Building a Comprehensive Database of Radiosonde Measurements for Climate Change Studies

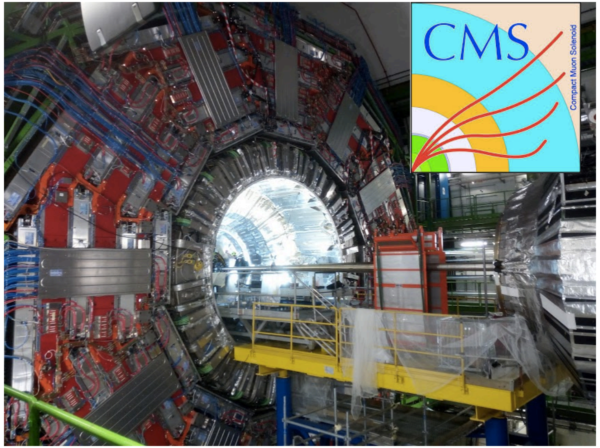
Federico Ambrogi

Leopold Haimberger, Michael Blaschek,
Ulrich Voggenberger

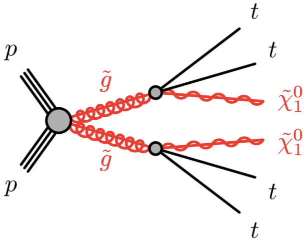
Master Seminar - IMG (Univie)
Vienna, 18/11/2020

Hello! Where I started...

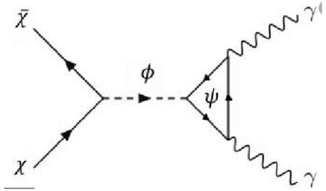
Tomorrow @ 14:00 , Astro Particle Physics WS2020:
https://cern.zoom.us/j/4940503352?pwd=TzNHcmUzM0Vvb2FsVFMyRG93ZDY2dz09



SModelS



MadDM v3



Interpretation of the Searches of Supersymmetry at the LHC

Automatic use of the Fermi-LAT satellite limits on Dark Matter annihilation

About Copernicus

"Copernicus is the European Union's Earth Observation Programme, looking at our planet and its environment for the ultimate benefit of all European citizens. It offers information services based on satellite Earth Observation and ***in situ (non-space) data***

https://www.copernicus.eu/

…

The information services provided are freely and openly accessible to its users"

Our Service: DC3S311c_Lot2

Copernicus Europe's eyes on Earth — Climate Change Service

Home Search Datasets Applications Toolbox FAQ Live

https://cds.climate.copernicus.eu (*not yet available, still dev phase)

Comprehensive upper-air observation network from 1905 to present

- The Copernicus Early Upper Air Subdaily Dataset is designed to contain all available upper air data from radiosondes and PILOT balloons prior to 1979
- ***Cleanest and most complete upper air data set for assimilation***

- Organized in station time series of records on significant levels / significant pressure levels
- Subject to strict quality control
- Data set homogenized in order to reduce systematic errors
- Observation errors estimated using Desroziers' method (based on reanalyses information).
- Monthly data set also available, containing gridded fields
- Additional historical radiosonde intercomparison data set to diagnose systematic errors between radiosonde types.

Everything we do —> https://github.com/MBlaschek/CEUAS

Overview

1. Build the most comprehensive database for upper air data (from historical rescued data to almost-most-recent)
2. Perform quality checks on the data available from different sources/databases
3. Estimate observation uncertainties (using reanalyses information with Desrozier's method)
4. Correct observation values for biases (using reanalyses information)
5. Produce gridded data for climatological studies
6. Provide historical intercomparison data (useful for error estimations)
7. Implement and maintain the CDS interface to allow users to download and use the data

1. Building the Database

The Common Data Model (CDM) https://github.com/glamod/common_data_model/

- It is a pre-defined, structured framework that allows to store *any* information in the database we build
- Both for data (e.g. measured quantities) and metadata (e.g. instrument type, station location)
- Common to other upper-air services -> easy to compare data from different lots, easy to build common software infrastructure
- How to store these tables is arbitrary (i.e. the specific file type —> we chose **netCDF**)

	element_name	kind	external_table	description
	observation_id	varchar (pk)		unique ID for observation
	report_id	varchar	header_table:report_id	Link to header information
	data_policy_licence	int	data_policy_licence:policy	WMOessential, WMOadditional, WMOother
	date_time	timestamp with timezone		timestamp for observation
	longitude	numeric		Longitude of the observed value, -180 to 180 (or other as defined by CRS). This may or may
	latitude	numeric		Latitude of the observed value, -90 to 90 (or other as defined by CRS)
	crs	int	crs:crs	Coordinate reference scheme use to encode location
e.g. pressure	z_coordinate	numeric		z coordinate of observation
	reference_z_coordinate	numeric		z coordinate of reference observation
	z_coordinate_type	int	z_coordinate_type:type	Type of z coordinate
	observation_height_above_station_surface	numeric		Height of sensor above local ground or sea surface. Positi
e.g. temp.	observed_variable	int	observed_variable:variable	The variable being observed / measured
	secondary_variable	int	secondary_variable:variable	Secondary variable required to understand observation, e.g
e.g. 290 K	observation_value	numeric		The observed value

1. Building the Database

Several existing sources of radiosonde data (ECMWF, USA, WMO …)

ERA5_1: The analysis input for ERA5 [1979-01-01,2019-12-31]

ERA5_2: The analysis input for ERA5 [1950-01-01,1978-12-31]

ERA5_1759 / ERA5_1761: Analysis input for ERA5 received from NCAR (MIT tapes)

ERA5_3188: Analysis input for ERA5 received from NCAR (CHUAN database)

NCAR: The NCAR upper air data base (1920 to present), NCAR catalog ds370.1

IGRA2: The IGRA version 2 data base (1905 to present), NOAA, NCEI

Pangaea: Radiosonde intercomparison data collected by UBERN and RIHMI for this service

- *No data set is really comprehensive*
- *Not user friendly*
- *Intercomparisons do not have a real dataset*
- …

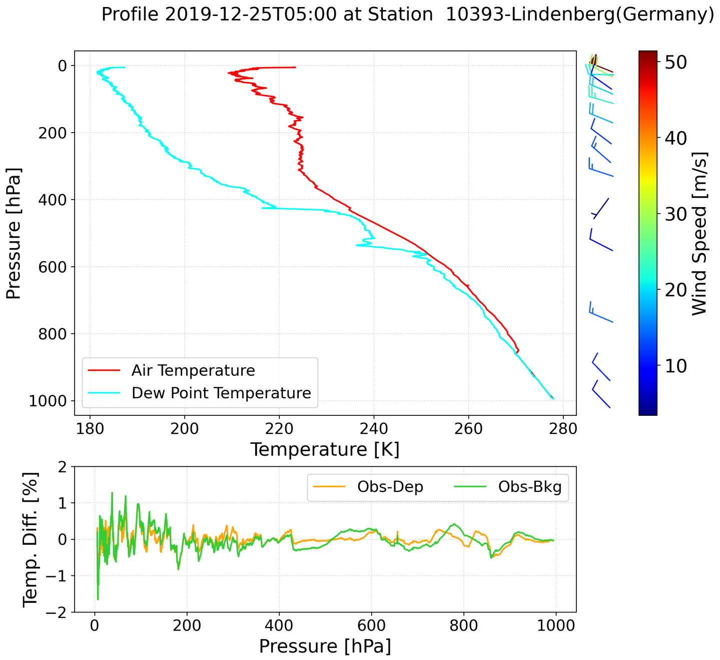
"Harvesting"

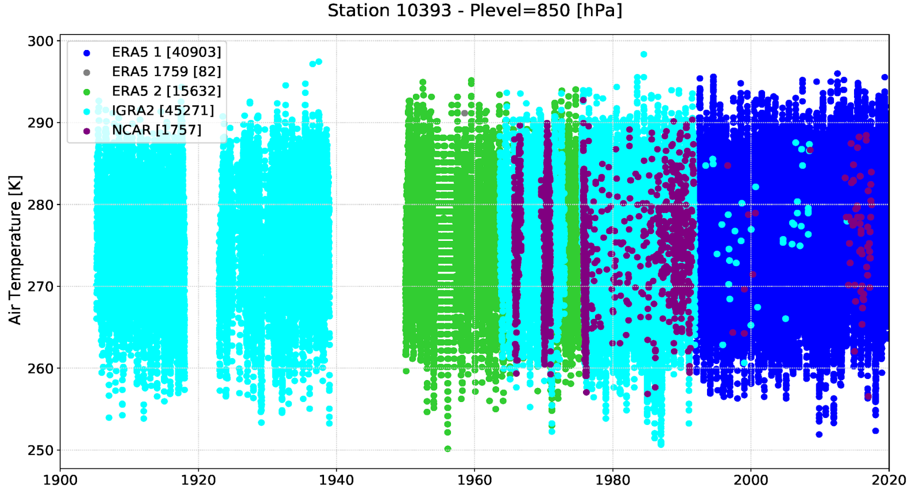
1. Collect all the available data/metadata
2. Take a common file format (**HDF netCDF4**) and implement the **CDM***
   One file per station

