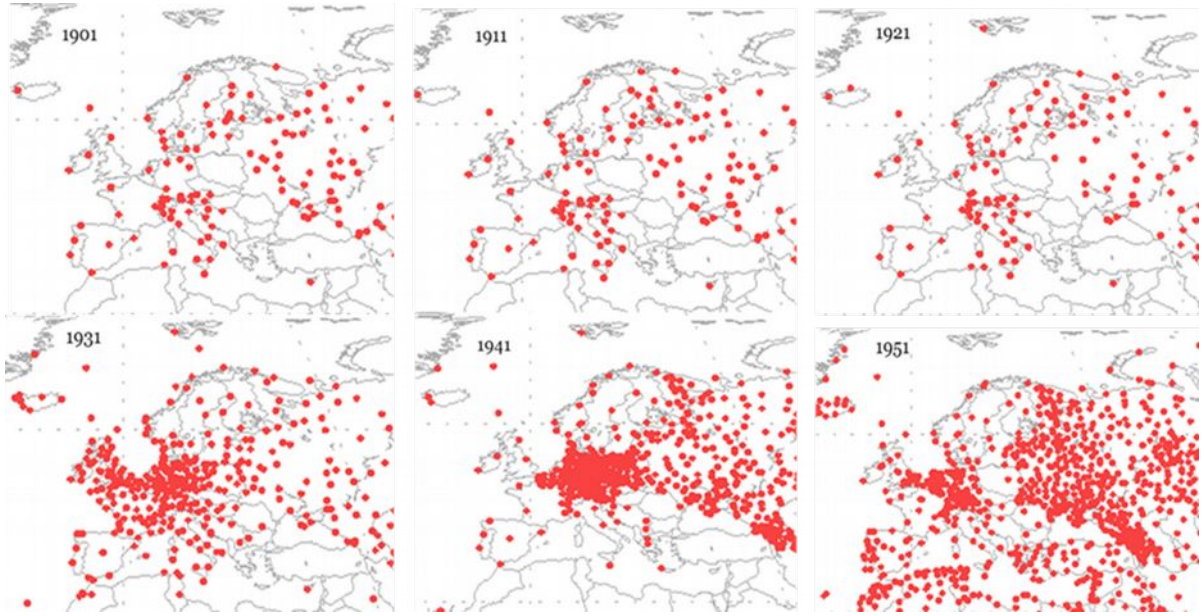
UERRA DATA RESCUE EFFORTS TO ENHANCE AVAILABILITY OF QUALITY CONTROLLED SYNOPTIC OBSERVATIONS TO SUPPORT REGIONAL REANALYSIS IN EUROPE

By Manola Brunet

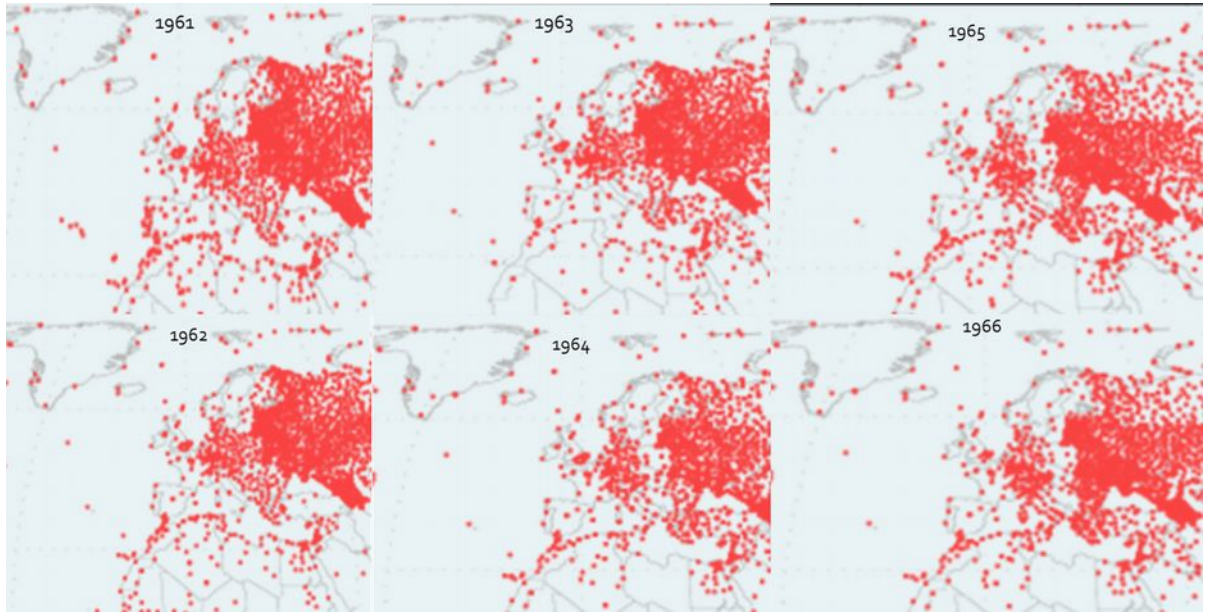
Centre for Climate Change (C3), University Rovira i Virgili, Tarragona, Catalonia

