential Collaboration with Copernicus

- Provide input into construction of a new data model (e.g., formats, methods for traceability) for land surface data
- Identify potential source datasets that can be integrated into global datasets and help reformatting efforts
  - –with special attention to the potential for resynchronization with national source archives and regular updates
  - –(regional efforts like ECA&D are really helpful!)
- Promote the concept of a multi-element land surface station database similar to ICOADS so that a recognized international database for land data can be established

Thank You!

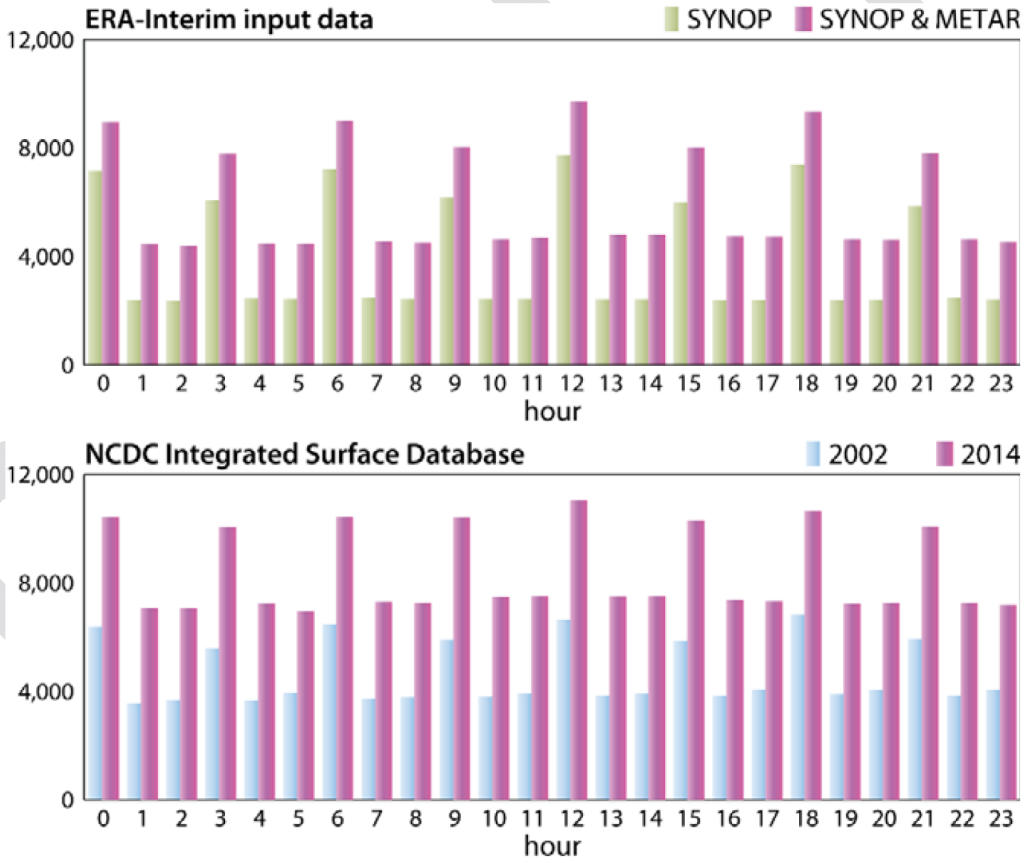


Figure 6: Average counts of surface air temperature observations over land for each hour of the day for October 2014 from ECMWF's operational receipt of data, as processed in ERA-Interim following basic quality-control checks (upper), and for October 2002 and 2014 from the NOAA NCDC Integrated Surface Database after duplicate removal and elimination of sub-hourly data (lower). ERA-Interim counts are shown for SYNOP reports alone, and as supplemented by METAR reports. NCDC data were downloaded from the ISD-Lite data stream on 22 January 2015.