The UERRA approach and results to enhance data availability and accessibility to support European Regional Reanalysis of high-resolution

By Joan Ramon Coll[1], M. Brunet[1], L. Ashcroft[1], A. Gilabert[1], P. Domonkos, E. Aguilar[1], J. Sigró[1], P.D. Jones[2], Ian Harris[2] and P. Undén[3]

1.Centre for Climate Change (C3), University Rovira i Virgili, Vila-seca, Catalonia

2. Climatic Research Unit (CRU) University of East Anglia, Horwich, UK

3. Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

11th EUMETNET Data Management Workshop. Placing climate data to social service: From observations to archives, 18-20 October 2017, Zagreb, Croatia

Outline

1. The effort undertaken in UERRA; The DARE approach

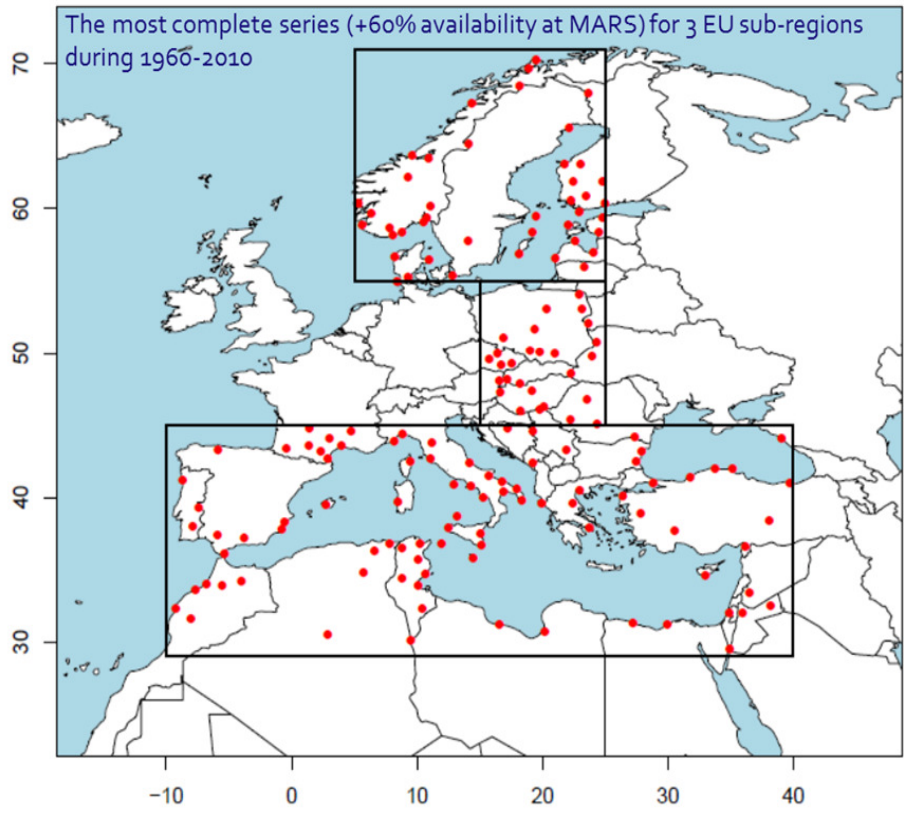
2. A summary of UERRA's digitisation & gathered data sources effort

3. Data development:
    - 3.1. The Automatic Quality Control (AQC); tests and results
    - 3.2. The Spatial Consistency Check (SCC); tests and results

4. Conclusions

1. The effort undertaken in UERRA: the data rescue (DARE) approach

- <u>Setting the targets</u>:
  - Produce a comprehensive list of historical climate data holders and sources with relevant un-digitised, but imaged data across Europe.
  - Exploring MARS Archive & cross-checking against the gathered data-sources to set the spatial-temporal gaps in climatic variables at the synoptic scale.
  - Coordination with other data rescue activities to avoid duplication (e.g. Exploring ERACLIM2 registry & ISPD & coordination with ISTI-DARE, Meteo-France, ACRE, UBERN, JLU Giessen).
- <u>The targets for digitisation</u>:
  - Mainly spatial infilling over data-sparse sub-regions: **Southern part of Med, Eastern Europe & Balkans**
  - Variables: **PP, TT, FF, WD, RH, DP**, ***SD, FS, RR*** at the **hourly** and ***daily*** scales
  - Mostly for the **post-1950**, but also pre-1950

