The ROMEX Core Data Set – Statistics, Reprocessing, and Lessons Learned

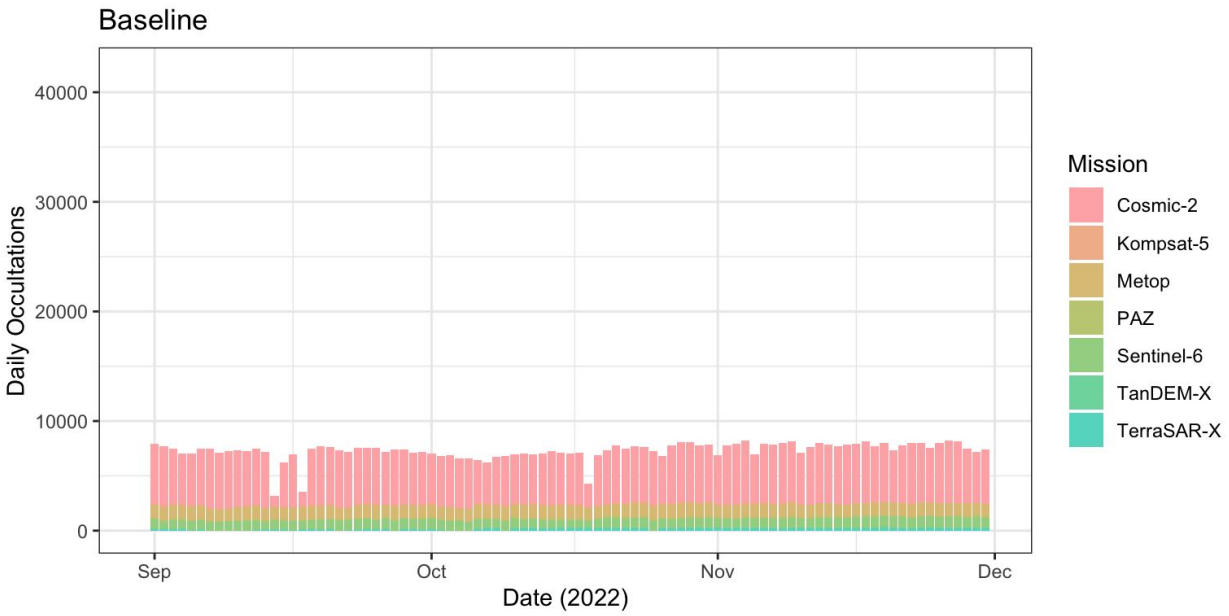
C. Marquardt, A. von Engeln, F. Martin Alemany, N. Morew, R. Notarpietro, S. Paolella, S. Padovan, V. Rivas Boscán, F. Sancho

