Ensuring quality and spatial consistency of the UERRA rescued dataset: Procedures and results

Work Package I

by Joan Ramon Coll
Linden Ashcroft, Manola Brunet, Enric Aguilar,
Javier Sigró and Alba Gilabert

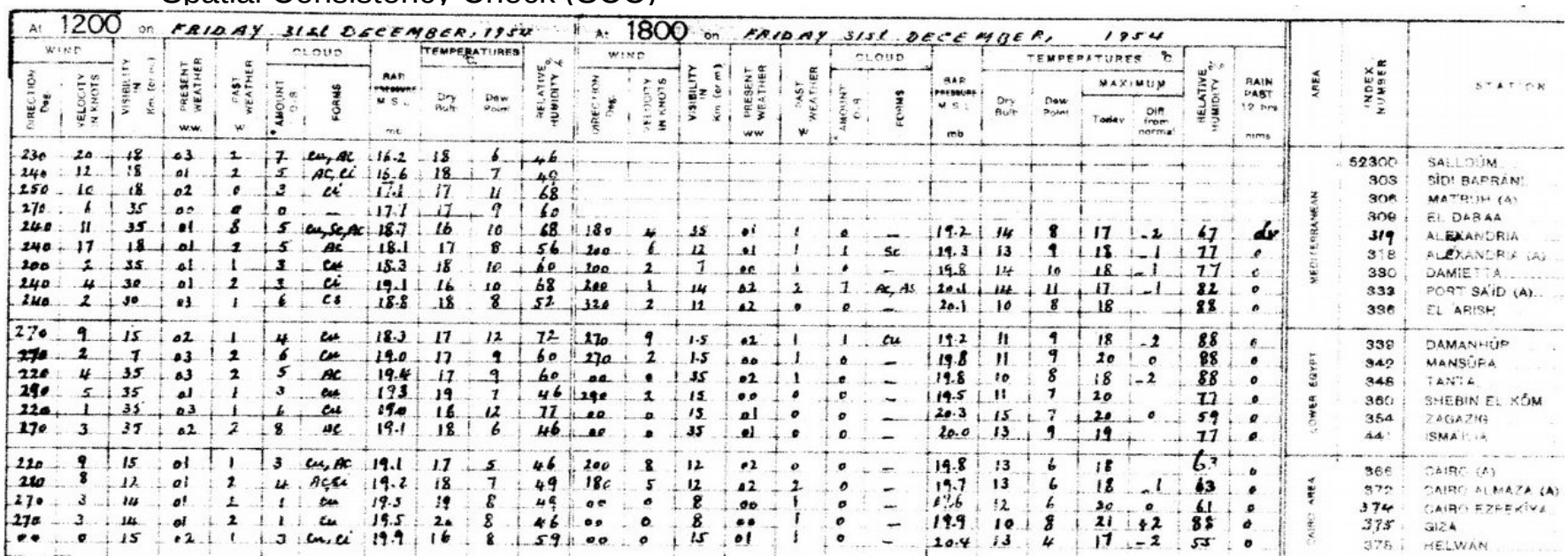
Centre for Climate Change (C3), University Rovira i Virgili, Tortosa, Catalonia

The UERRA General Assembly 4 (GA4) and Project Meeting, 21-23 November 2016, ECMWF, Reading, UK

Contents

1. Introduction

- The Quality Control (QC) procedure is crucial to identify errors derived from original sources, digitisation process, data manipulation or transmission.
- Type of errors:
  - Source error
  - Typing error
  - Skewed values
  - Hard to read or misread
  - Confusions (incorrect station, variable, observation time,…)
- The QC applied to the UERRA rescued dataset is divided in 3 steps;
  - Visual Quality Control (VQC)
  - Automated Quality Control (AQC)
  - Spatial Consistency Check (SCC)



2. Procedures I; Visual Quality Control (VQC)

- **Main purpose:** Detection and correction of skewed values introduced from an erroneous calendar day for each station. The VQC was carried out during/after the digitisation process.
- **Procedure applied:**

1. Visual cross-checking of the digitised values compared with original sources every 10th, 20th and 30th day per month to ensure maximum precision on data digitisation (applied by the digitizers during the digitisation process).
2. Visual cross-checking choosing randomly 4 digitised whole months every 5 years to be compared with original sources for each station.
3. Checking the continuity of original sources to find potential errors (caused by no sheet, missing/repeated day, no data, no station,…).

