Current practice, tools, and users for observation feedback at ECMWF

Paul Poli

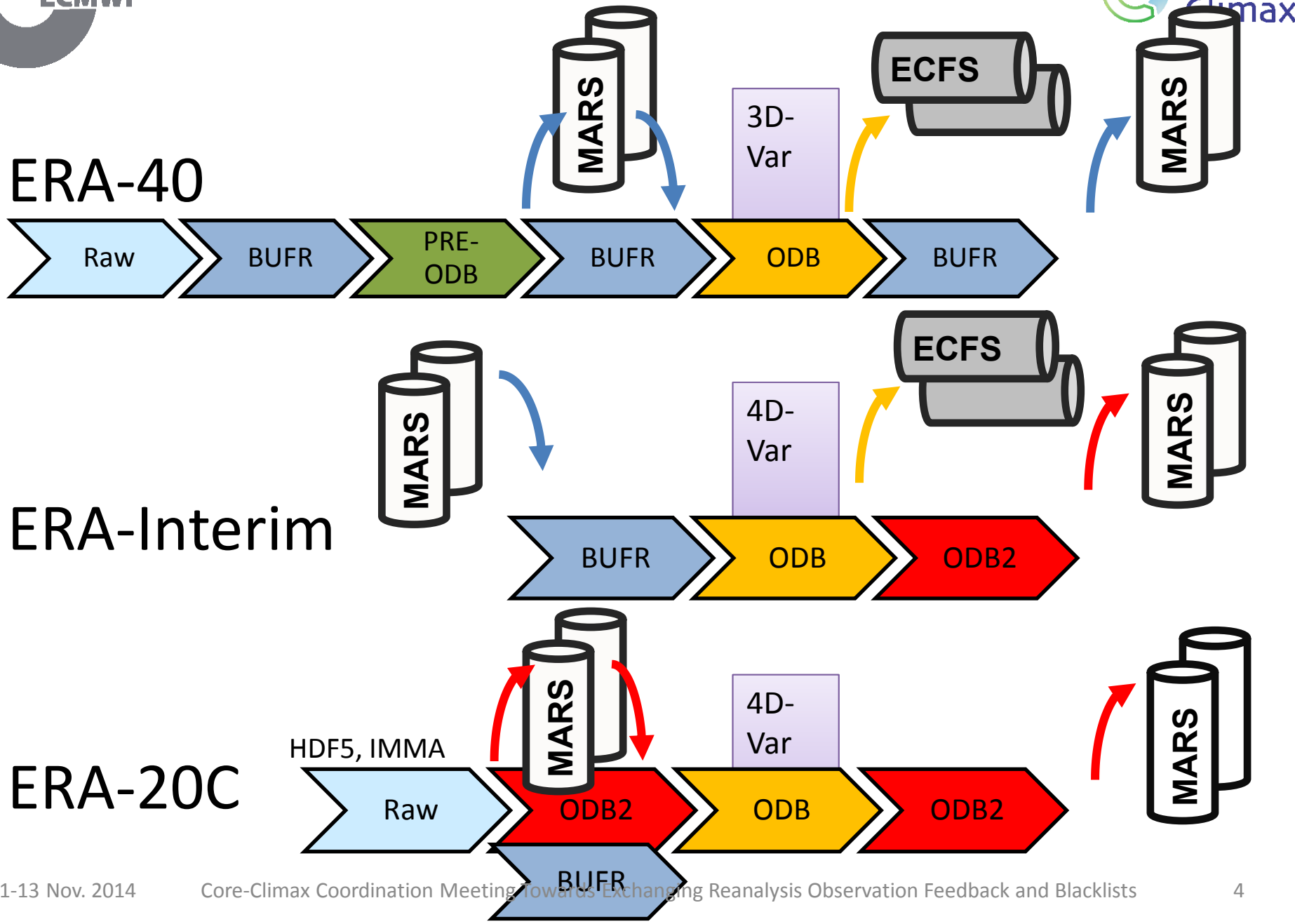
Which of our reanalyses have observation feedback?