THE NEED FOR ENHANCING THE LAND-SURFACE INPUT DATA TO SUPPORT CLIMATE REANALYSIS (BOTH GLOBAL AND REGIONAL)

A look to surface climate data availability for pre-1951 in the ISPD: the basic land-surface input data for the MARS Archive

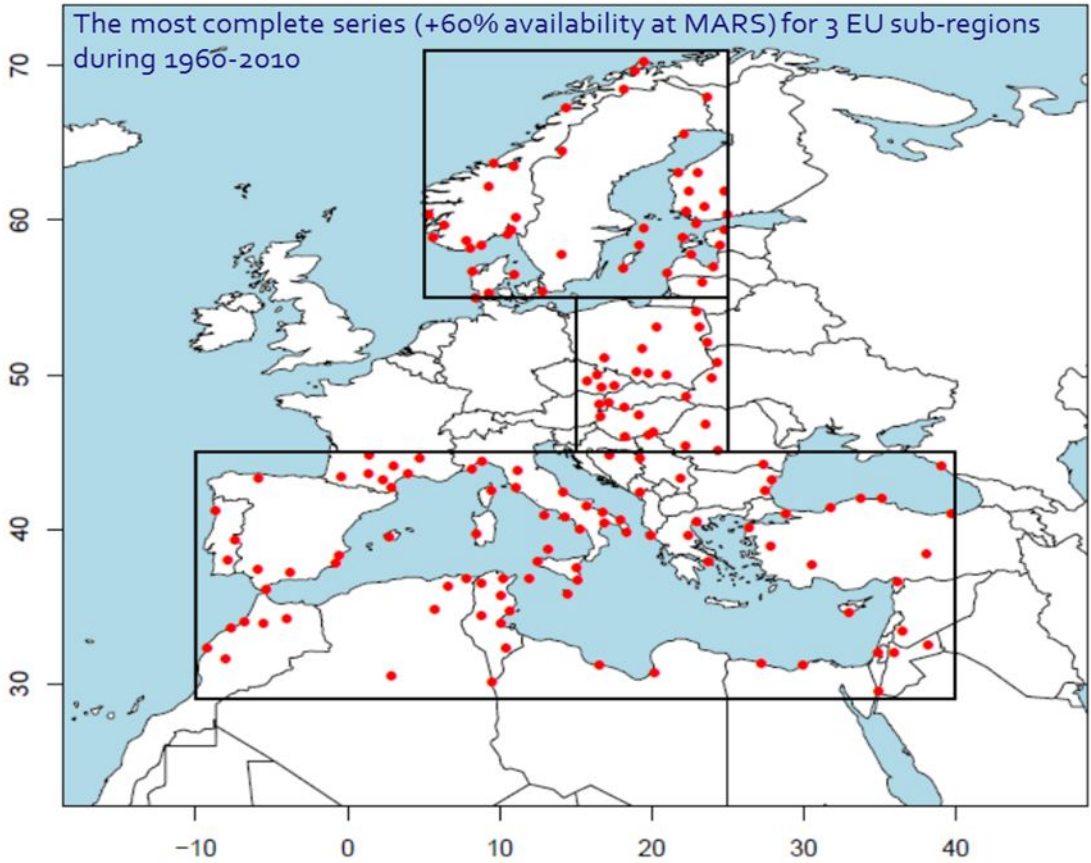


And for the poorest data availability in the 1960s over several sub-regions (e.g. Scandinavia, Western and Eastern Europe)