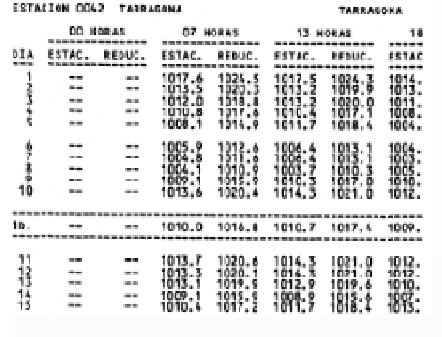


1. The effort undertaken in UERRA: the data rescue (DARE) approach

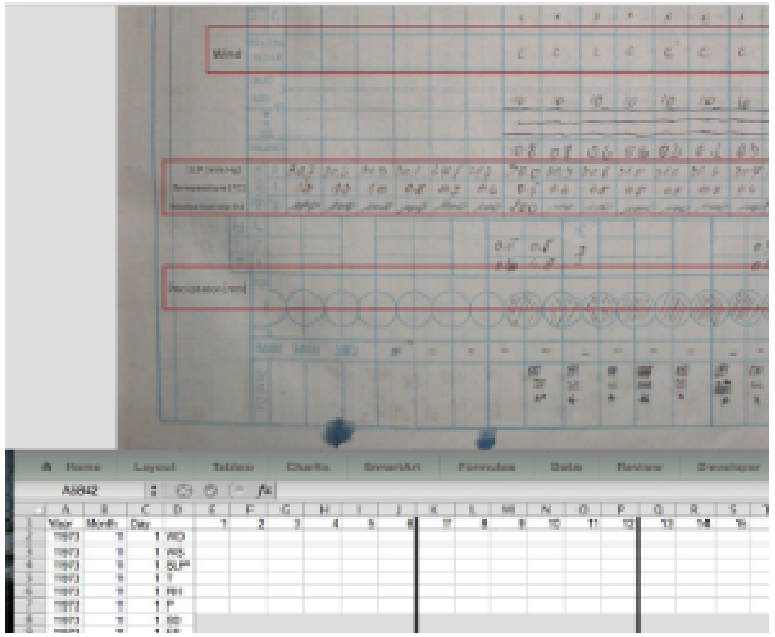
- <u>Setting the digitisation plan</u>
  - Contracting 11 science students (2 years), training & elaborating tutorials for a more efficient digitisation.
  - Produce **spreadsheet templates** that match the data source to reduce digitisation errors.
- <u>Ensuring the quality of the digitisation work</u>: **Visual cross-checking** (data source vs data digitised) to minimise errors. Two steps:
  - Step 1: **Digitisers self-assessment:** the 10th, 20th and 30th of each month crosschecked; Monthly means calculated and compared with source means, when possible; All missing data and illegible observations recorded in metadata
  - Step 2: **Systematic checks**: 1st, 15th and 25th (3 days) of Jan, April, Aug and Dec (4 months) checked every five (5) years;
  - Common **errors** identified and **reported back to digitisers**. E.g: Columns and rows miss-aligned; Missing data recorded as blank
  - **The 95,7% of skewed values found in the UERRA rescued dataset were corrected and the 4,3% were set to missing**