- ERA-15
- ERA-40
- ERA-Interim
- ERA-20C

Observation sources

- "ERA-40 observation input" includes GTS and many other datasets acquired specifically, listed in ERA-40 paper
- GTS since 2002
- Reprocessed datasets:
  - Significant wave-heights from ERS-1,-2 altimeters
  - AMV winds from Meteosats
  - Ozone profiles from ERS-2 GOME
  - Bending angles from CHAMP GPS radio occultation
- ICOADS v2.5.1
  - All parameters observed from ships and buoys
- ISPD v3.2.6
  - Surface and mean-sea-level pressures

Observation flow in ECMWF reanalyses



BUFR

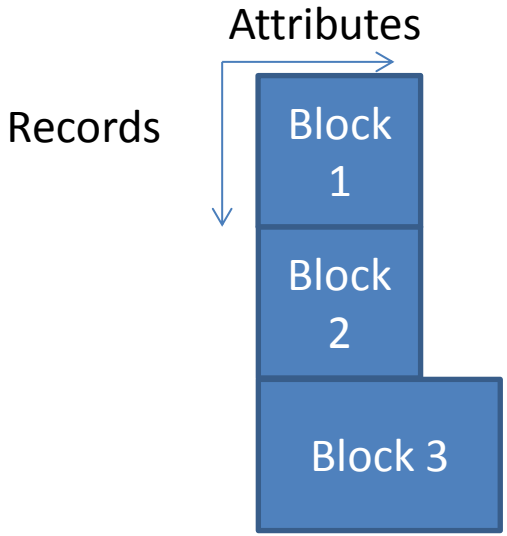
- Series of messages – can be aggregated (`cat`), regardless of their contents
- Each message is arranged in an order that requires detailed documentation to decode
- A near-real-time (NRT) data exchange format:
  - Tries to save as much space as possible, when bandwith is the most costly
- Keeps the data the way they were originally sent by the data provider, and received* (most NRT transmissions being in BUFR)
  - In fact there is a BUFR processing step at reception

ODB1

- Hierarchical database
- Several ODB1s cannot be merged easily (over-arching headers)
- Data organized by broad type
  - CONV, AMSUA
  - Each type shares the same attributes
    - For example, all AMSUA data specify a “scan position”, all GPSRO data specify the “local radius of curvature of the Earth”
  - Within each type, database tables allow to avoid “data normalization” in database language (avoids information duplication)
  - Another layer (pool) allows for parallel processing – important for exploitation on high-performance computers
- Format ideal for large data assimilation problems
  - It allows to read/write **only the part of the data** that each processor choses to work on

ODB2

- Flat format – series of blocks of records
- Blocks can be aggregated together (UNIX ‘cat’ command)
- Within each block, the list of attributes is invariant
  - Regardless of whether some attributes are constant
  - No notion of hierarchy
  - Enormous duplication
  - Each block compresses information therein using a CODEC



- Format good for archiving (can handle observation data streaming, similar to video, even in large flows) and for comparing – since all attributes are common, expanded to the lowest common denominator

Tools, by format

- BUFR: for linux and HPC
  - ‘dump’ and (linux:) ‘viewer’
  - spaghetti Fortran programs that need tailoring for each type of BUFR message
- ODB1: for linux and HPC (preferred)
  - Fortran interface (quite heavy to implement)
  - SQL command-line (dump, search, sort, aggregate functions…)
  - Linux: viewer in metview/Magics++
- ODB2: for linux and HPC
  - API with Fortran, C++, and python bindings
  - SQL command-line (dump, search, sort, aggregate functions…)
  - Several wrappers at command-line for sub-setting, merging (adding blocks or attributes), comparing (attributes and numerical contents)
  - Linux: viewer in metview/Magics++

Observation feedback data policy

- ERA-40 observation feedback: BUFR format makes it difficult/impossible for users to access anyway – no need to worry too much
- ERA-Interim observation feedback:
  - Origin of data: some ERA-40 data (some of which were purchased and cannot be redistributed) and GTS
  - For GTS: WMO resolution 40 Adopts (3): “Members should provide to the research and education communities, for their non-commercial activities, free and unrestricted access to all data and products exchanged under the auspices of WMO with the understanding that their commercial activities are subject to the same conditions identified in Adopts (2) above;”
  - Aiming to get the same license as the other ERA products (meteorological fields); the *feedback* part is what makes this an ERA product, not the observations alone (these are incomplete anyway – people interested solely in these should go back to the original data provider for complete metadata etc…)
- ERA-20C and after:
  - All observation data used were acquired under the explicit premises that they can be redistributed for free for research, alongside the feedback