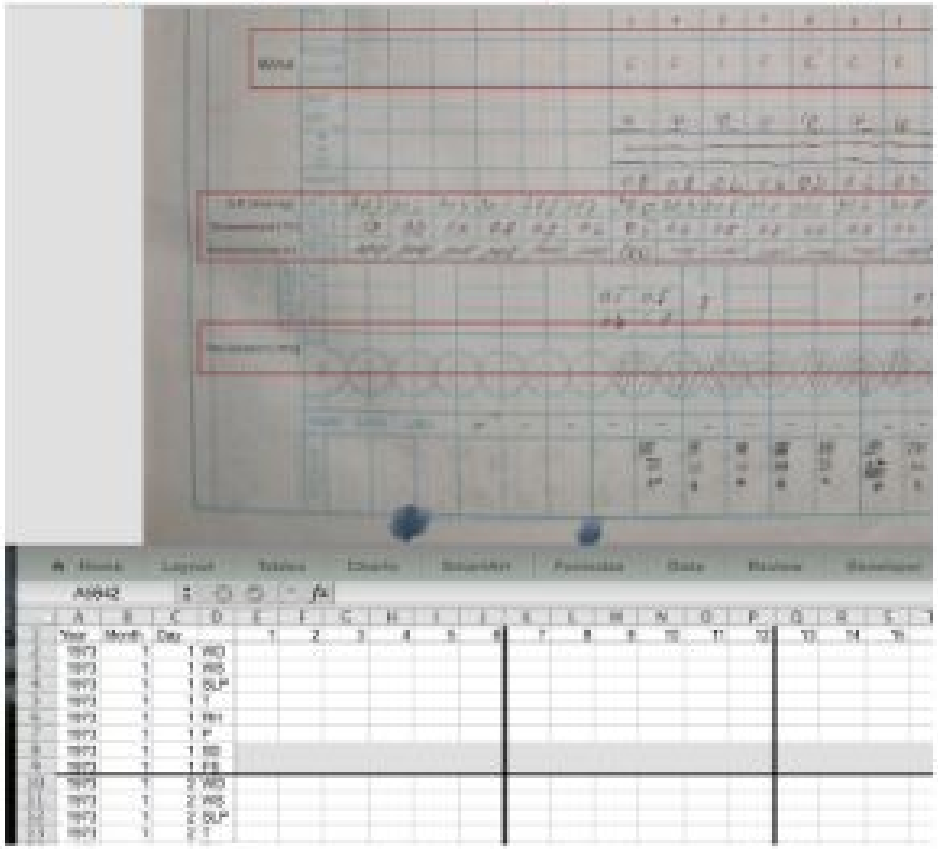


THE EFFORT UNDERTAKEN IN UERRA: THE DATA RESCUE (DARE) APPROACH

- Setting the targets (assessing spatial & temporal gaps):
  - Using the EURO4M gathered data sources to identify new ones & produce a comprehensive list of historical climate data holders and sources with relevant un-digitised, but imaged, data over Europe. Large variety of data-source formats
  - Exploring MARS Archive & crosschecking against the gathered data-sources to set the spatial-temporal gaps in climatic variables at the synoptic scale
  - Involving UERRA reanalysis people to identify climatic variables and locations that can have a higher impact to enhance RRA in Europe
  - 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)
- The targets for digitisation:
  - Mainly spatial infilling over data-sparse sub-regions: Southern part of Med, Eastern Europe & Balkans
  - Variables: SLP, TMP, WS, WD, RH, DP, SD, FS, RR at the hourly and daily scales
  - For post-1950 (the target with the focus on 60s & 70s), but also pre-1950 (data sources availability constrain)



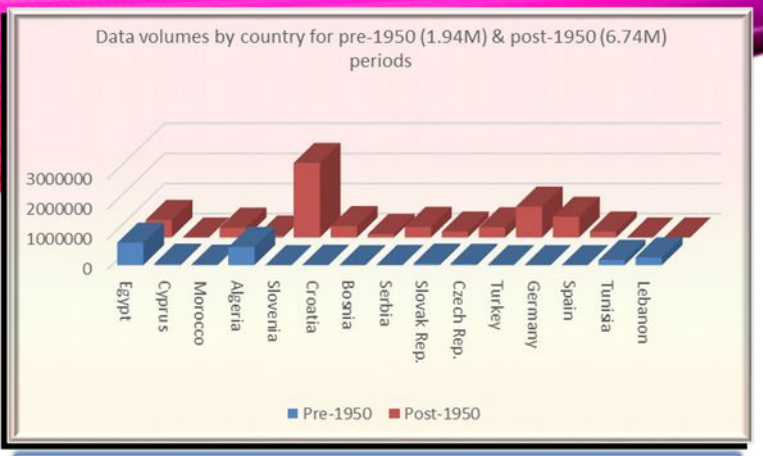
- Setting the digitisation plan
  - Contracting 11 motivated science students digitising 15 hours a week during 2 years
  - Training digitisers & elaborating tutorials for a more efficient digitisation
  - Producing spreadsheet templates that match the data source to reduce digitisation errors (each data source requires a different template)
- Ensuring the quality of the digitisation work: 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 following 3/4/5 rule: 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. zeros missing in pressure readings: 121.9 instead of 1021.9; Columns and rows miss-aligned; Missing data recorded as blank or zero