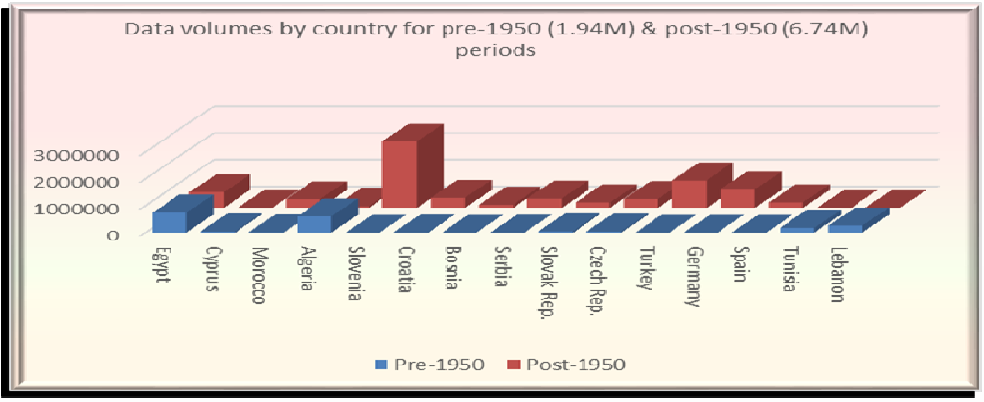
Egypt, 4–5 September 1938

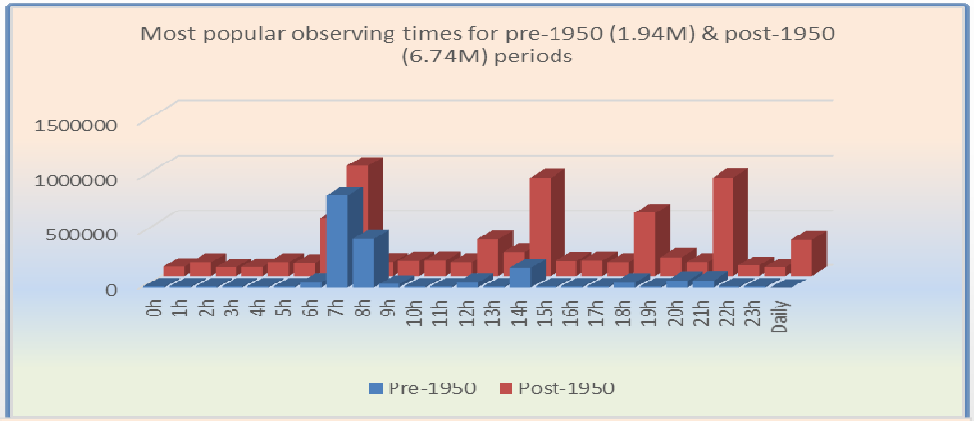
c) Tarragona, Spain, July 1977.