The 95,7% of skewed values found in the UERRA rescued dataset were corrected and the 4,3% were set to missing

UERRA
Uncertainties in Ensembles of Regional ReAnalyses

2. Procedures II; Automated Quality Control (AQC)

- **Main purpose:** Detect and correct/remove non-systematic errors from the UERRA rescued dataset by using a battery of efficient QC tests specifically designed for the WP1 of the UERRA Project to check hourly observations of main climatic variables.
- Any individual or group of records that exceed a threshold previously established will be flagged as outliers or potential errors.
- Hourly variables tested:
  - TT=Surface air temperature
  - SLP=Sea Level Pressure
  - RH=Relative Humidity
  - WD=Wind Direction
  - WS=Wind Speed
  - DP=Dew Point temperature
  - FS=Fresh Snow (daily)
  - SD=Snow Depth (daily)
  - RR=Precipitation (daily)

UERRA
Uncertainties in Ensembles of Regional ReAnalyses

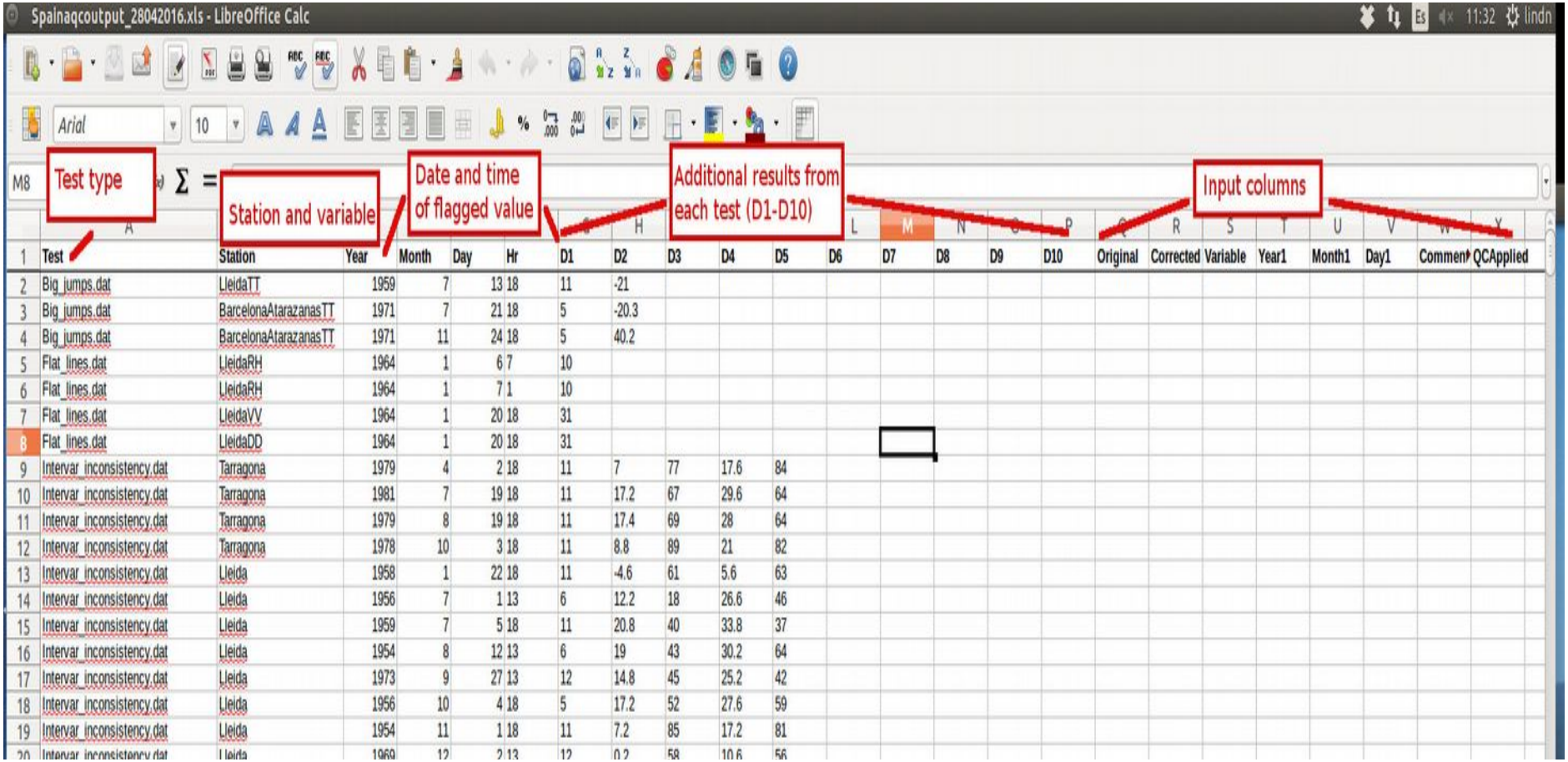
2. Procedures II; Automated Quality Control (AQC)