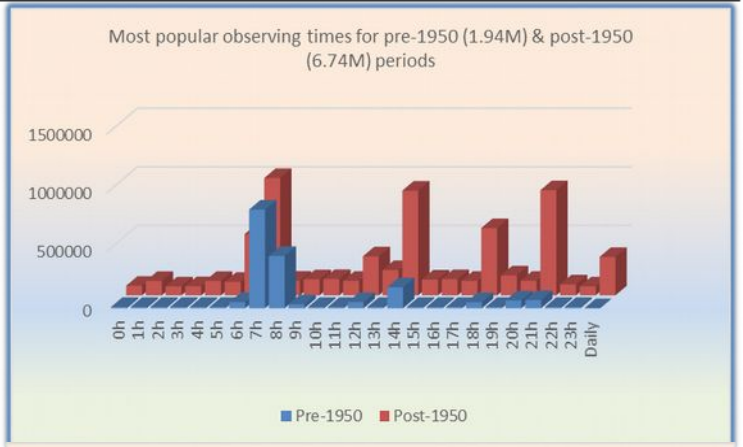


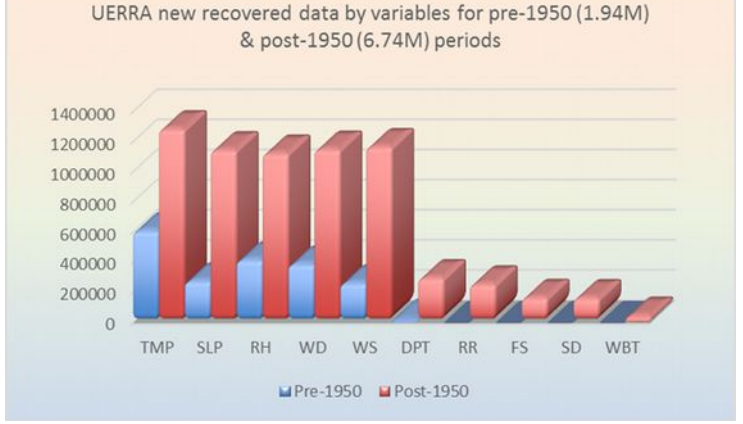
THE UERRA/DARE DOUBLE STRATEGY TO ENHANCE REANALYSIS INPUT DATA

- **Digitisation effort:**
  - Observations (at the synoptic and daily scales) from the gathered & available data sources were used for digitisation, increasing in about **5.6M of new obs.**
  - The constraint of data sources availability for the post-1950 period brought us to the need of involving EU NMSs to get access to their scanned assets: Proposal of an exchange exercise (e.g. scanned data provision and return of digitised and quality controlled data) sent to 13 relevant EU NMSs
  - **Catalonia, Germany and Slovenia** *NMSs* provided access to some of their scanned, but un-digitised, data sources, increasing in about **3.4M of new obs.** the UERRA recovery effort. About **9M of new obs.** digitised in UERRA to fill in spatial gaps
  - Others couldn't provide access because either internal data policies or difficulties to have scanned - duplicated their recent obs.
- **Gathering digitised data from NMSs with open data policies**
  - Accessing digitised synoptic data was another unforeseen effort kin the UERRA proposal, but successful and with meaningful results
  - **Catalonia, Norway & Sweden NMSs** agreed & provided us about **178.1M** of their digitised synoptic data

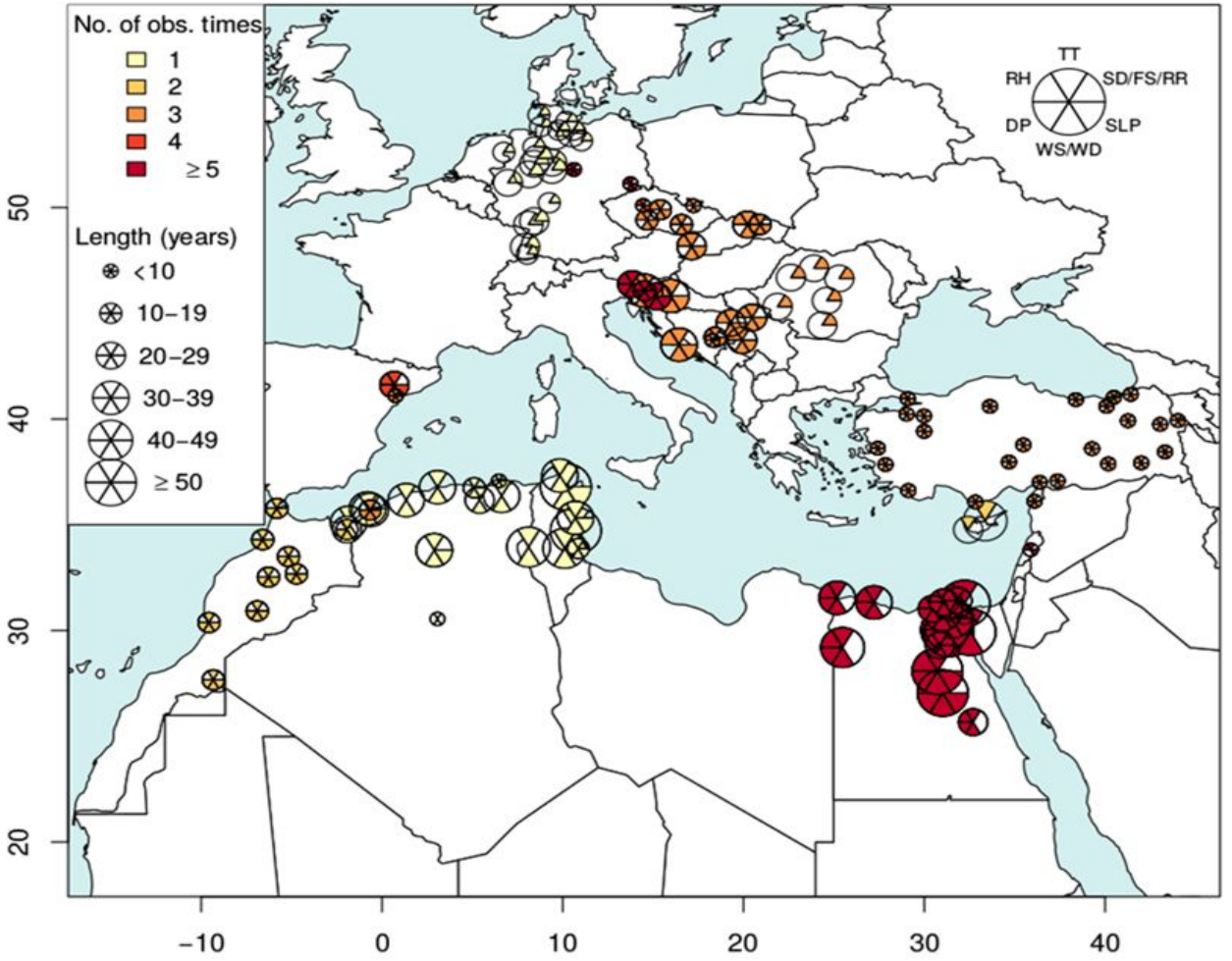
SUMMARISING UERRA'S DIGITISATION EFFORTS (~9M OF NEW OSERVATIONS)