"Merging"

3. Compare data from different sources, remove duplicates, check quality, combine the data

1. Building the Database (from the Lindenberg station)



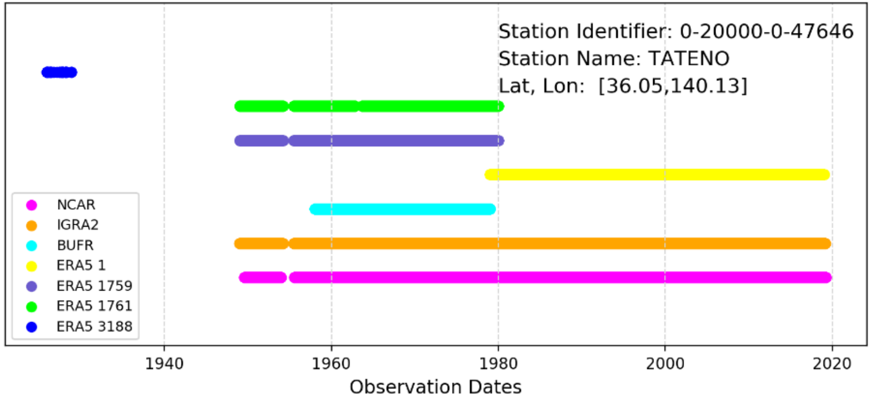


- Example of some variables (Temperature, dew point temperature, wind speed, wind direction) for many pressure levels i.e. full ascent data
- Additionally, ERA5 reanalyses data are available

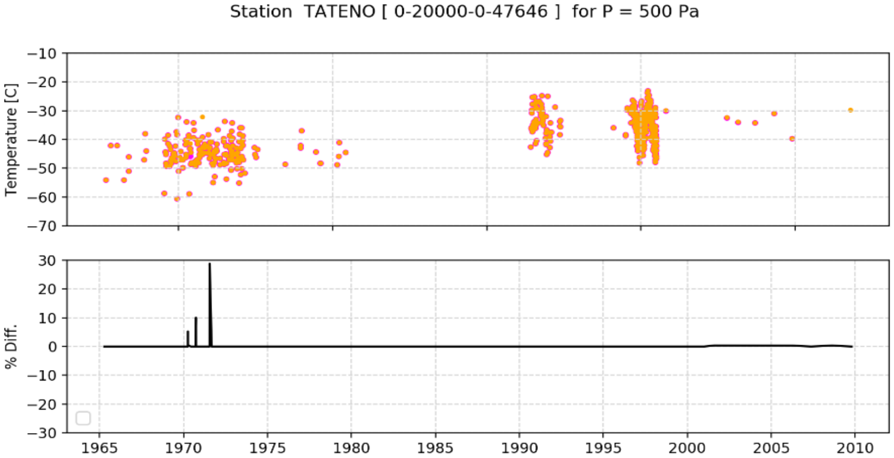
- Since~hundred thousands of records are available, we can build time series
- Each record comes from a different “best” source

1. Building the Database: Selection of the Best Source

- Data for the same station is available in multiple sources
- We define a specific set of rules to select the “best” records (i.e. data for a specific sonde ascent) from the available data sources and avoid duplicated data

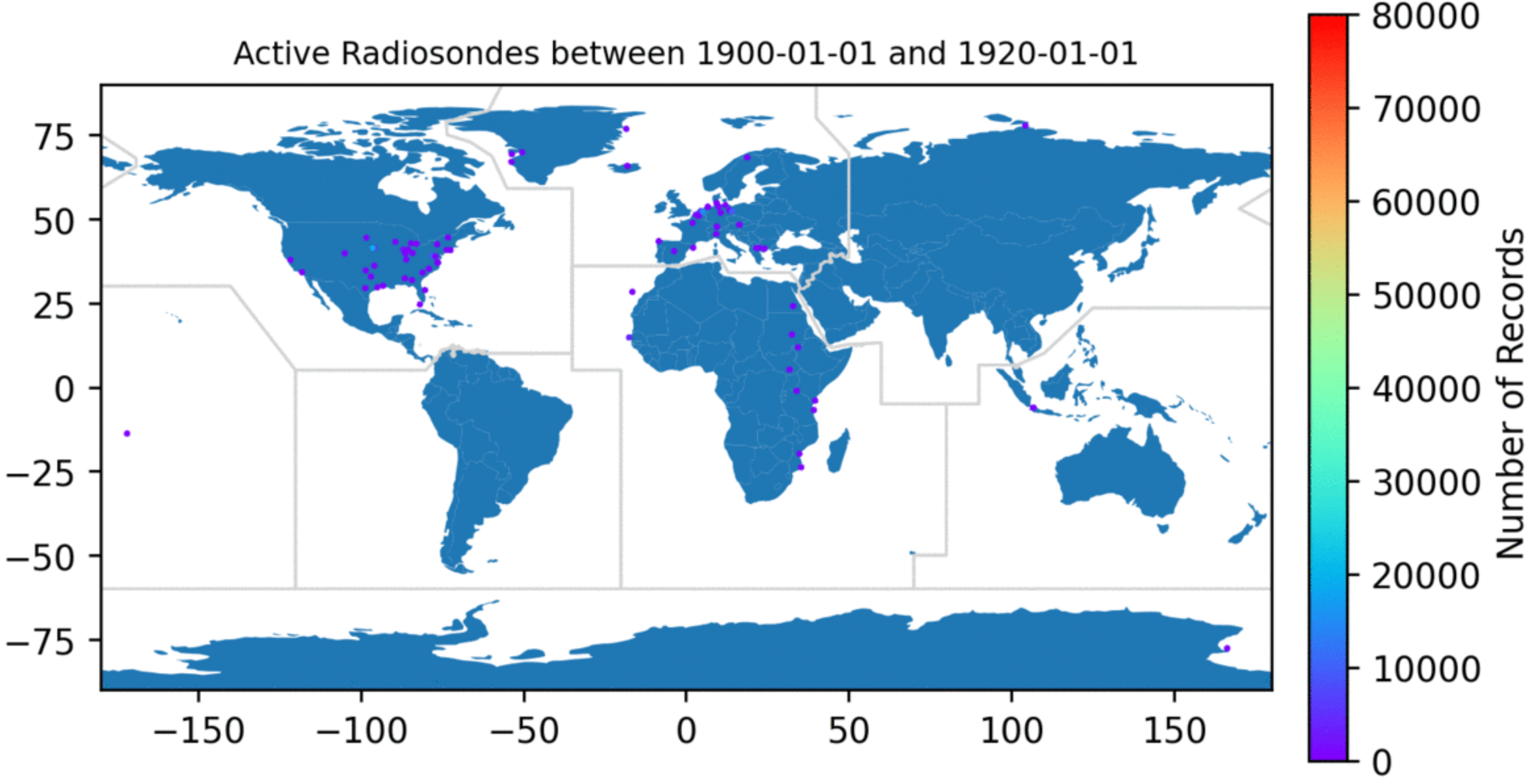


Availability of data from different datasets at station 47646 (Tateno, Japan)