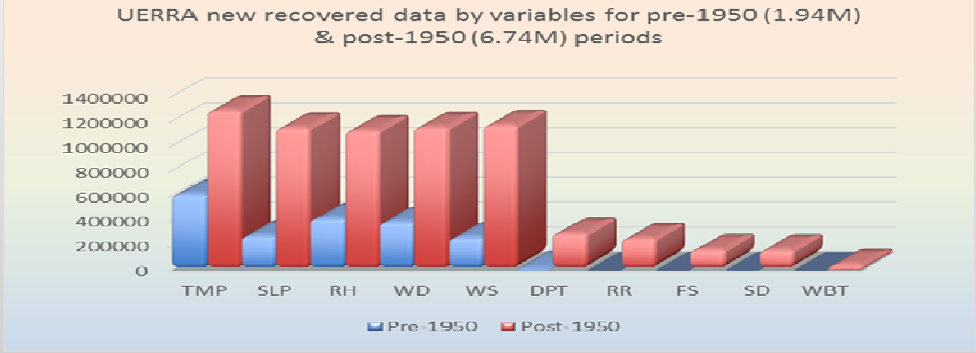


Lubljana, Slovenia, 1 January 1973

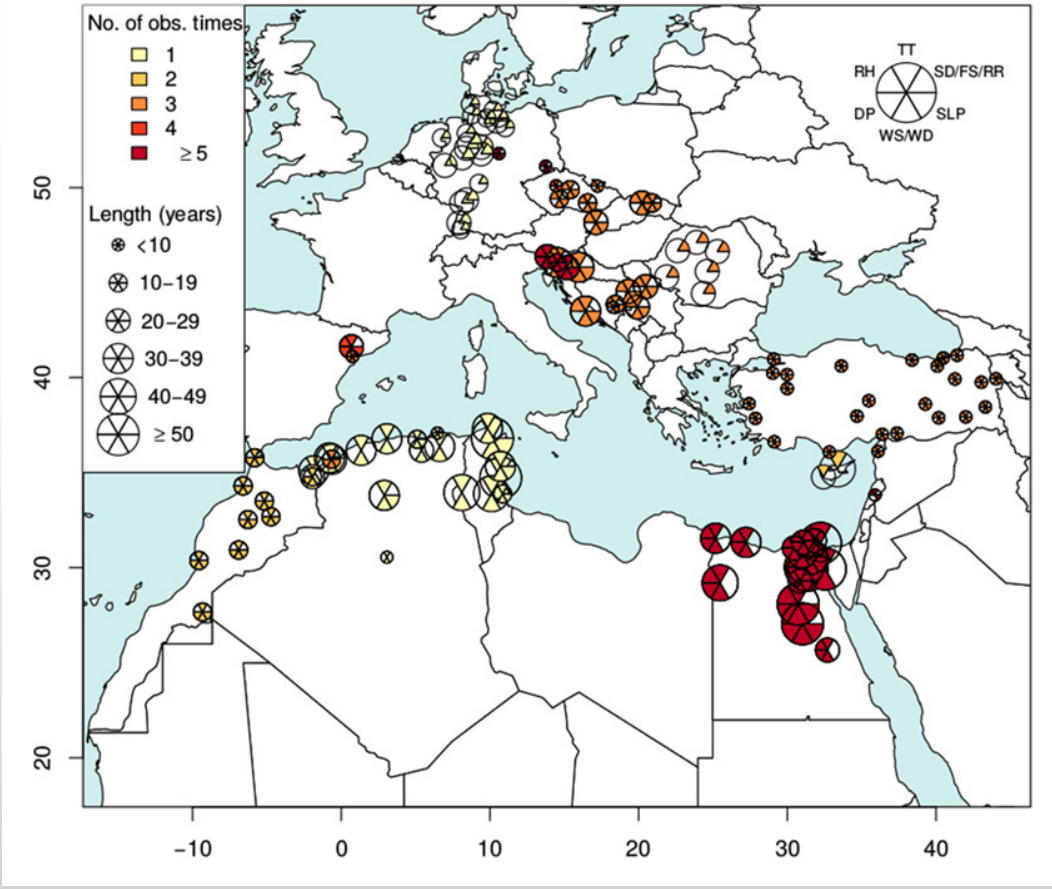








2. Summarising UERRA's digitisation efforts (~9M of new observations)



2. Summarising the UERRA gathering effort: 178.1M of digitised data to support ERRAs

Table 1. Statistics of data provided by MetNo, SMHI and MeteoCat for UERRA WP1. Variable acronyms represent air temperature (TT), atmospheric pressure (SLP), rainfall (RR) relative humidity (RH), cloud cover (CC), snow depth (SD), wind direction (WD) and wind speed(WS).

Provider	Number of stations	Time period covered	Variables provided	Frequency of observations	Number of total observations
SMHI	146	1945–2009	TT, SLP, RR, RH, SD, CC	Precipitation daily, other observations from 3 times daily to hourly	42.0 million
MetNo	93	1960–1980	TT, WD, WS, RR	Generally 3-4 times a day, some stations hourly	7.2 million
MeteoCat	76	1988–2015	TT, SLP, WD, WS, RR, RH	Hourly (SD daily)	128.9 million
Total	315				178.1M

3. DATA DEVELOPMENT: TWO APPROACHES

- The Automatic Quality Control (AQC)

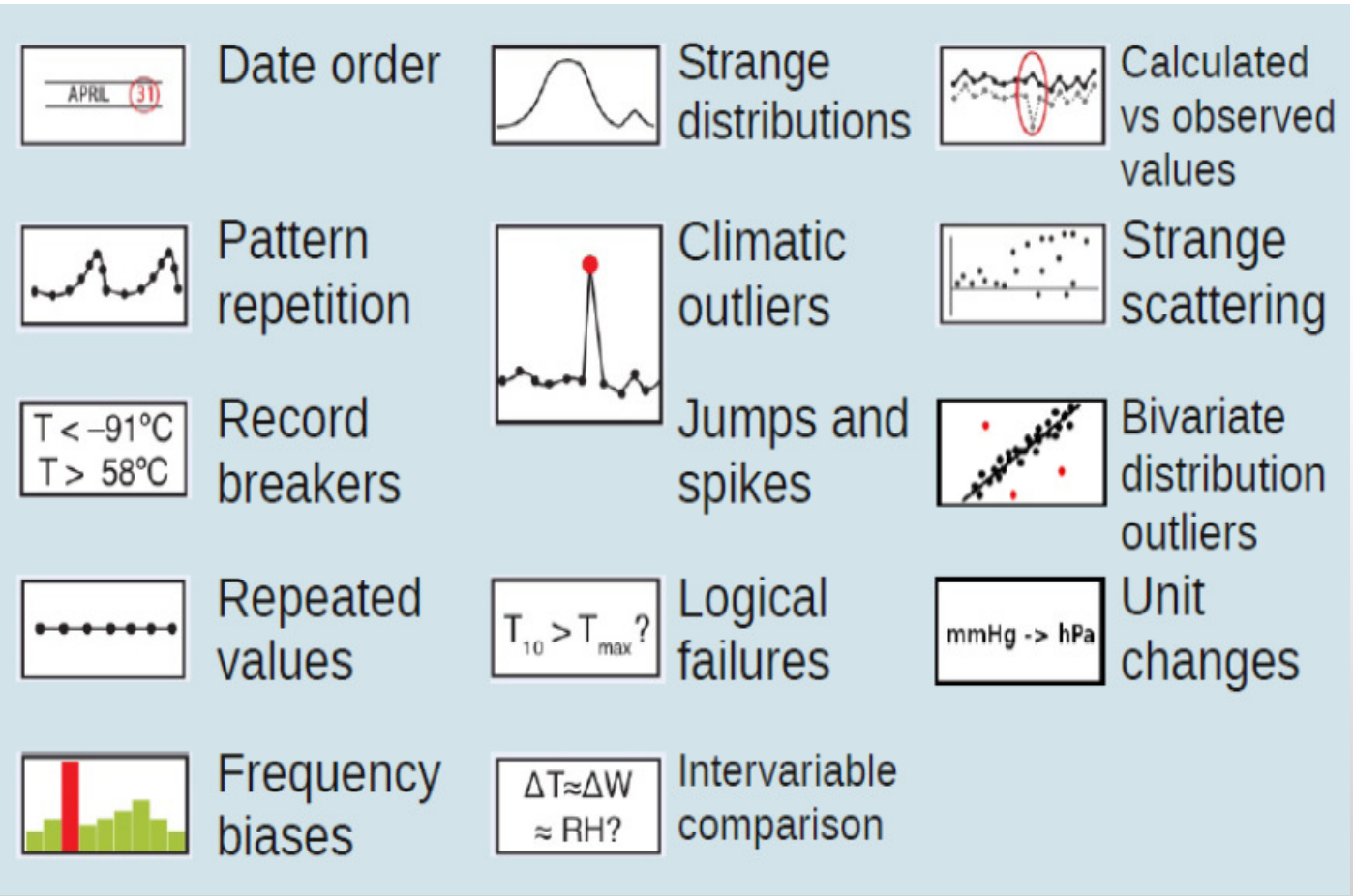
Main purpose: Detect and correct/remove intra-station errors from the UERRA rescued dataset by using a battery of efficient QC tests specifically designed by C3/URV to check temporal consistency at synoptic scale for TT, PP, DP, RH, WD, FF, RR, FS and SD.