- **Spatial effort**: Balkans (Slovenia, B&H, Serbia), Turkey, Germany, Egypt & Morocco for post-1950 & Southern Med (Egypt, Algeria, Tunisia) for pre-1950
- **Most popular observing times**: morning (7am), afternoon (2pm), evening (9pm & 6pm) for post-1950 & 7am & 8m, 2pm, 9pm and 8pm for pre-1950
- **Most popular climatic variables**: TMO, SLP RH, WD/WS for both periods



SUMMARISING THE UERRA GATHERING EFFORT: BRINGING 178.1M OF DIGITISED & PUBLICLY AVAILABLE 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

Infilling in Scandinavian gaps:

Norway: 7.2M of TMP, WD/WS, RR obs. from 93 stations for 1960-1980

Sweden: 42M of TMP, SLP, RR, RH, SD, CC obs. from 146 stations for 1945-2009

Catalonia: ~129M of TMP, SLP, WD/WS, RR, RH, SD from 76 stations for 1988-2015

And total figures

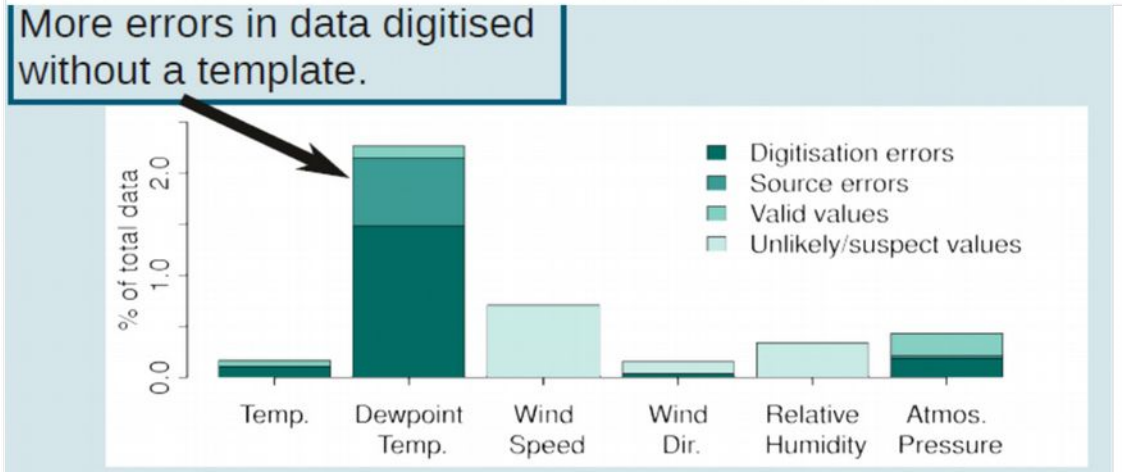
DATA DEVELOPMENT: A TWO-PRONGED QUALITY CONTROL (QC) APPROACH

Our dual approach: Visual Quality Control (**VQC**) & Automatic Quality Control (**AQC**)