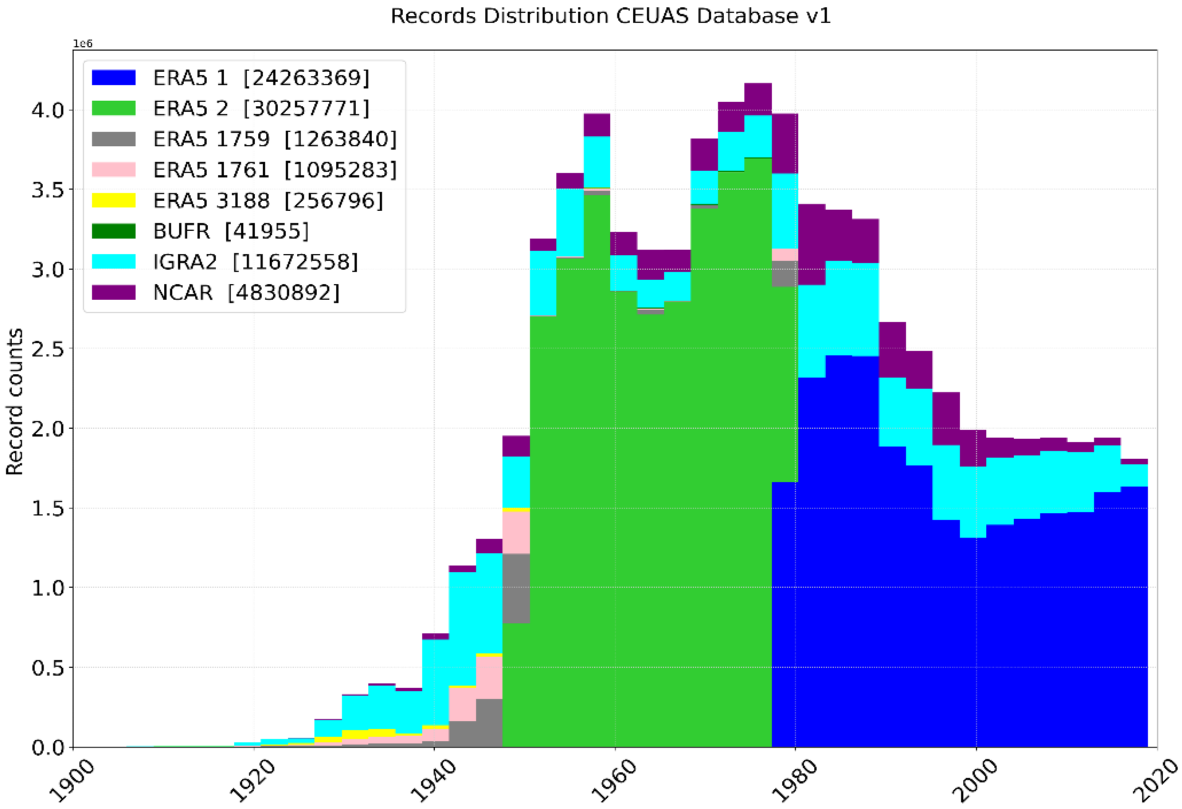
Comparing and checking for inconsistency in the temperature data, for 500 Pa, in the NCAR and IGRAv2 datasets

1. Building the Database



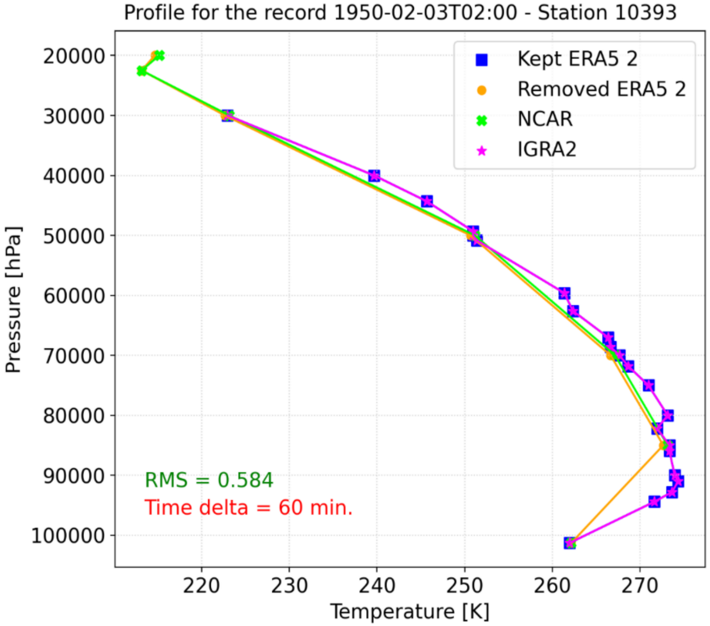
1. Building the Database

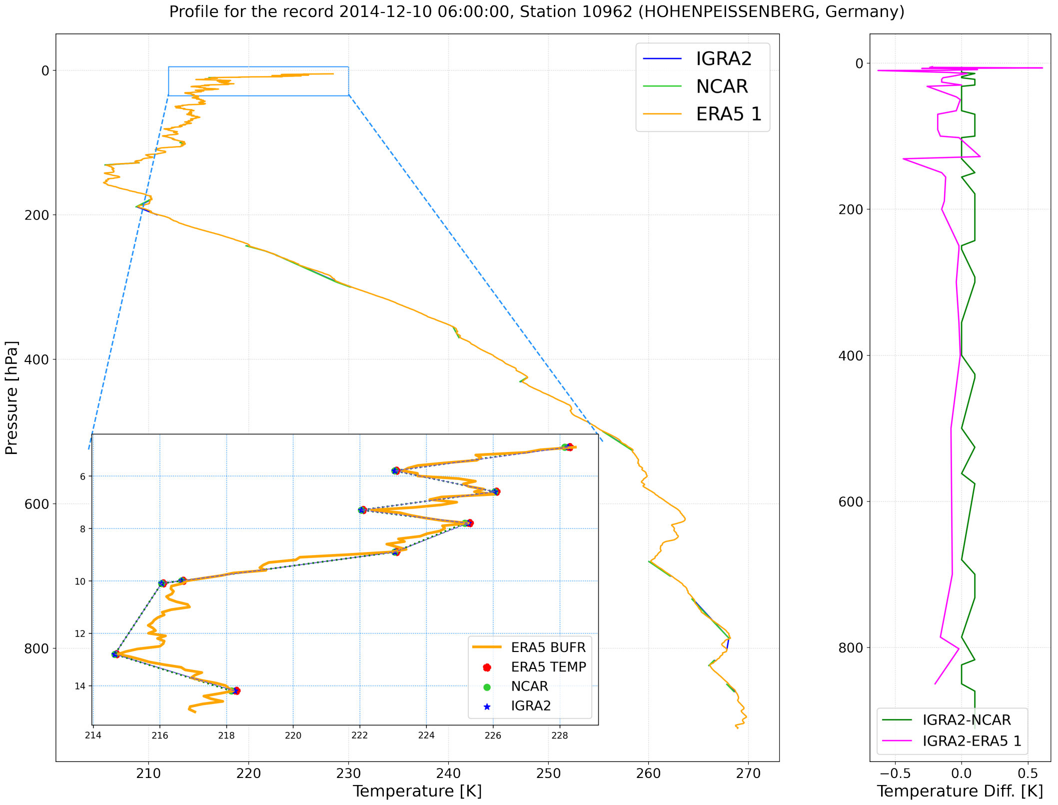
6850 stations
> 60 million records
period: [1905-2020]



2. Quality Checks

- Find inconsistencies between datasets
- Flag “weird” values
- Check for bugs
- …



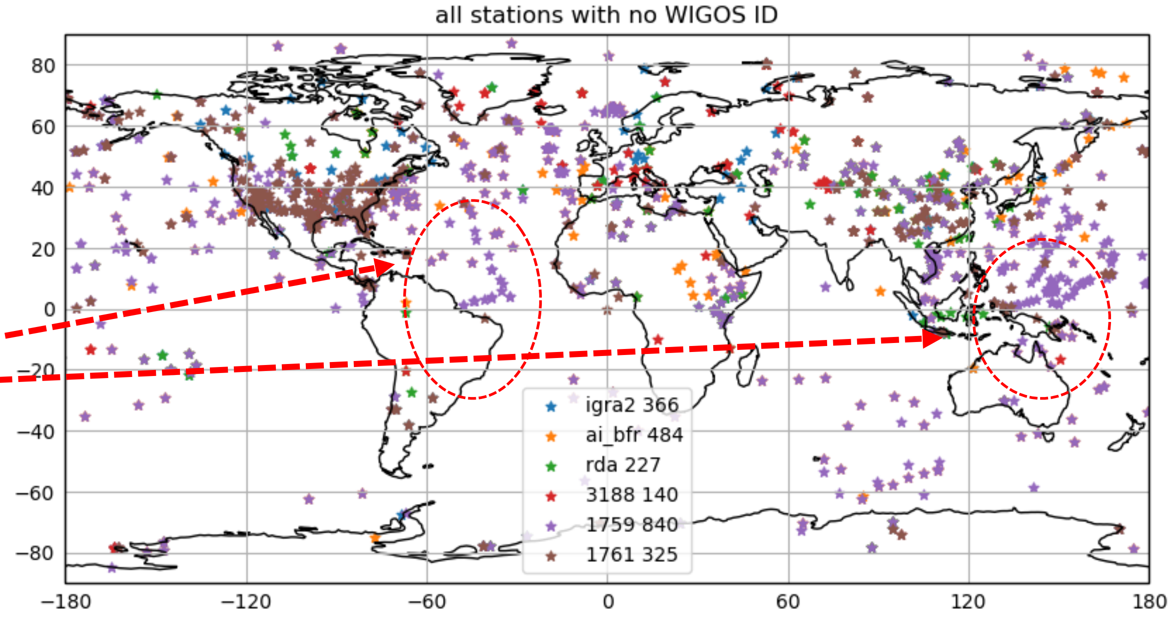


2. Quality Checks

Identification of stations for WIGOS standards

- For a number of stations, WMO identifiers could **NOT** be retrieved
- Multiple stations id assigned to the same station (in different datasets)
- Newly coined WIGOS identifiers needed
- Erroneous latitude for several “1759” stations