Main QC tests applied to the UERRA rescued dataset:

- Date order check: Correct order of calendar days.
- Unrealistic values: Values out of absolute extremes.
- Outliers: Values out of established thresholds.
- Big jumps: Large differences between adjacent values.
- Flat lines: Sequence of identical consecutive values.
- Repetitions of streaks: Explore repeated digitisation of entire months with erroneous dates.
- DP-inconsistency: Detect and flag DP>TT. Differences higher than 1ºC between the digitised DP and computed DP (from TT and RH) are also flagged.
- Fresh snow check: When FS>0, check RR>0 and TT<4ºC.

2. Procedures II; Automated Quality Control (AQC)

- AQC output file (manually check) by each country



Test	Station	Year	Month	Day	Hr	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Original	Corrected	Variable	Year1	Month1	Day1	Comment	QCApplied
Big_jumps.dat	LleidaTT	1959	7	13	18	11	-21																
Big_jumps.dat	BarcelonaAtarazanasTT	1971	7	21	18	5	-20.3																
Big_jumps.dat	BarcelonaAtarazanasTT	1971	11	24	18	5	40.2																
Flat_lines.dat	LleidaRH	1964	1	6	7	10																	
Flat_lines.dat	LleidaRH	1964	1	7	1	10																	
Flat_lines.dat	LleidaVV	1964	1	20	18	31																	
Flat_lines.dat	LleidaDD	1964	1	20	18	31																	
Intervar_inconsistency.dat	Tarragona	1979	4	2	18	11	7	77	17.6	84													
Intervar_inconsistency.dat	Tarragona	1981	7	19	18	11	17.2	67	29.6	64													
Intervar_inconsistency.dat	Tarragona	1979	8	19	18	11	17.4	69	28	64													
Intervar_inconsistency.dat	Tarragona	1978	10	3	18	11	8.8	89	21	82													
Intervar_inconsistency.dat	Lleida	1958	1	22	18	11	-4.6	61	5.6	63													
Intervar_inconsistency.dat	Lleida	1956	7	1	13	6	12.2	18	26.6	46													
Intervar_inconsistency.dat	Lleida	1959	7	5	18	11	20.8	40	33.8	37													
Intervar_inconsistency.dat	Lleida	1954	8	12	13	6	19	43	30.2	64													
Intervar_inconsistency.dat	Lleida	1973	9	27	13	12	14.8	45	25.2	42													
Intervar_inconsistency.dat	Lleida	1956	10	4	18	5	17.2	52	27.6	59													
Intervar_inconsistency.dat	Lleida	1954	11	1	18	11	7.2	85	17.2	81													
Intervar_inconsistency.dat	Lleida	1969	12	2	13	12	0.2	58	10.6	56													