Observation feedback isn't complete

- In all current reanalyses: **data from the following separate analyses are not found** in our obs. feedback
  - land-surface analysis of T2m, RH2m, snow cover, soil moisture observations
  - ocean wave analysis of significant wave heights observations
  - In ERA5 these feedbacks will be saved
- **Data lost at the pre-processing step**:
  - At data reception
    (e.g., several issues with snow cover data in recent years)
  - Thinning
    (e.g., removing 1 out of N satellite pixels)
  - Super-obing
    (e.g., averaging observations)

Observation feedback contents

- Assimilation experiment attributes
- Database book-keeping attributes
- Observation identification, metadata, and data
- Assimilation feedback attributes (departures before/after assimilation, usage flags,…)

Example 1: Radiosonde temperature

- **Assimilation experiment attributes**: expver@desc andate@desc antime@desc creadate@desc creatime@desc creaby@desc moddate@desc modtime@desc modby@desc mxup_traj@desc
  - ' 1758' 20130101 000000 20130223 120957 ' eras' 20130223 134338 ' eras' 2
- **Database book-keeping attributes**: ddrs.offset@desc ddrs.len@desc timeslot_index.offset@desc timeslot_index.len@desc latlon_rad@desc ddrno@ddrs wordno@ddrs bulkdata@ddrs target@index hdr.offset@index hdr.len@index seqno@hdr body.offset@hdr body.len@hdr errstat.offset@hdr errstat.len@hdr entryno@body
  - 0 6144 0 25 0 2 3047 538976288 25 2047 1 5461155 65459 504 65459 504 65
- **Observation identification, metadata, and data**: obstype@hdr codetype@hdr insttype@hdr retrtype@hdr areatype@hdr zone@hdr date@hdr time@hdr sortbox@hdr sitedep@hdr source@hdr statid@hdr ident@hdr sonde_type@hdr lat@hdr lon@hdr country_code@hdr stalt@hdr modoro@hdr instspec@hdr sensor@hdr rdbflag@hdr subtype@hdr bufrtype@hdr satinst@hdr satname_1@hdr satname_2@hdr satname_3@hdr satname_4@hdr varno@body vertco_type@body rdbflag@body
  - 5 35 NULL 0 1 0 20130101 000000 0 0 ' ' ' 71924' 71924 80 74.7 -94.97 0 46 40.33777234841 0 0 0 101 2 0 'TEMP Lan' 'd Report' ' ' ' ' ’ 2 1 0
- **Assimilation feedback attributes**: procid@index tslot@index status@hdr event1@hdr event2@hdr anflag@body status@body event1@body event2@body press@body press_rl@body obsvalue@body biascorr@body biasctrl@body tcwv_fg@body blacklist@body ppcode@body an_depar@body fg_depar@body an_sens_obs@body surfemiss@body final_obs_error@errstat obs_error@errstat repres_error@errstat pers_error@errstat fg_error@errstat
  - 133 19 1 0 2 0 1 67108864 1 83900 NULL 249.89 -0.0099987792223556 0 0 0 0 0.014472140730075 0.50996667233544 0 0 0.86813322147508 0.86813322147508 -0.0099987792223556 NULL 1.3246080918025