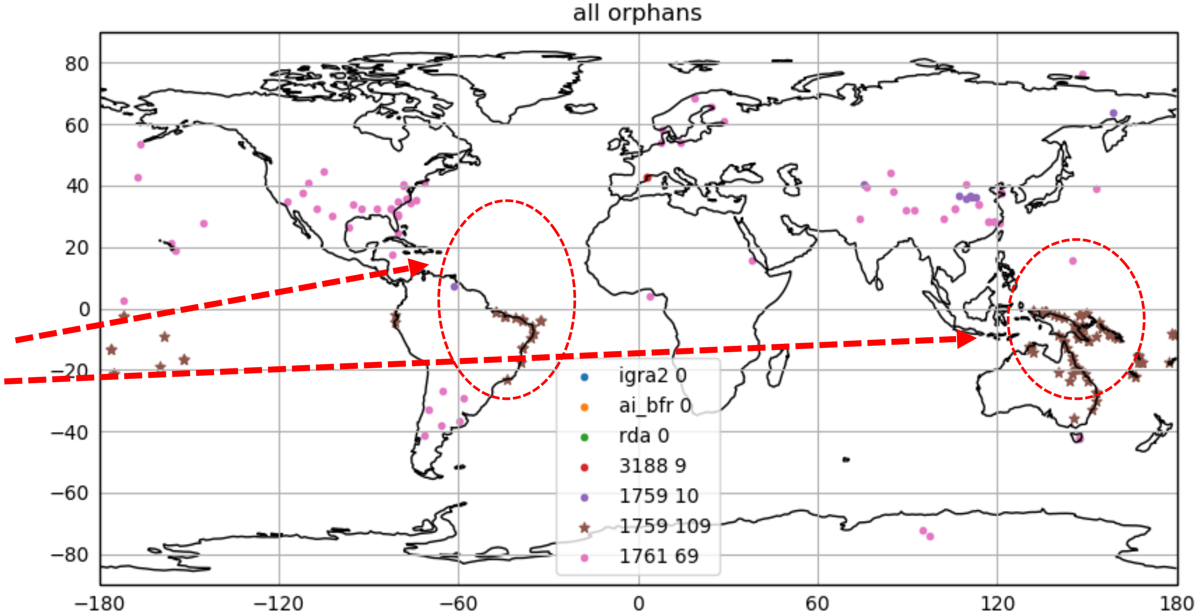
➢ Typically short data record
➢ Extra effort required for identification
➢ However: found inconsistency with latitude data
➢ -> minus sign in latitude missing



Identification of stations for WIGOS standards

- For a number of stations, WMO identifiers could **NOT** be retrieved
- Multiple stations id assigned to the same station (in different datasets)
- Newly coined WIGOS identifiers needed
- Erroneous latitude for several “1759” stations

➢ Typically short data record
➢ Extra effort required for identification
➢ However: found inconsistency with latitude data
➢ -> minus sign in latitude missing



3. Observation Uncertainty

- Early + many contemporary UA measurements do not contain uncertainty estimates
- Only dedicated networks have some, e.g. GRUAN
- Inter-comparison campaigns: difficult to find the original digital results

Goal: estimation of the upper air (UA) observation errors (measurement + representation) for *temperature*, *humidity* and *wind* from reanalyses departures —> Desrozier's diagnostics

Observation errors can be derived using
- Observation minus background ($\mathbf{d}_{O\text{-}B}$) departure
- observation minus analysis ($\mathbf{d}_{O\text{-}A}$) departure

* $$\left(\sigma^{oij}\right)^2 = \sum_{k=1,\ldots,p} (d_k^{oi} - d_k^{aj})(d_k^{oi} - d_k^{bj})/p$$

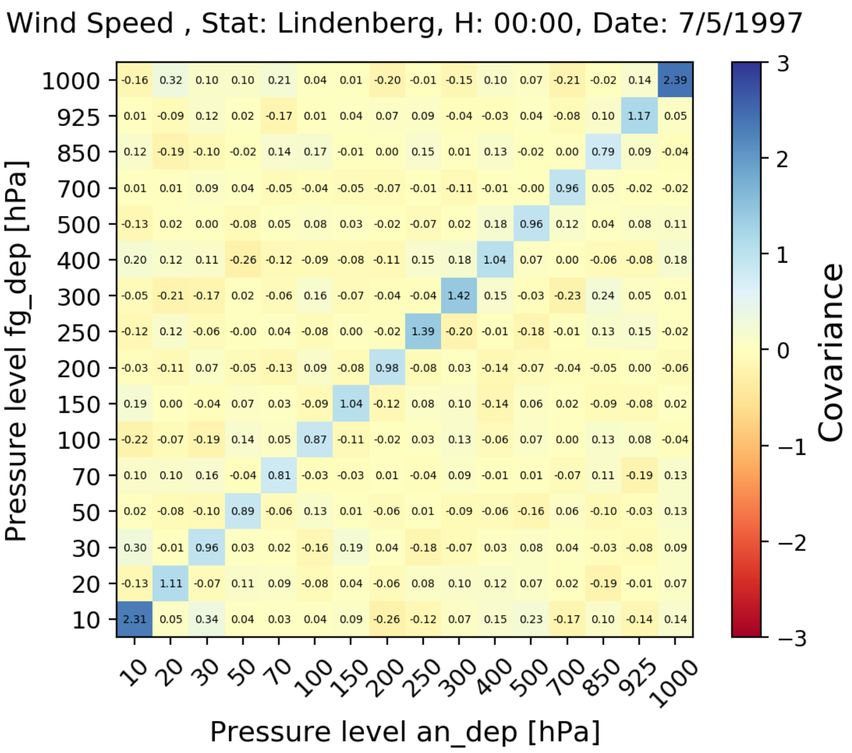
* $$E\left[\boldsymbol{d}_a^o\left(\boldsymbol{d}_b^o\right)^T\right] = \boldsymbol{R}$$

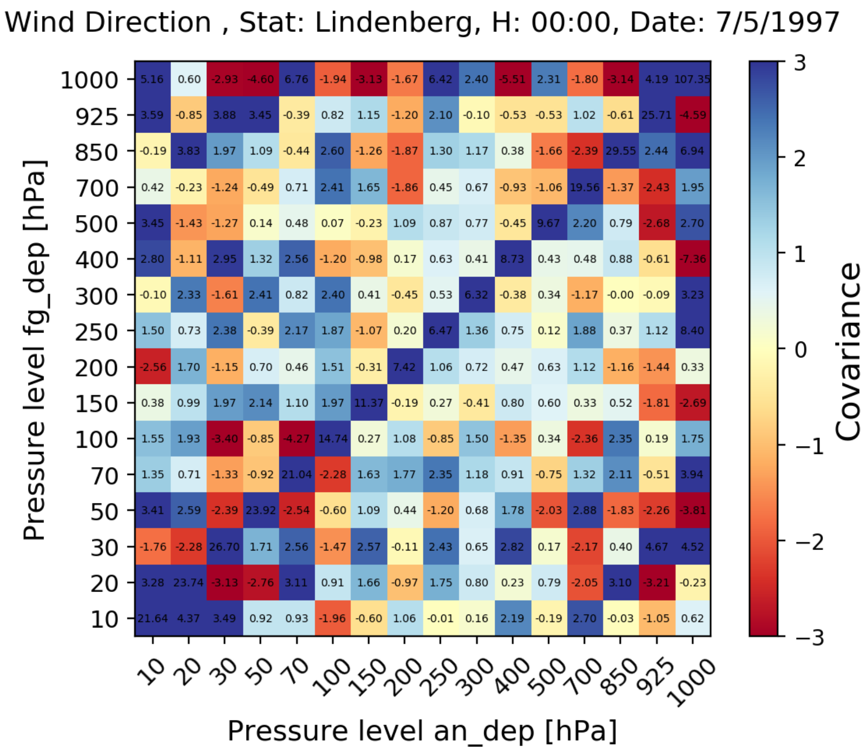
Expectation value (time-average) of the observation error