- The Spatial Consistency Check (SCC)

Main Purpose: Ensure the quality of the UERRA rescued dataset by checking the spatial consistency once temporal consistency is already assessed.
A battery of automatic QC tests were applied to the dataset by using the methodology developed by Dunn et al., 2012 (adapted). Intra- and inter-station consistency were checked at sub-daily scale for TT, PP, DP, WD and WS.

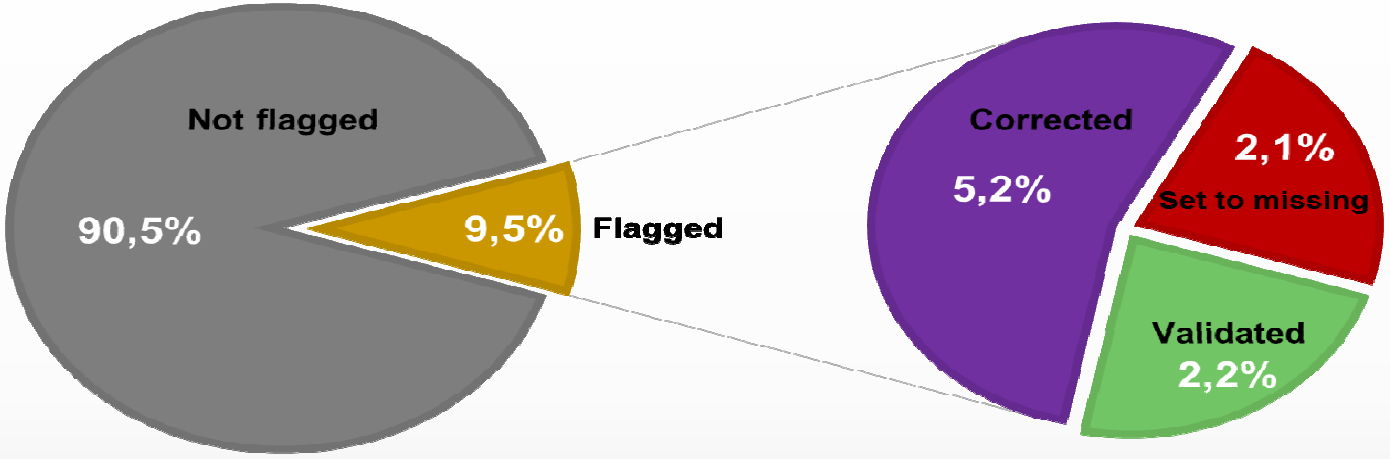
RJH Dunn, KM Willett, PW Thorne, EV Woolley, I Durre, A Dai, DE Parker, RS Vose. HadISD: a quality controlled global synoptic report database for selected variables at long-term stations from 1973-2010. Climate of the Past Discussions 8, 1763-1833 (2012) http://www.clim-past.net/8/1649/2012/cp-8-1649-2012.html