UERRA
Uncertainties in Ensembles of Regional ReAnalyses

2. Procedures II; Automated Quality Control (AQC)

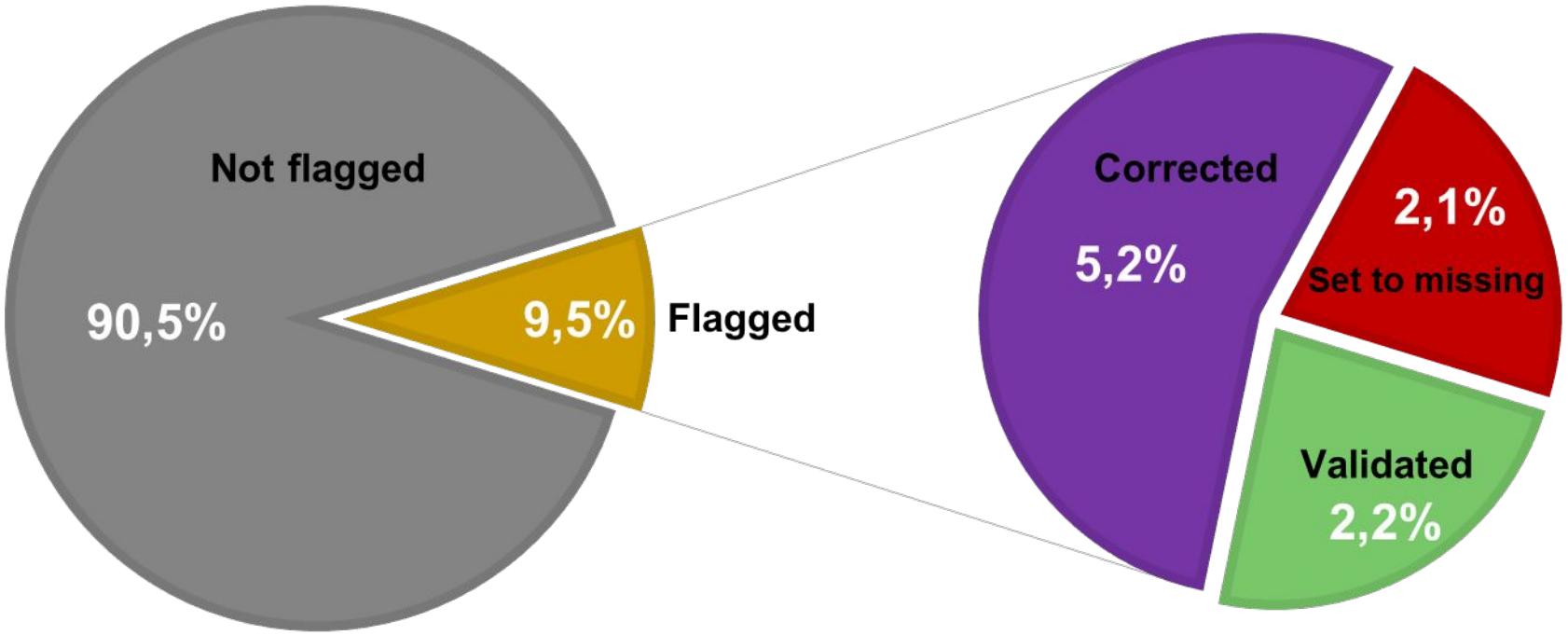
- All flagged observations have been manually cross-checked by using the original sources.

- Flagged values are not always errors. So, we should…
  - Validate them: the value is correct.
  - Correct them: the value is not correct and we are sure about the correct value.
  - Reject them: the value is not correct and we cannot offer another value.