- Systematic visual crosschecking have been conducted looking for common mistakes
- Common errors include:
  - Consistent misreading of confusing values (e.g. a 3 as 8)
  - Skipped dates, leading to out of order data
  - Incorrect rows or columns digitised (these last two problems are minimised with the use of templates)
  - ***These errors are hard to detect using automatic methods***

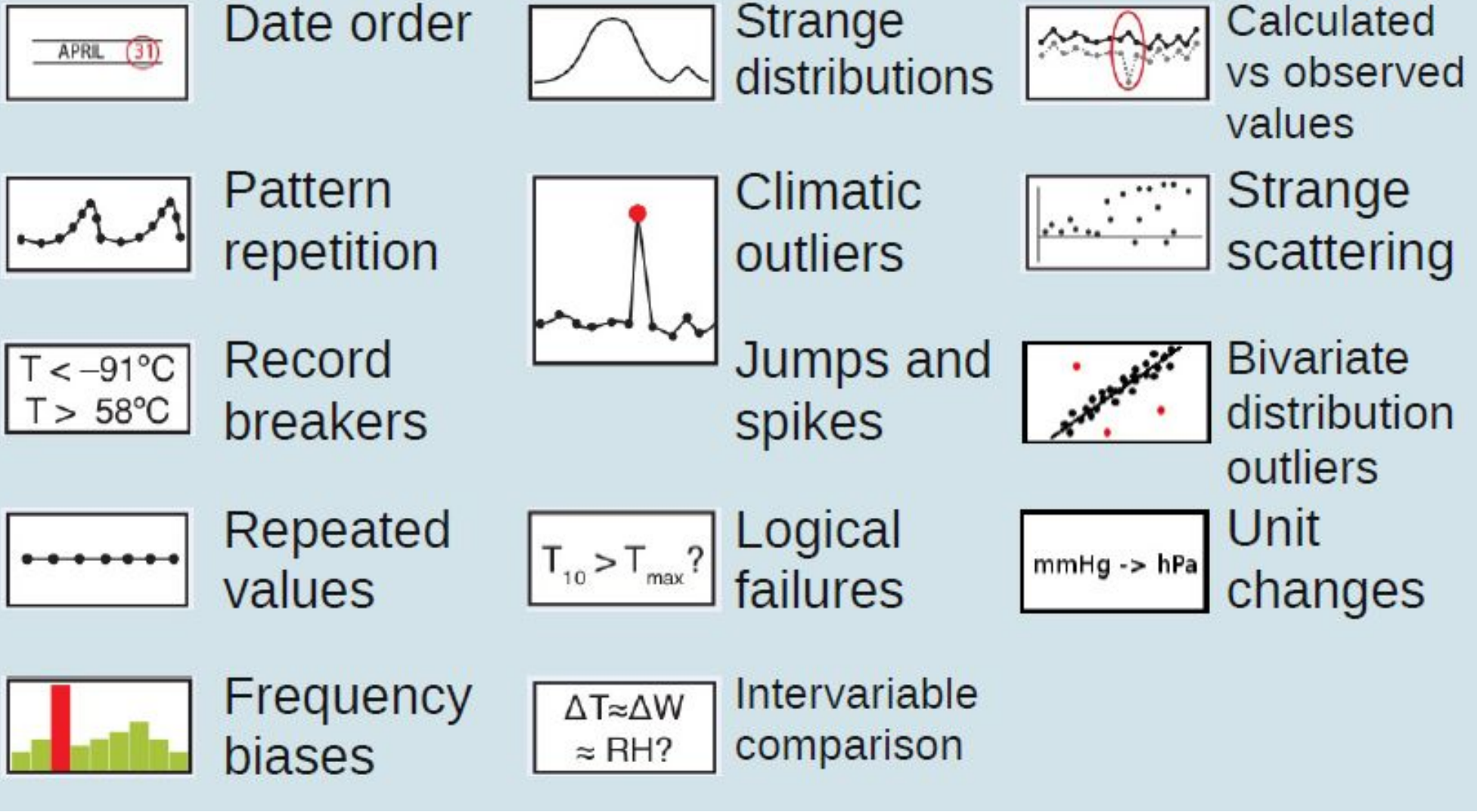
For example, digitized data from Tarragona, Spain:

- 07:00h, 13:00h, 18:00h for 1977–1984, 6 variables = 52,020 station-values
- Less than 3% of observations were flagged as errors
- 62% of flagged observations were corrected as typos or removed as clear source errors