Errors cross-covariance matrix

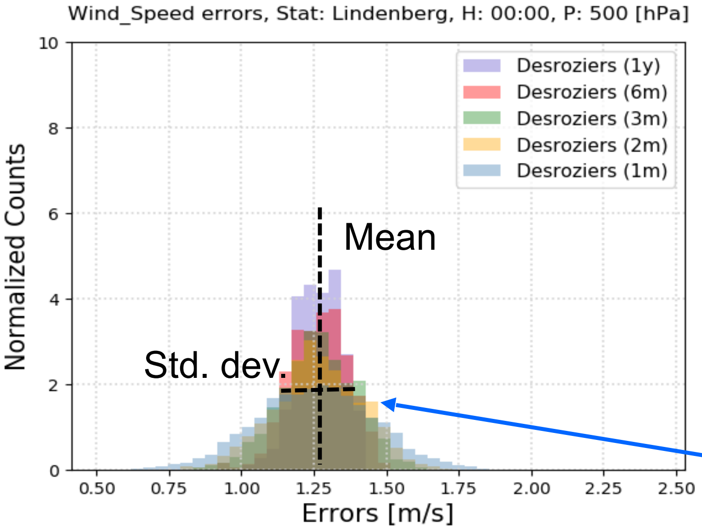
- Statistical method independent of metadata
- Can be extended to measurements back in time
- Include representation errors
- Can be compared with inter-comparison campaigns results

3. Observation Uncertainty: Desrozier's Diagnostics





3. Observation Uncertainty: Desrozier's Diagnostics

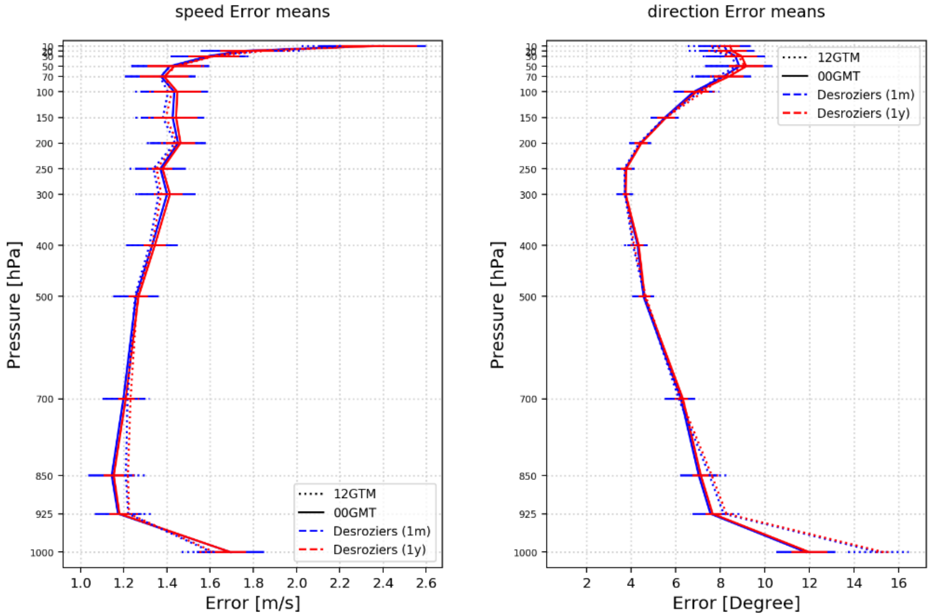


Distributions of the estimated error for the wind speed
(GMT00, 500 hPa pressure level)

- Mean value of the distributions ~independent on the time average
- Distribution more peaked if averaged over longer periods

Global results

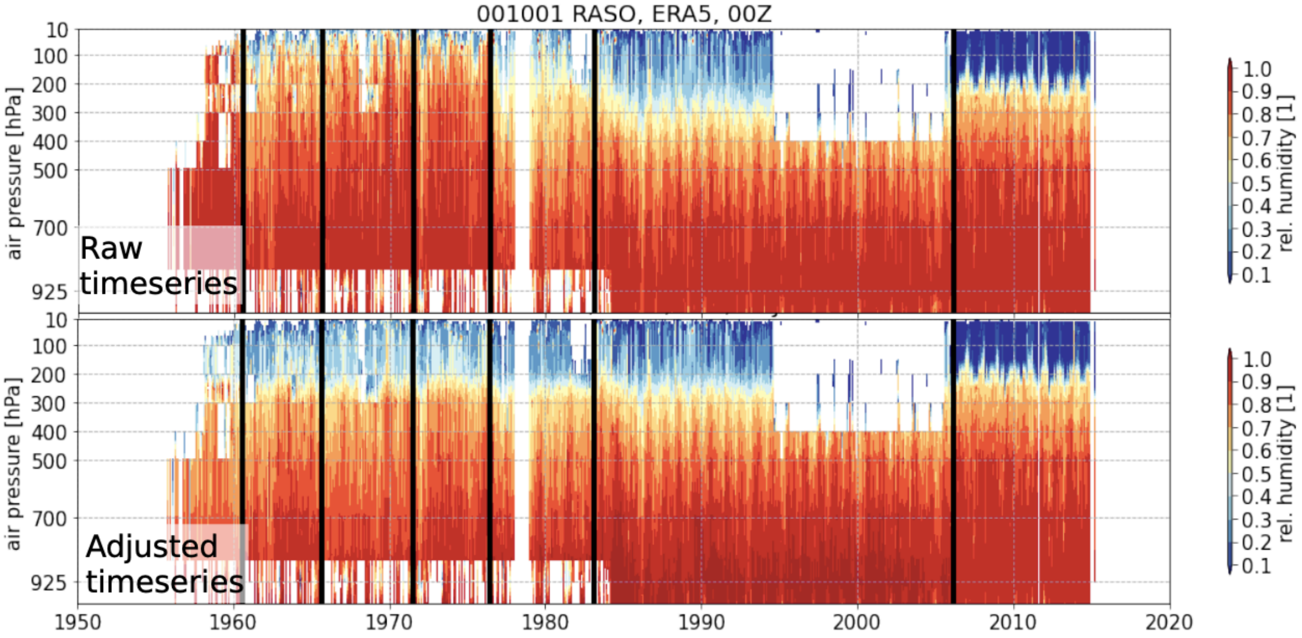
- Temperature error : 0.3-1.2 [K]
- Speed error : 1.1-2.4 [m/s]
- Direction error : 4-14 degree



4. Bias Correction and Homogenization

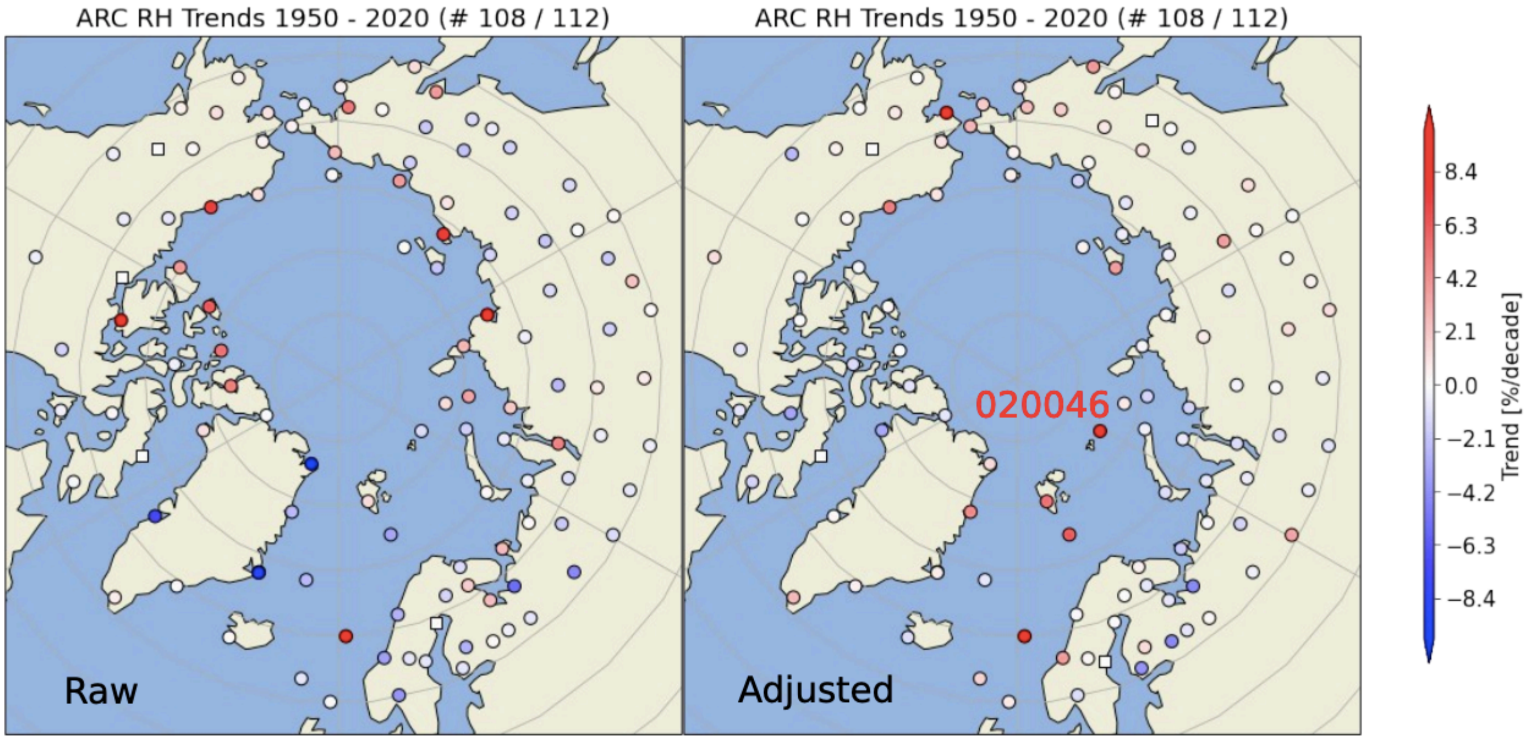
- Operational changes in sonde type, instrumentation, technology
- Problems: discontinuities, data gaps, sensitivity biases
- Needs: Adjustments – Employing a reanalysis based departure-statistics adjustment procedure
- Data gaps pose a challenge (can result in spurious trends)