Example 2: MSU brightness temperature

- **Assimilation experiment attributes**: expver@desc andate@desc antime@desc creadate@desc creatime@desc creaby@desc moddate@desc modtime@desc modby@desc mxup_traj@desc
  - ' 1510' 19830101 000000 20110101 223441 ' eras' 20110101 231523 ' eras' 2
- **Database book-keeping attributes**: ddrs.offset@desc ddrs.len@desc timeslot_index.offset@desc timeslot_index.len@desc latlon_rad@desc ddrno@ddrs wordno@ddrs bulkdata@ddrs target@index hdr.offset@index hdr.len@index seqno@hdr body.offset@hdr body.len@hdr errstat.offset@hdr errstat.len@hdr entryno@body
  - 0 6144 0 25 0 1 1 3072 24 0 1 534 0 4 0 4 1
- **Observation identification, metadata, and data**: obstype@hdr codetype@hdr insttype@hdr retrtype@hdr areatype@hdr zone@hdr date@hdr time@hdr sortbox@hdr sitedep@hdr source@hdr statid@hdr ident@hdr sonde_type@hdr lat@hdr lon@hdr country_code@hdr stalt@hdr modoro@hdr instspec@hdr sensor@hdr rdbflag@hdr subtype@hdr bufrtype@hdr satinst@hdr satname_1@hdr satname_2@hdr satname_3@hdr satname_4@hdr varno@body vertco_type@body rdbflag@body scanline@atovs fov@atovs zenith@atovs
  - 7 210 NULL 0 2 0 19821231 150000 0 0 ' ' ' ' 706' 706 0 -73.5625 93.93069 0 841000 3509.897012 0 1 0 54 3 623 'NOAA ' ' 6 706' ' SENSOR=' 'MSU ' 119 3 0 127 5 10.731356366
- **Assimilation feedback attributes**: procid@index tslot@index status@hdr event1@hdr event2@hdr anflag@body status@body event1@body event2@body press@body press_rl@body obsvalue@body biascorr@body biasctrl@body tcwv_fg@body blacklist@body ppcode@body an_depar@body fg_depar@body an_sens_obs@body surfemiss@body final_obs_error@errstat obs_error@errstat repres_error@errstat pers_error@errstat fg_error@errstat cldcover@atovs cldptop_1@atovs cldptop_2@atovs cldptop_3@atovs cldne_1@atovs cldne_2@atovs cldne_3@atovs
  - 97 1 1 0 0 196608 44 0 0 1 NULL 213.72 1.0327265021583 0 0 6 0 5.8517725440586 5.9099425328733 0 0.80000122085215 10 10 NULL 0.80000122085215 3 0 0 0 0 0 0 0

Observation identification: example of traceability in ERA-20C

- **Source**: ISPD 3.2.6, ICOADS 2.5.1
- **Collection identifier**: each source assembled data records from various origins. For example, collection number 761 in ICOADS 2.5.1 contains
  "Japanese Whaling Ship Data (CDMP/MIT digitization)"
- **Station identifier**: as defined by the source, not necessarily WMO ID.
- **Unique identifier**: each source has its own way of defining this variable. It is usually a number which needs to be used along with something else (observation date and time for example)
- **Report type**: Radiosonde launched from land, Island surface station, etc…
- **Timeseries_index**: added by us, allows later to pull out easily a single station time-series, across the entire observation feedback archive
- **Geophysical variable number**: actual observable (temperature, surface pressure…)

Observation identification tables at ECMWF

- Reportype:
  - Examples: **16002** indicates Manual Land SYNOP, **1005** indicates NOAA 19 AMSUA Radiances, …
  - http://data-portal.ecmwf.int/odbgov/ReportType/

- Geophysical variable numbers:
  - Examples: **7** indicates Specific humidity, **119** indicates Brightness temperature, …
  - http://data-portal.ecmwf.int/odbgov/Varno/

Identification of in situ observations

- By WMO station identifier
  - Does not work for old stations
  - As anything else, they are sometimes incorrect (some identifiers re-used across the network, e.g., “SHIP”)
- By observation type
  - Not so clear-cut as it seems; this only refers mostly to the way the report was transmitted: some Ps data were exchanged as SYNOP, but also as METAR, and even possibly within a RS message
- By geolocation to discriminate between duplicated identifiers
- By date/time
- By pressure or vertical level for upper-air data
- By geophysical variable

Identification of satellite observations

- By **satellite identifier** *241 DMSP F-8*
  - Ideally WMO Publication No. 306 “Manual on codes” Common code table C-5 “Satellite identifier”
  - Issues for old satellites, some missing (e.g. GMS, GMS-2)
- By **instrument** *905 SSM/I*
  - Ideally WMO Publication No. 306 “Manual on codes” Common code table C-8 “Satellite instruments”
  - Issues for old instruments, some missing (e.g. HIRS before HIRS/2)
- By **sensing epoch**
  - E.g., orbit or revolution number, scan line, scan position *603, 200, 5*
- **Vertical positioning**: *3*
  - For sounders: by channel
    - Which channel number convention to use?
  - For AMVs: by pressure level of the retrieval
  - For GPSRO: by impact parameter (or reduced impact parameter)
- **Spatio-temporal position** (date/time/lat/lon) *19870716 03:05:01 20.50 N 58.7 E*
  - Bearing in mind that this may be change slightly with reprocessing