- Taking decisions about flagged values:
  - **Code 1:** Removed due to digitisation error (e.g. incorrect station digitised, incorrect variable,…)
  - **Code 2:** Corrected digitisation error (e.g. typing error)
  - **Code 3:** Removed after expert consideration
  - **Code 4:** Validated after expert consideration (e.g. a false positive)
  - **Code 5:** Corrected after expert consideration (e.g. a clear typographical error in the data source)
  - **Code 6:** Not manually checked, assumed correct
  - **Code 7:** Corrected to missing (e.g. there is no value in the data source)

3. Results I:

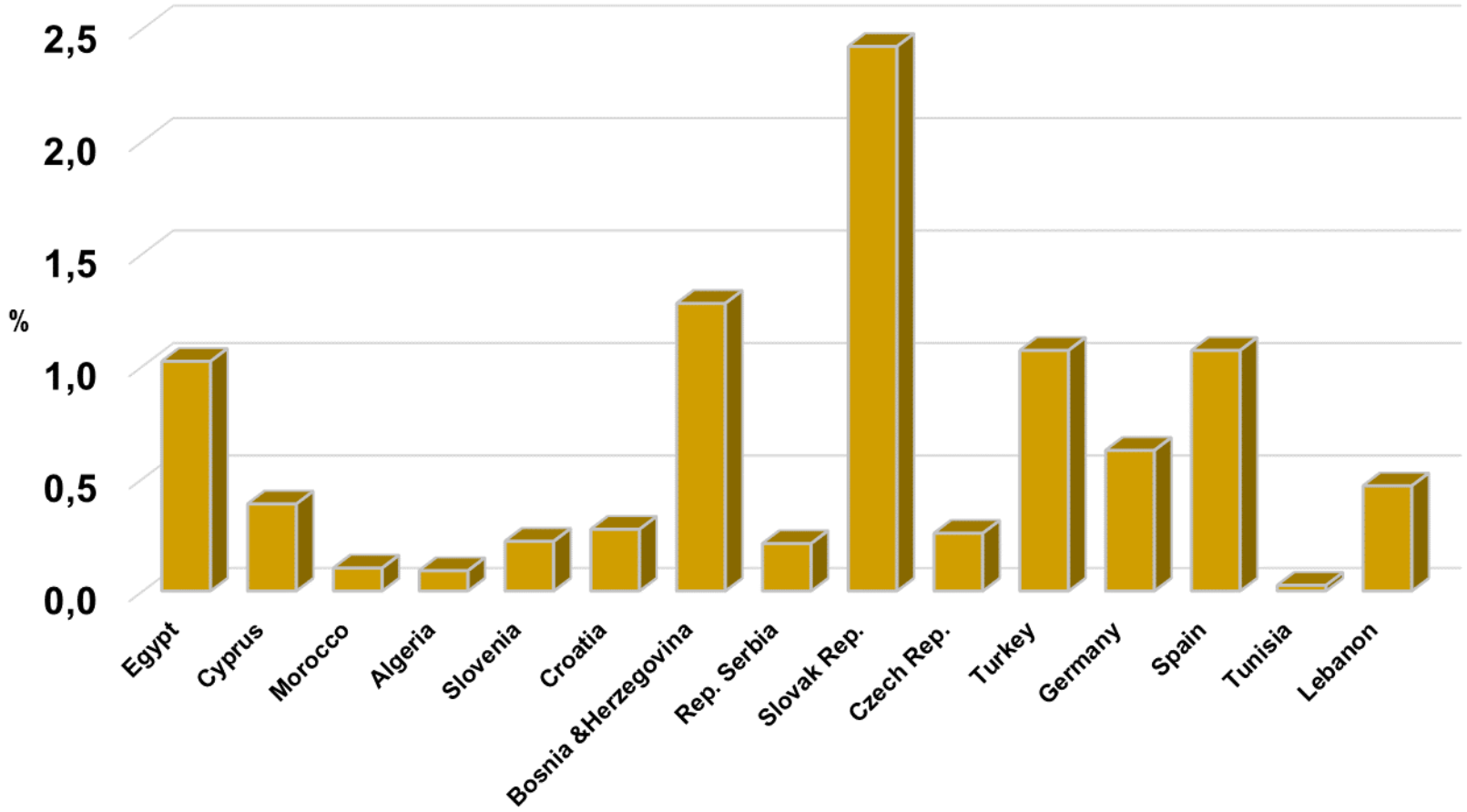
- **Main statistics derived from AQC results for the UERRA rescued dataset**