from Michi Blaschek, @EGU 2020



4. Bias Correction and Homogenization

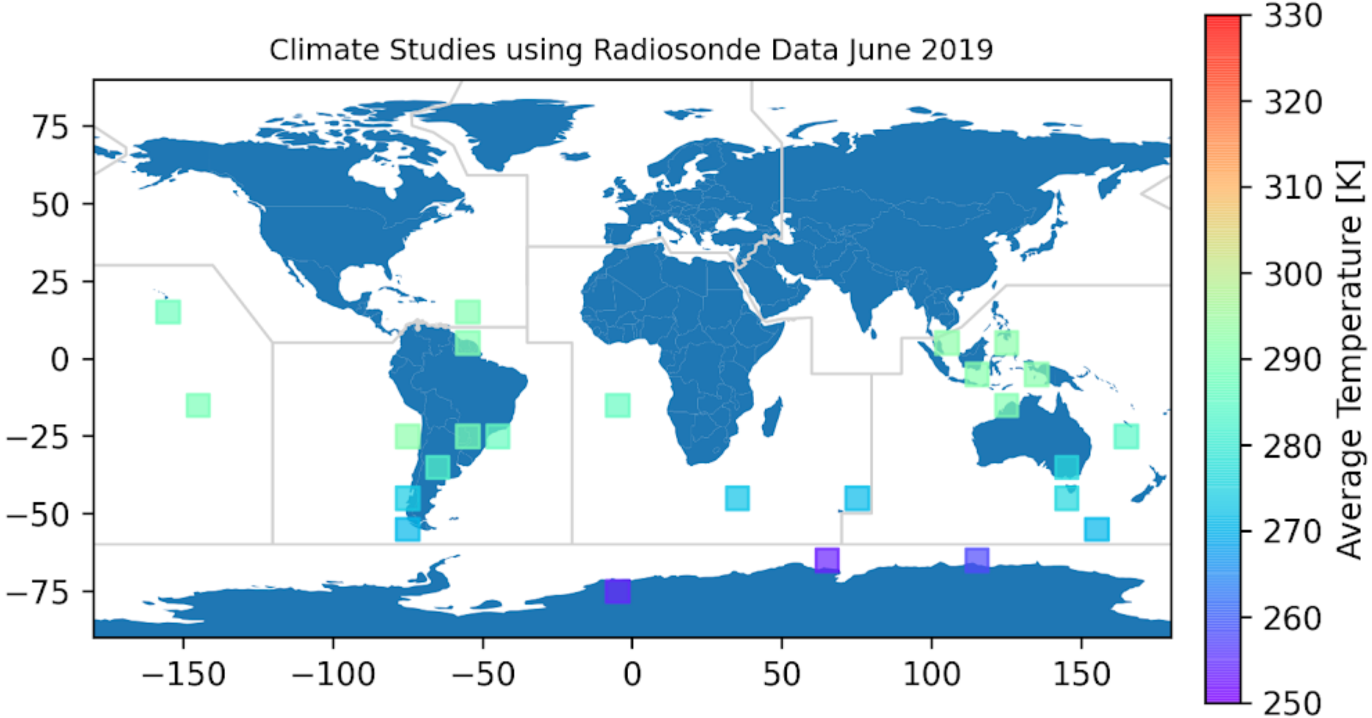
from Michi Blaschek,
@EGU 2020



5. Gridded Data

Monthly Gridded Data (ongoing, right now)

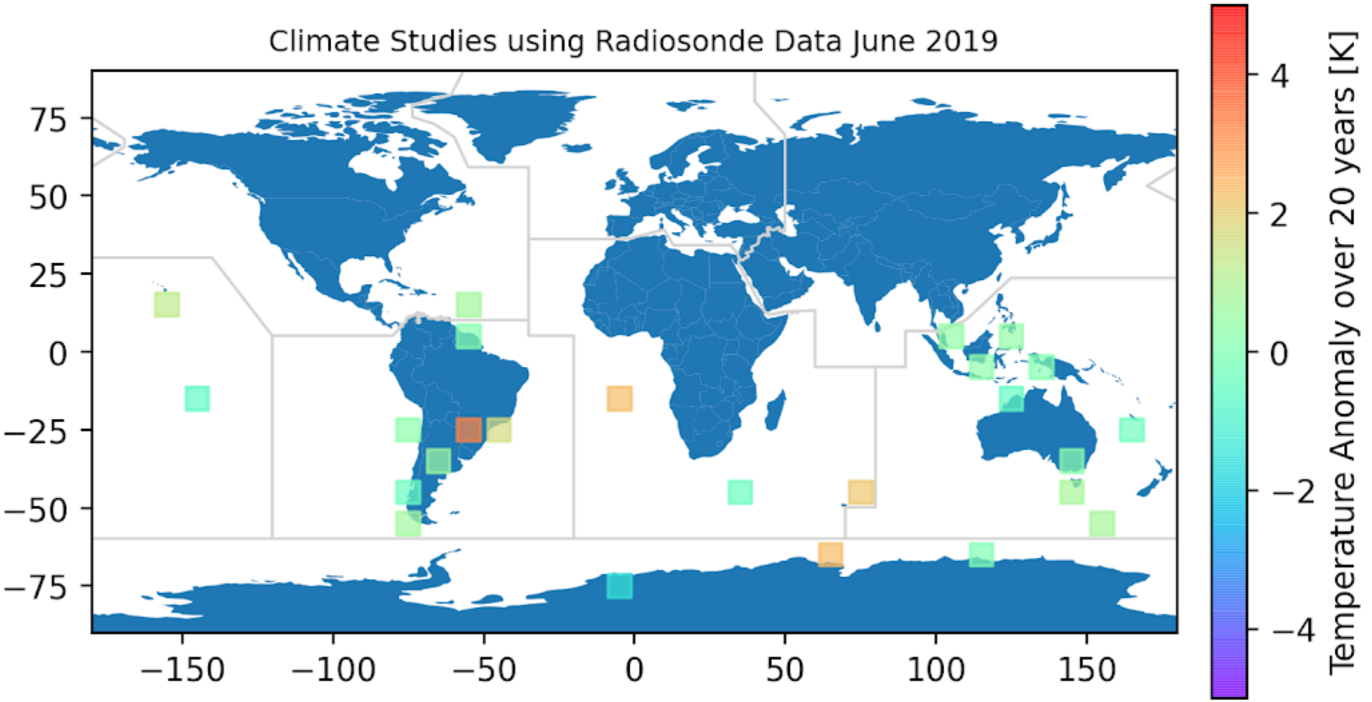
- Data for standard pressure levels
- Easy to compare with other data/studies



5. Gridded Data

Monthly Gridded Data (ongoing, right now)

- Data for standard pressure levels
- Easy to compare with other data/studies



6. Intercomparison Campaigns (with Ubern)

Radiosonde intercomparisons give independent information on systematic differences between radiosonde types.