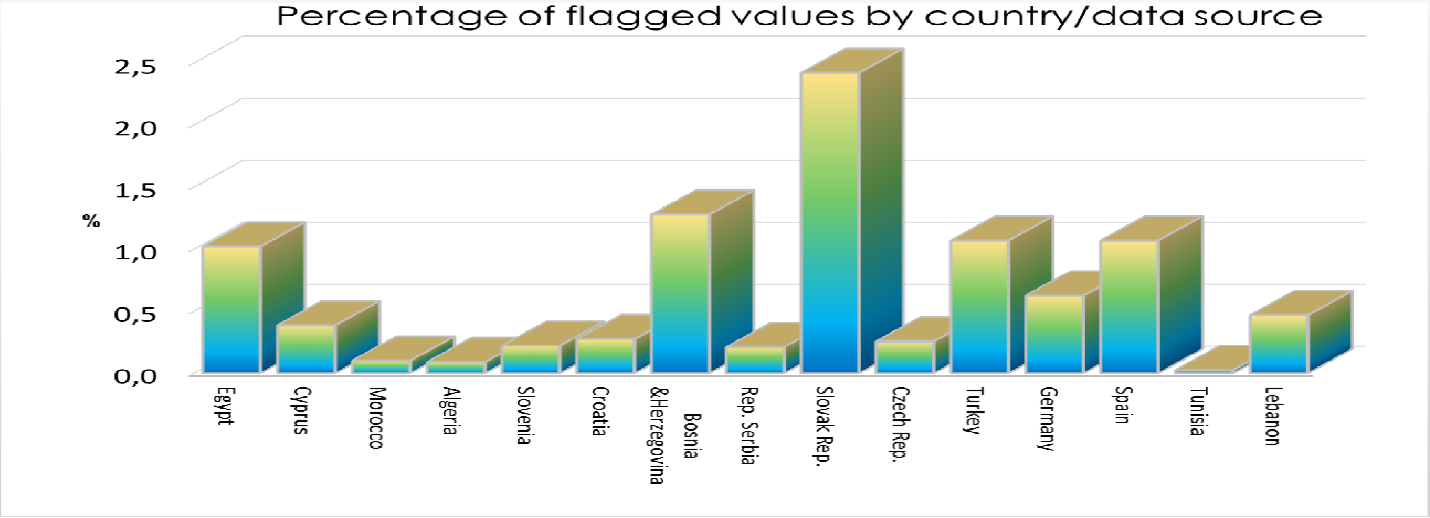
3.1. The Automatic Quality Control (AQC) tests and results



- 14 tests to identify & label suspicious values & chains of values
- Checking the data-sources to validate, correct or set to missing the labelled values (time consuming, but useful exercise)