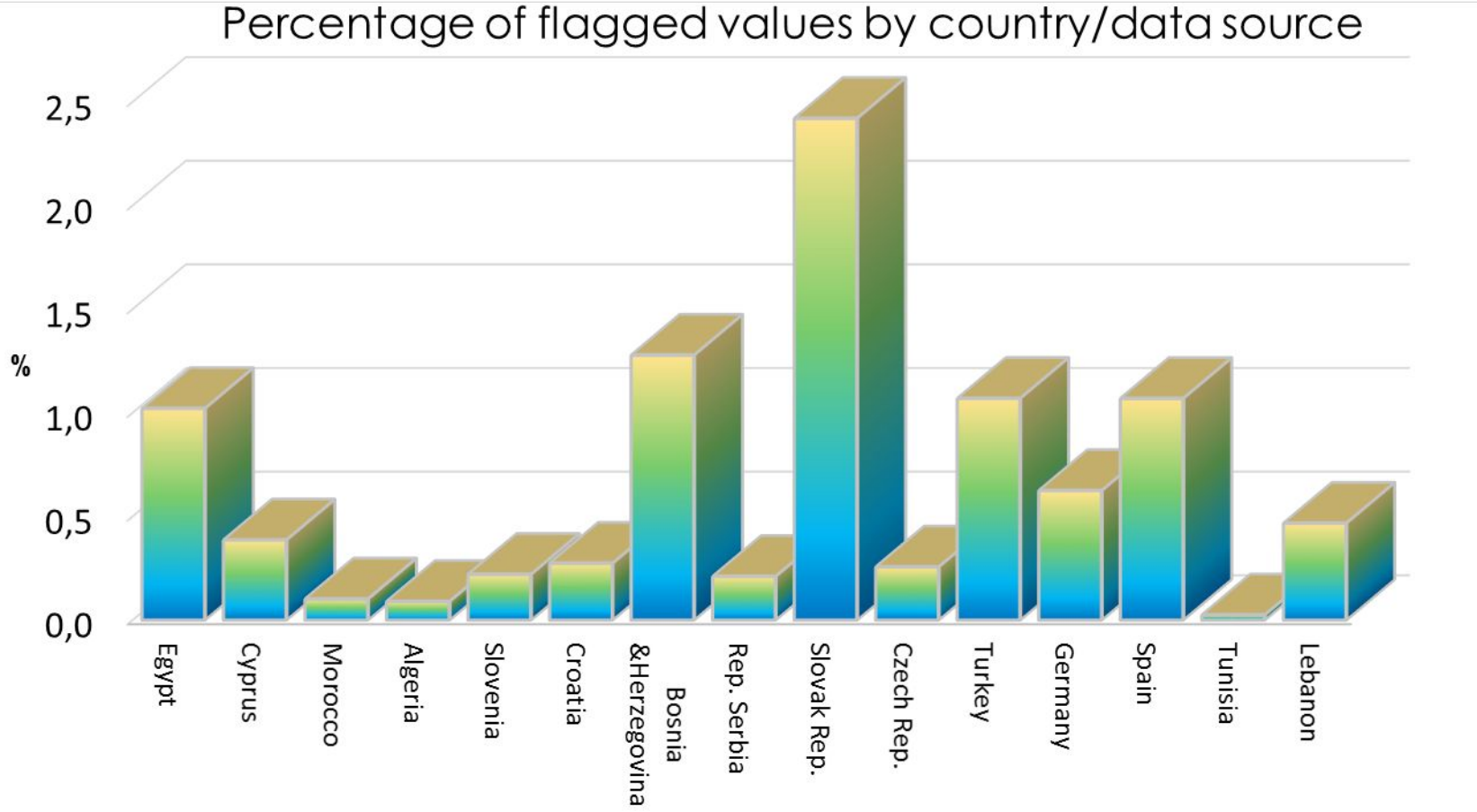
Percentage of total Tarragona sub-daily data flagged by the automatic quality control procedure

THE AUTOMATIC QUALITY CONTROL (AQC) **TESTS** AND RESULTS



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

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

DATA ACCESSIBILITY

- All the observations recovered (either digitised or gathered in digital format) will be provided, first, to UERRA partners to be in use in the last year of the project and later (early 2017) will be made publicly available through:
  - European Climate Assessment & Dataset (ECA&D) & MARS Archive
  - ZENODO repository
  - The International Surface Pressure Databank (ISPD)
  - The Met Office Hadley Centre observations datasets: the HadISDH dataset
  - The International Surface Temperature Initiative (ISTI) databank
  - And given to the relevant NMSs to encourage them in data sharing