UERRA
Uncertainties in Ensembles of Regional ReAnalyses

3. Results II:

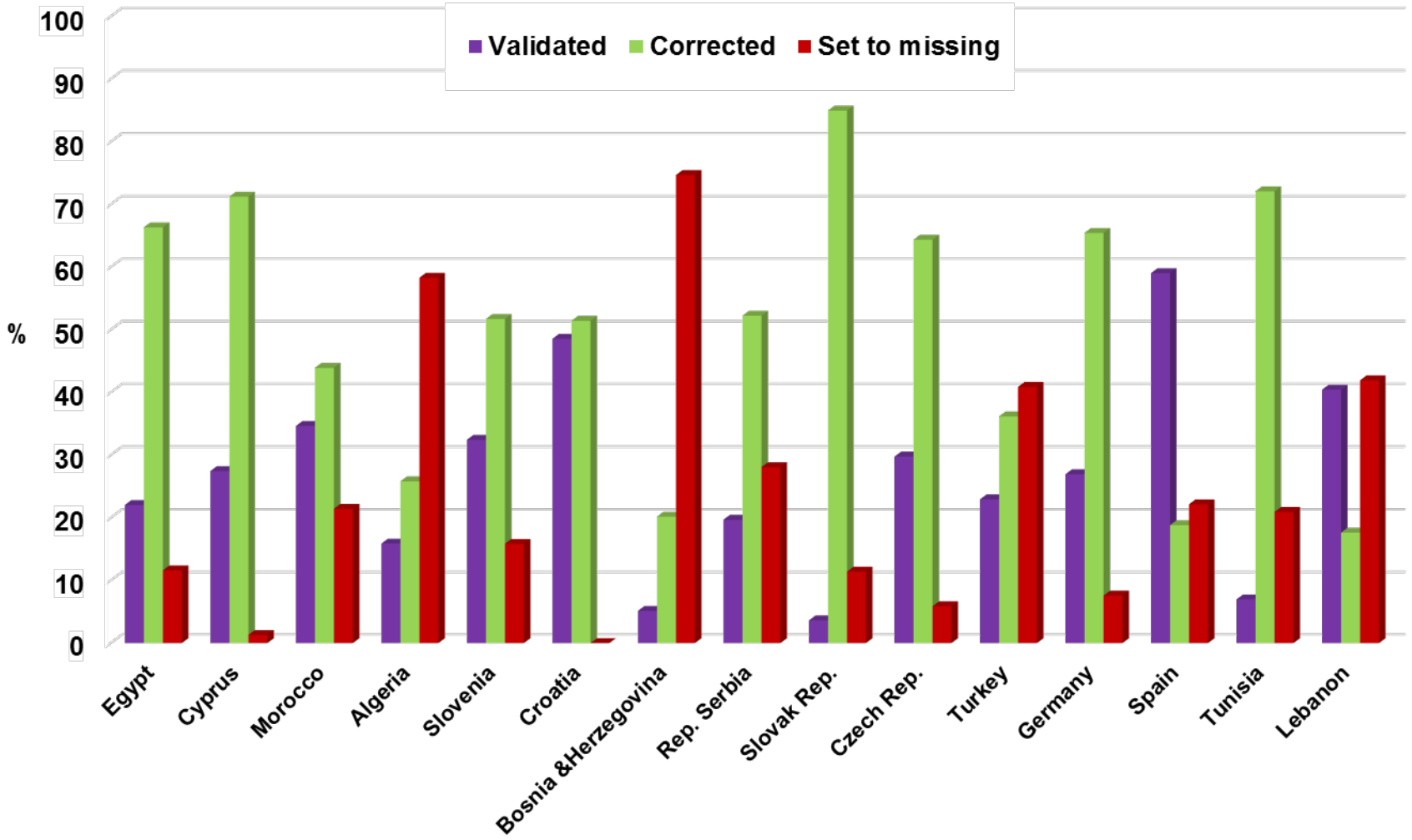
- **Percentage of flagged values by countries (Total=9,5%)**