Feedback added-value

- Obs – f.g. (after bias correction)
- Obs – analysis (after bias correction)
- Bias correction estimate (not always)
- Usage flags, superposition of bits: meaning:
  - 0/1 datum active (was submitted to 1st minimization)
  - 0/2 datum passive (had no weight in the assimilation)
  - 0/4 datum rejected (was kicked out of the assimilation, either right before assimilation, if not active, or after the 1st minimization, if active)
  - 0/8 datum blacklisted (blacklist said so!)
- Assumed observation error (standard deviation)

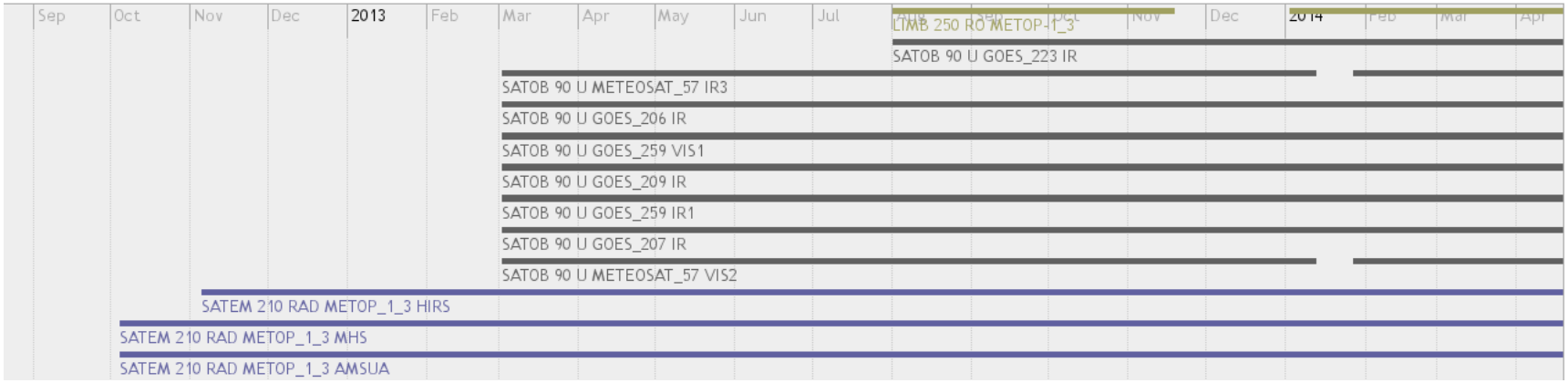
Observation feedback summaries

- Radiosondes: a CSV file (specifying used/non used) http://old.ecmwf.int/research/era/do/get/index/29/28
- Observation timelines: http://intra-dav.ecmwf.int/assets/js/timeline/timeline_tools.html
- Statistical summaries: count, mean, stdev., rms of departures before/after assimilation: (PostgreSQL) observation statistics database http://ibstone.ecmwf.int:8080/main2?dbname=db2&dbuser=erc&dbhost=ibstone&dbport=5432&dbschema=sch_stats5.&dbengine=e1
- Monitoring plots in many shapes and forms, e.g. http://old.ecmwf.int/products/forecasts/d/charts/monitoring/satellite/era/

Observation timelines

(modified) SIMILE TIMELINE widget, driven by JSON files.

Can be searched and diff'd.



Data volumes

- ERA-40 (28Tb for 45 years of fields)
  - 2Tb in BUFR
- ERA-Interim (78Tb for 35 years of fields)
  - 23Tb in ODB1
  - 66Tb in ODB2
  - Only 4% of this volume is conventional data

Observation Feedback Archive

- A product of ERA-CLIM
- Offers an **open-access web interface**
- Major improvement as it enables users to 'see' the observation systems and data locations, without transferring large data amounts across the network

- Still limited by export format (ASCII and ODB2)
- Publicly available ODB2 tools limited, but an ODB2 decoder is in the works
- Needs manual intervention to create the catalogue (not self-made)

- Ongoing developments:
  - Native support for **NetCDF** in MARS (ERA-CLIM2)
  - Searchable metadatabase of events (CHARMe)