- The calibration of different sensors (pressures: [1000,5] hPa, temperature: [-90,+50] C, relative humidities [~10,100] %), plus solar radiation etc. has always been challenging
- International intercomparison campaigns allowed to assess systematic differences of observed meteorological variables measured with different sensors

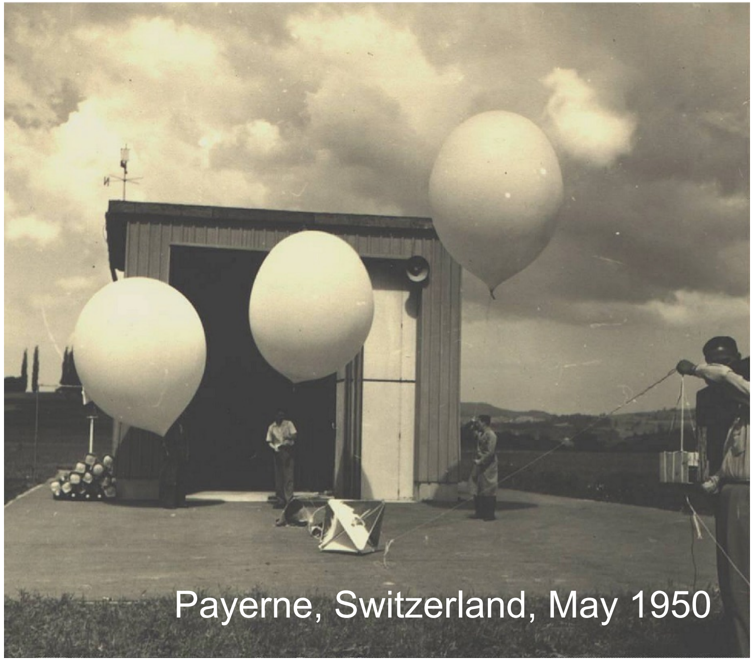
Sources of inter comparison data:

- list of global (WMO organized), regional, and national intercomparison campaigns
- Archive search at the MeteoSwiss aerological station in Payerne (Switzerland): raw data from some of the first international intercomparison campaigns
- At the WDC/RIHMI Obninsk (Russia): national and international intercomparison campaigns held in the Former Soviet Union (FSU), important thanks to the large coverage of the soviet radiosondes network

Data from 15 international radiosonde intercomparisons [1935-1993] have been digitized or collected by UNIBE and RIHMI

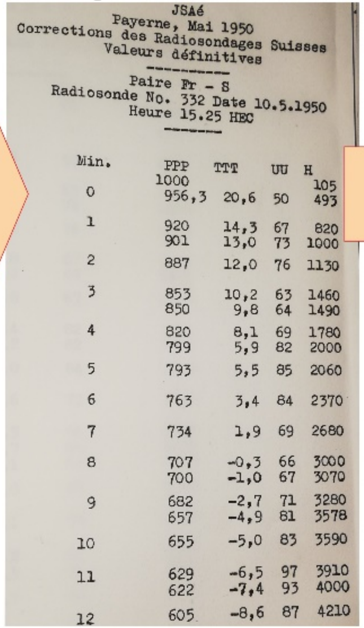
6. Intercomparison Campaigns (with Ubern)

Digitizing and building a database of historical Int. Radiosonde Intercomparison campaigns available to the users —> input to estimating data uncertainties



Payerne, Switzerland, May 1950

Imaged Data Sheet



JSAé
Payerne, Mai 1950
Corrections des Radiosondages Suisses
Valeurs définitives

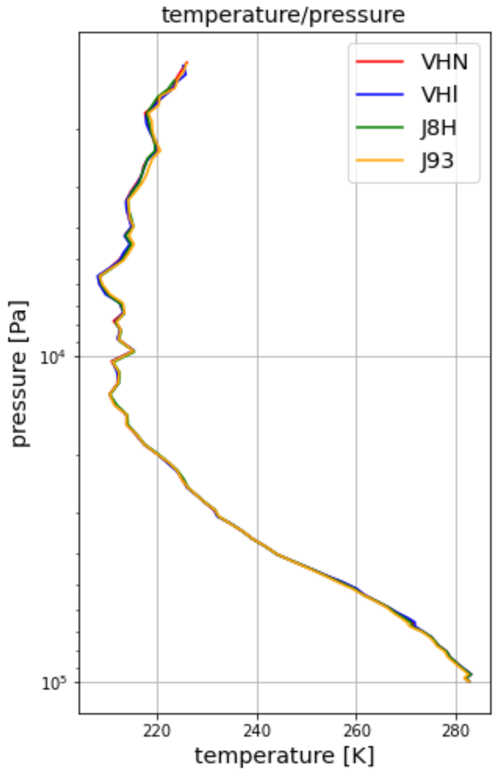
Paire Fr - S
Radiosonde No. 332 Date 10.5.1950
Heure 15.25 HEC

Min.	PPP	TTT	UU	H
	1000			105
0	956,3	20,6	50	493
1	920	14,3	67	820
	901	13,0	73	1000
2	887	12,0	76	1130
3	853	10,2	63	1460
	850	9,8	64	1490
4	820	8,1	69	1780
	799	5,9	82	2000
5	793	5,5	85	2060
6	763	3,4	84	2370
7	734	1,9	69	2680
8	707	-0,3	66	3000
	700	-1,0	67	3070
9	682	-2,7	71	3280
	657	-4,9	81	3578
10	655	-5,0	83	3590
11	629	-6,5	97	3910
	622	-7,4	93	4000
12	605	-8,6	87	4210

Digitised Profile

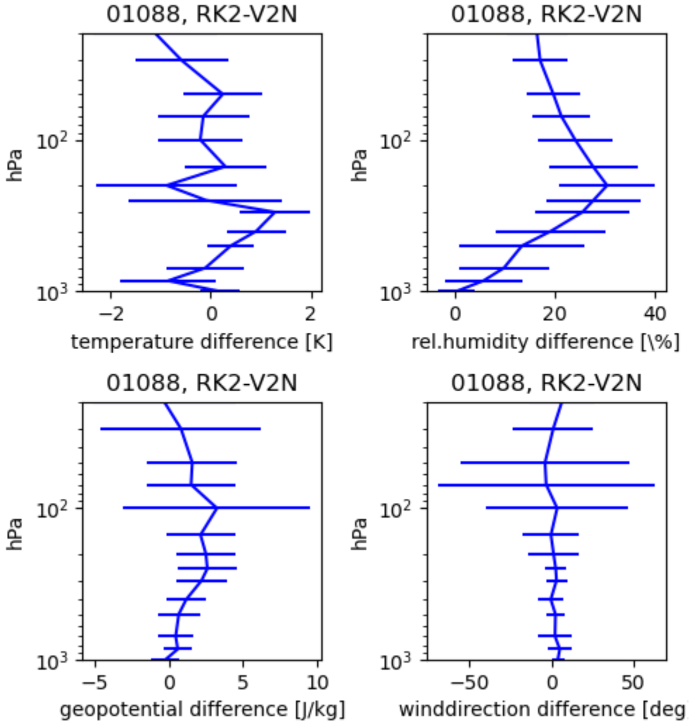
E	F	G	H	I
Min.	**PPP**	**TTT**	**UU**	**H**
	1000			105
0	956.3	20.6	50	493
1	920	14.3	67	820
	901	13	73	1000
2	887	12	76	1130
3	853	10.2	63	1460
	850	9.8	64	1490
4	820	8.1	69	1780
	799	5.9	82	2000
5	793	5.5	85	2060
6	763	3.4	84	2370
7	734	1.9	69	2680
8	707	-0.3	66	3000
	700	-1	67	3070
9	682	-2.7	71	3280
	657	-4.9	81	3578
10	655	-5	83	3590
11	629	-6.5	97	3910
	622	-7.4	93	4000
12	605	-8.6	87	4210

6. Intercomparison Campaigns (with Ubern)



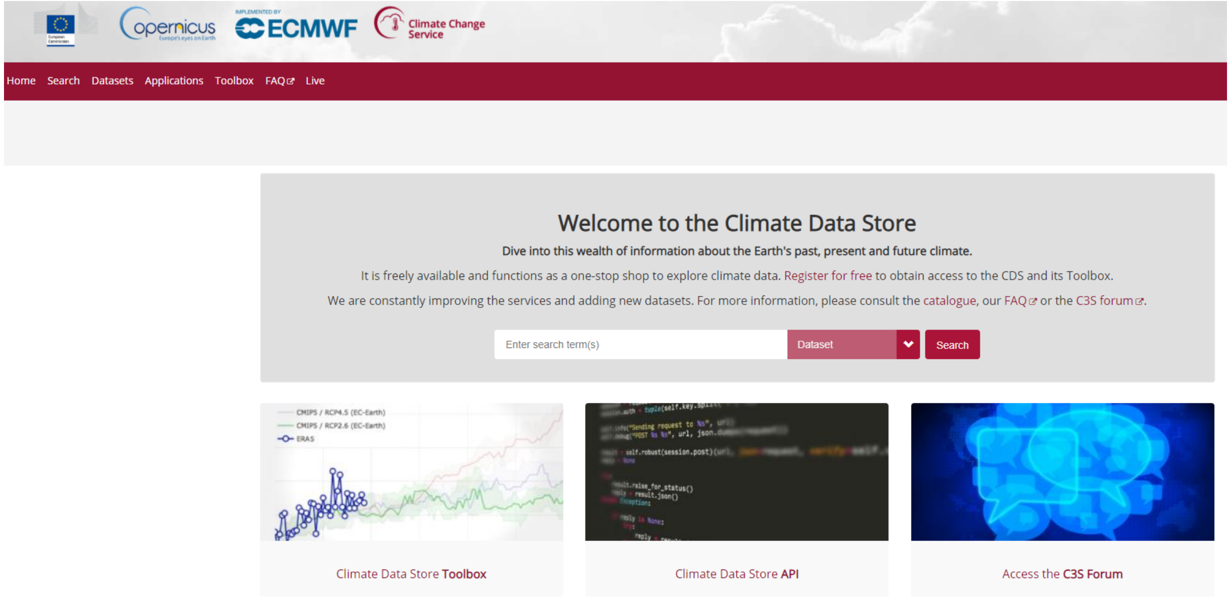
Example of a single ascent during the WMO Phase IV (Tsukuba) campaign 1993

Mean differences between ascents of USSR RKZ-2 (RK2) and Vaisala RS21-12CN (V2N) radiosondes during a regional intercomparison in 1984. Errors: sample standard deviations.