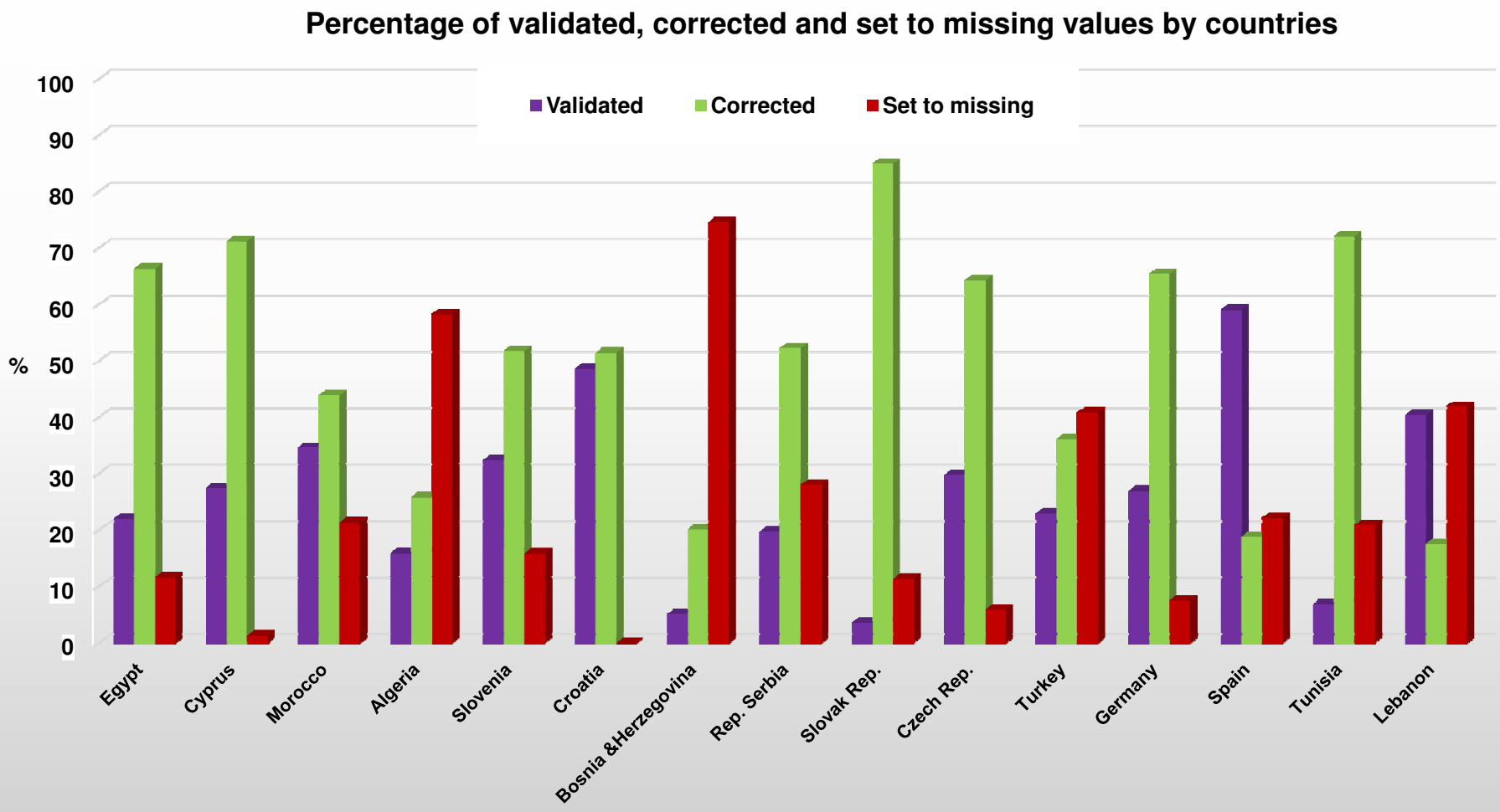
3.1. The Automatic Quality Control (AQC) tests and results





- A remarkable amount of labelled values as potentially wrong
- Half of them recovered & corrected after data-source examination
- A ¼ false positive & then validated & another ¼ set to missing
- Data source format impact on the fraction of labelled values