UERRA
Uncertainties in Ensembles of Regional ReAnalyses

3. Results III:

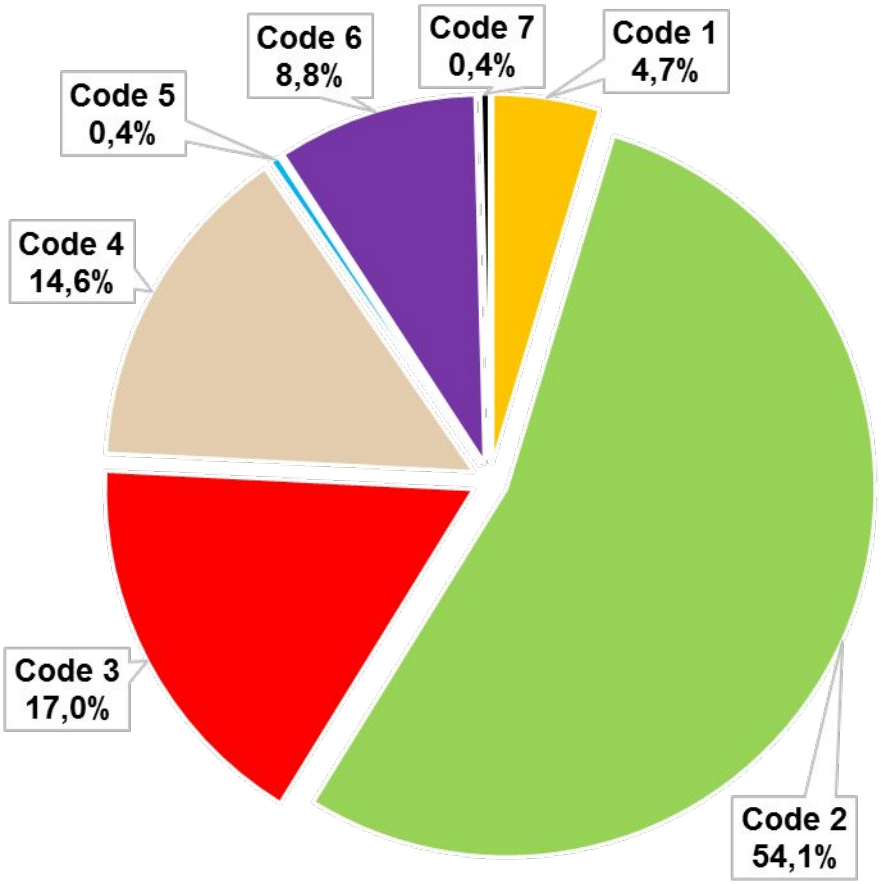
- **Percentage of validated, corrected and set to missing values by countries**



UERRA
Uncertainties in Ensembles of Regional ReAnalyses

3. Results IV:

- **Taken decisions about flagged values by using specific codes**



Code 1 Removed digitisation error
Code 2 Corrected digitisation error
Code 3 Removed after expert consideration
Code 4 Validated after expert consideration
Code 5 Corrected after expert consideration
Code 6 Assumed correct
Code 7 Corrected to missing

UERRA
Uncertainties in Ensembles of Regional ReAnalyses

4. 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 automated quality control tests will be applied to the dataset by using the methodology developed by Dunn et al., 2012. Intra- and inter-station consistency will be checked at sub-daily scale for TT, SLP, DP, WD and WS. Suspect values will be flagged, but not automatically deleted.

QC tests to be 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 medians 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.

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

UERRA Uncertainties in Ensembles of Regional ReAnalyses

Thank you!!

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