7. CDS - Climate Data Store Interface

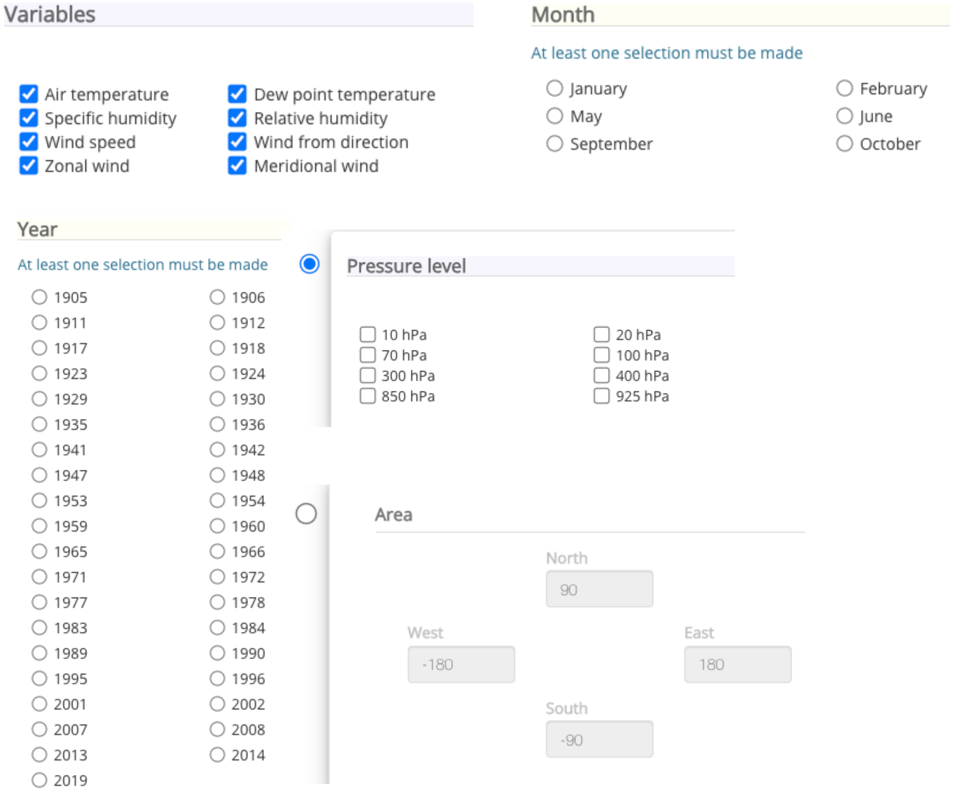
- Our data resides in a dedicated virtual machine at ECMWF
- Currently, data can be downloaded only from the CDS-develop interface (private)
- We are responsible for the implementation of the front- and backend functionalities
- ***“Comprehensive upper-air observation network from 1905 to present”***



MAIN VARIABLES		
Name	Units	Description
Eastward wind component	m s-1	Atmospheric eastward wind component at given pressure level
Geopotential height	m	Geopotential altitude from corrected pressure product
Northward wind component	m s-1	Atmospheric northward wind component at given pressure level
Relative humidity	%	Ratio of actual water vapor pressure to saturation water vapor pressure
Specific humidity	Kg/kg	Mass fraction of water vapor in atmospheric volume
Temperature	K	Atmospheric temperature at given pressure level
Wind from direction	Deg (0=N, 90=E)	Atmospheric wind direction at given pressure or height level, following the metorological convention, i.e. wind direction is wind from direction, it increases clockwise and starts with 0 deg if wind comes from North
Wind speed	m s-1	Atmospheric wind speed at given pressure or height level

7. CDS - Climate Data Store Interface

… using the web interface widgets



… or python API

```
import cdsapi
c = cdsapi.Client()
c.retrieve(
    'insitu-comprehensive-upper-air-observation-network',
    {
        'variable': 'air_temperature',
        'year': '1905',
        'day': [
            '01', '19', '31',
        ],
    },
    'download.data')
```

To Do (end of the project Copernicus 1.0: June 2021)

- Extension of the database with additional data/metadata, + gridded data for easier comparison with other CDS products

- Quality checks

- Data assimilation ready —> input for ERA6 reanalysis

- Revised homogenisation procedure (temperature, humidity and wind)

- Comparison with other upper-air or satellite data

- Complete integration of the dataset within the CDS, add toolbox functionalities

Thank You !

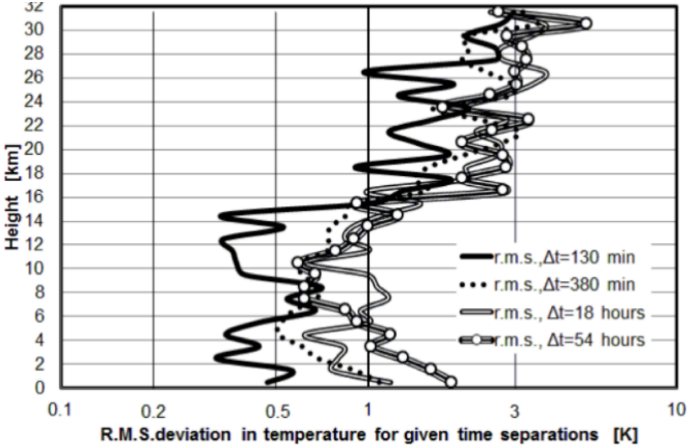
Backup

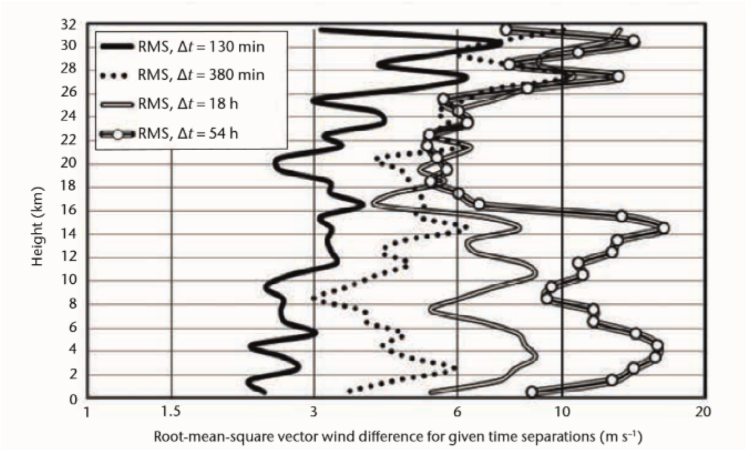
3. Observation Uncertainty

- Early + many contemporary UA measurements do not contain uncertainty estimates
- Only dedicated networks have some, e.g. GRUAN
- Inter-comparison campaigns: difficult to find the original digital results

Goal: estimation of the upper air (UA) observation errors (measurement + representation) for *temperature*, *humidity* and *wind* from reanalyses departures

Example: RMS (root mean square) deviations from the Yangjiang inter-comparison campaign (China,2011) for different time intervals





3. Observation Uncertainty: Desrozier's Diagnostics

Observation errors can be derived using
- Observation minus background ($\mathbf{d_{O\text{-}B}}$) departure
- observation minus analysis ($\mathbf{d_{O\text{-}A}}$) departure

* $$\left(\sigma^{oij}\right)^2 = \sum_{k=1,\dots,p} (d_k^{oi} - d_k^{aj})(d_k^{oi} - d_k^{bj})/p$$

Assumption: background errors and observation errors are independent ,
and "optimal assimilation systems" (i.e. correct background errors)

Desroziers Diagnostics (Desroziers 2005,2011)

* $$E\left[d_a^o\left(d_b^o\right)^T\right] = \boldsymbol{R}$$

Expectation value (time-average) of the observation error

Errors cross-covariance matrix

- Statistical method independent of metadata
- Can be extended to measurements back in time
- Include representation errors
- Can be compared with inter-comparison campaigns results

→ Used **ERA5** database which include background and analysis departures