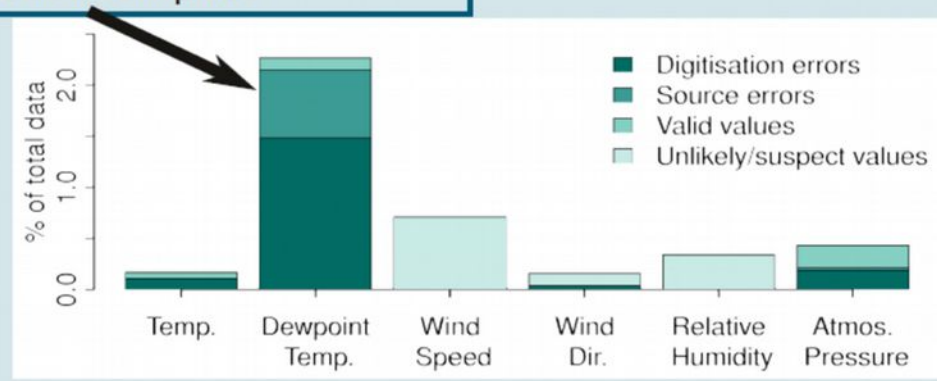
Thanks 4 your attention

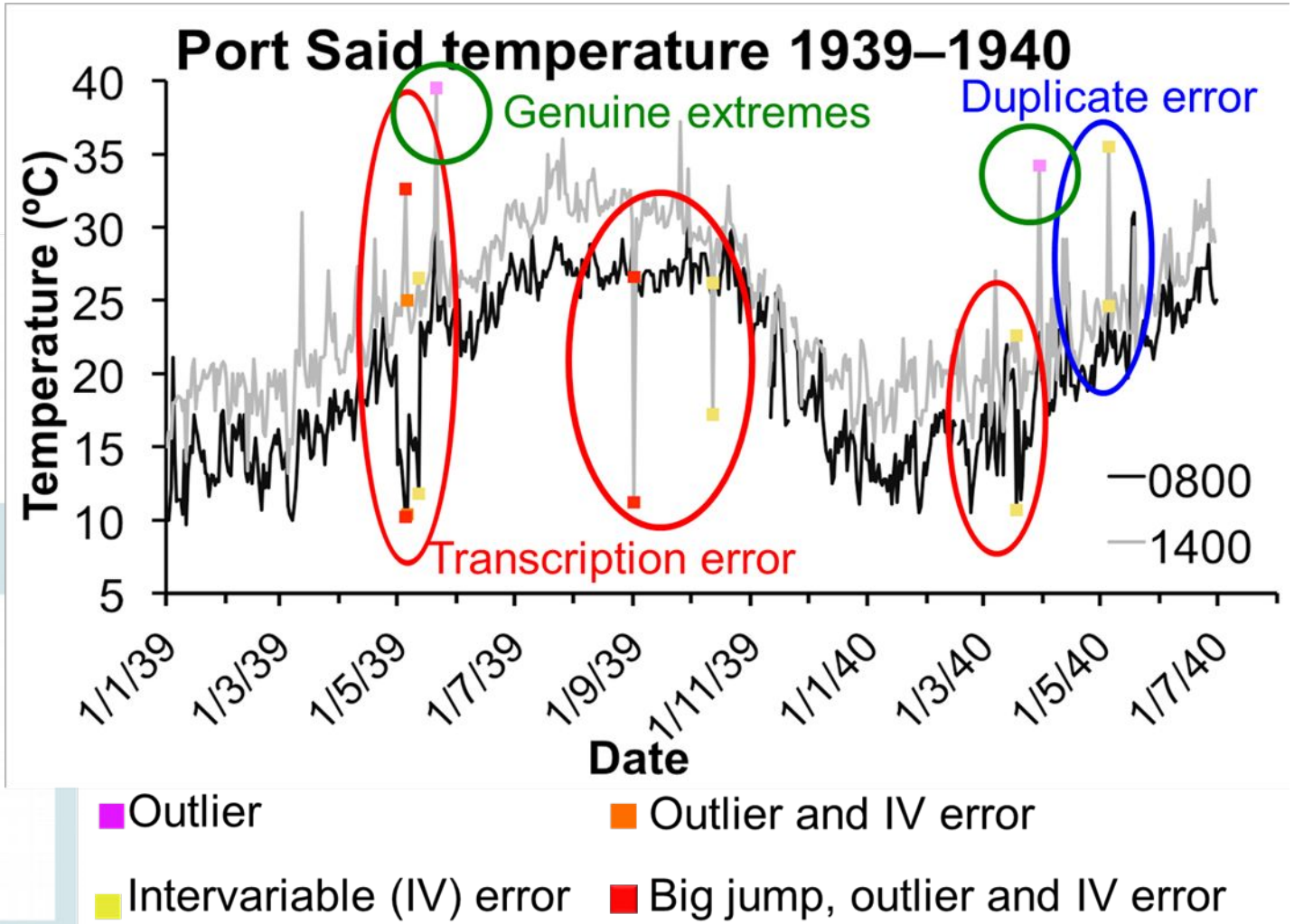
DATA DEVELOPMENT FROM UERRA DIGITISATION EFFORT: TWO-PRONGE QUALITY CONTROL APPROACH

For example, digitized data from Tarragona, Spain:

- 07:00h, 13:00h, 18:00h for 1977–1984, 6 variables = 52,020 station-values
- Less than 3% of observations were flagged as errors
- 62% of flagged observations were corrected as typos or removed as clear source errors

More errors in data digitised without a template.





THE AUTOMATIC QUALITY CONTROL (AQC) TESTS AND **RESULTS**