3.1. The Automatic Quality Control (AQC) tests and results



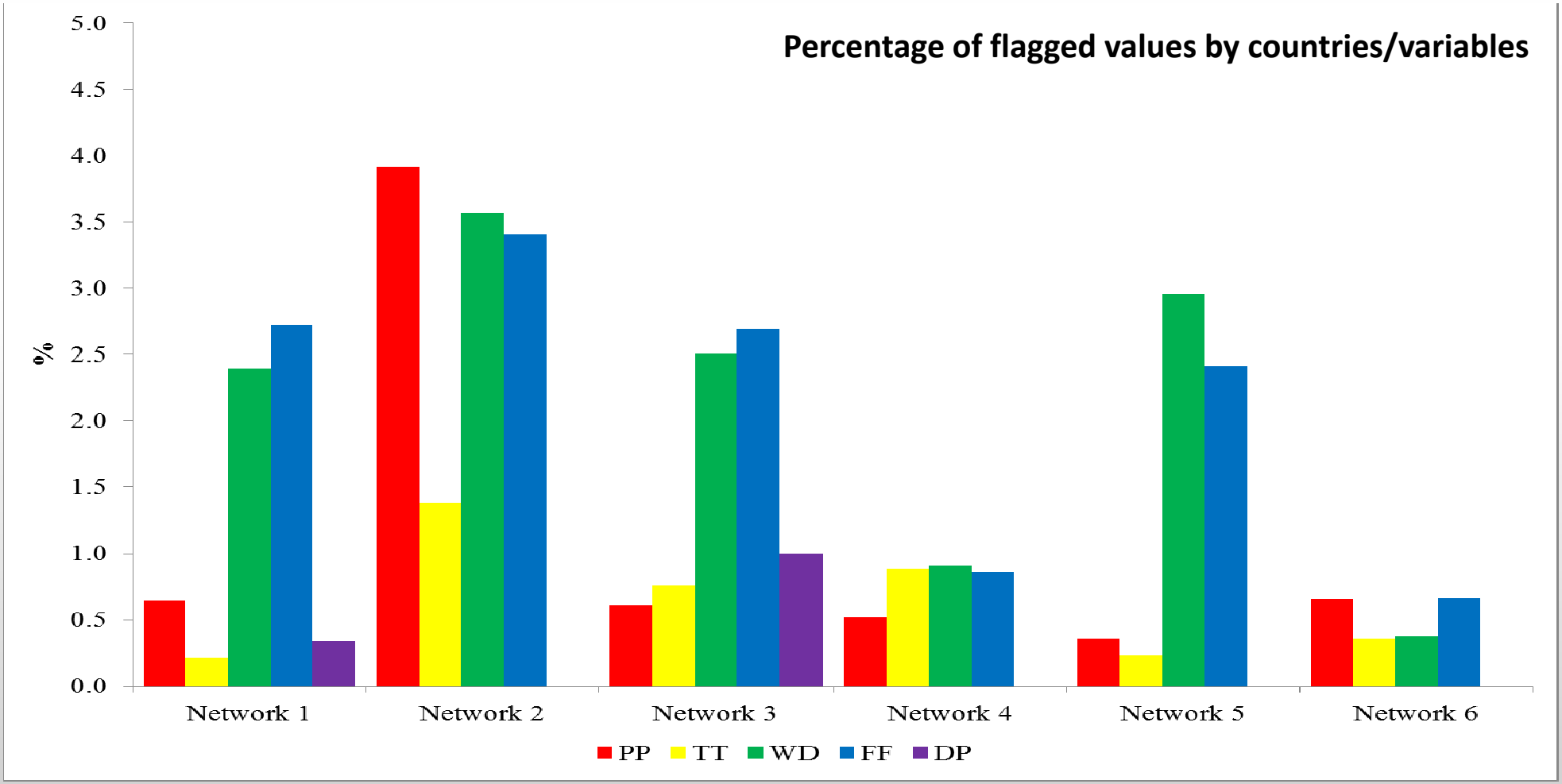
3.2. The Spatial Consistency Check (SCC) tests and results

Network	Countries	Nº of stations	Period	Variables	Observing times
1	MAR	8	1953-1968	PP, TT, WD, FF DP	6 or 7h/18h (2)
2	ALG and TUN	25	1886-1938	PP, TT, WD, FF	7h (1)
3	EGY	21	1907-1957	PP, TT, WD, FF DP	6h or 8h/12h or 14h/18h or 20h (4)
4	TUR	25	1962-1971	PP, TT, WD, FF	7h/14h/21h (3)
5	SLO, CRO, BOH, SER	10	1950-2012	PP, TT, WD, FF	7h/14h/21h (3)
6	SLV, CZE, GER (2 stations)	11	1948-1968	PP, TT, WD, FF	7h/14h/21h (3)

QC tests applied among others:

- Inter-station duplicate check: Each time series is compared iteratively with that of every other time series.
- Frequent value check: Detect far more observations of a given value than expected (e.g. specific codes,…).
- Distributional gap check: Detect erroneous portions of time series by computing monthly means and thresholds.
- Nearest neighbour data checks: A difference series is created for each candidate station minus neighbour pair. Any observation associated with a difference exceeding an established threshold of the whole difference series is flagged.

3.2. The Spatial Consistency Check (SCC) tests and results



4. Conclusions

- About 9M of station synoptic values were rescued over southern Mediterranean basin, Eastern Europe and Balkans region for TT, PP, RH, WD, FF, DP, RR, FS and SD mostly for the post-1950.

- Most of data errors were produced in digitization procedure. Training digitizers, having a strong feedback as well as the use of spreadsheets templates that match data sources were crucial to reduce typing errors.

- About 178M of digitized data provided by MetNo, SMHI and MeteoCat were gathered to support European Regional Reanalysis.

- The application of Automatic Quality Control tests and Spatial Consistency Checking ensured temporal and spatial data consistency to the new high-quality UERRA rescued dataset.

- The incorporation of about 9M of high-quality station synoptic values together with 178M of gathered data to the European Regional Reanalysis will surely help to produce accurate ensembles and estimate the associated uncertainties.

Thanks for your attention

Email: joanramon.coll@urv.cat

UERRA
Uncertainties in Ensembles of Regional ReAnalyses

C3 Centre for Climate Change

Universitat Rovira i Virgili
Campus Terres de l'Ebre