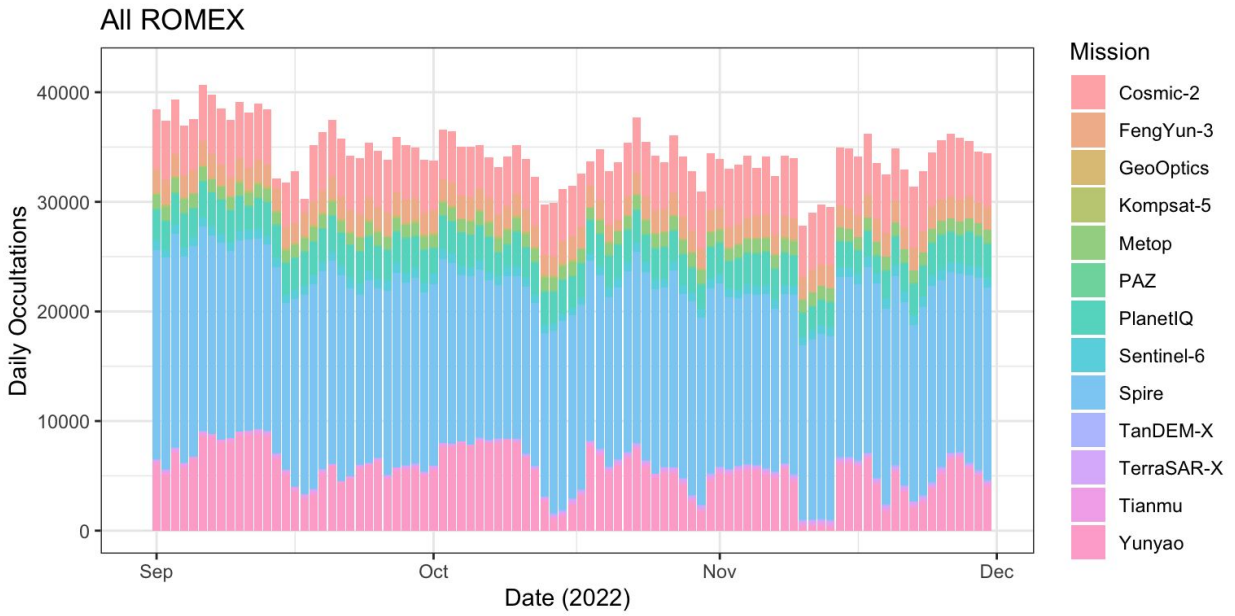
JCSDA Workshop / IROWG-10, Boulder

13. September 20

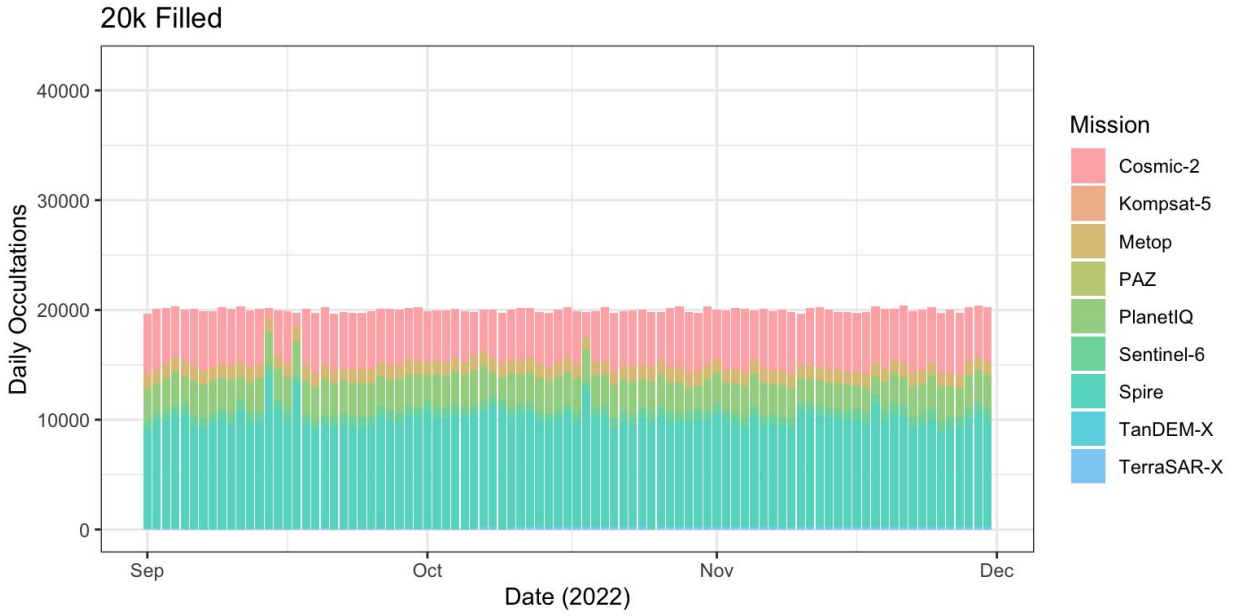
- Numbers and distributions
- Data quality
- Data Status
- Summary

ROMEX Data Sets and Numbers





- **Baseline**: currently available data, minus commercial and Chinese data
- **Core**: All ROMEX data available
- **20k**: Intermediate data set composed of Baseline, PlanetIQ and Spire data only



Core Data Set

- The initial idea was to take operational data sets as provided and pack them for ROMEX.
- That worked well for:
  - UCAR-processed data sets: COSMIC-2, Kompsat-5, PAZ, TerraSAR-X, TanDEM-X, GeoOptics
  - EUMETSAT-processed data sets: GRAS, Sentinel-6 (reprocessed already)

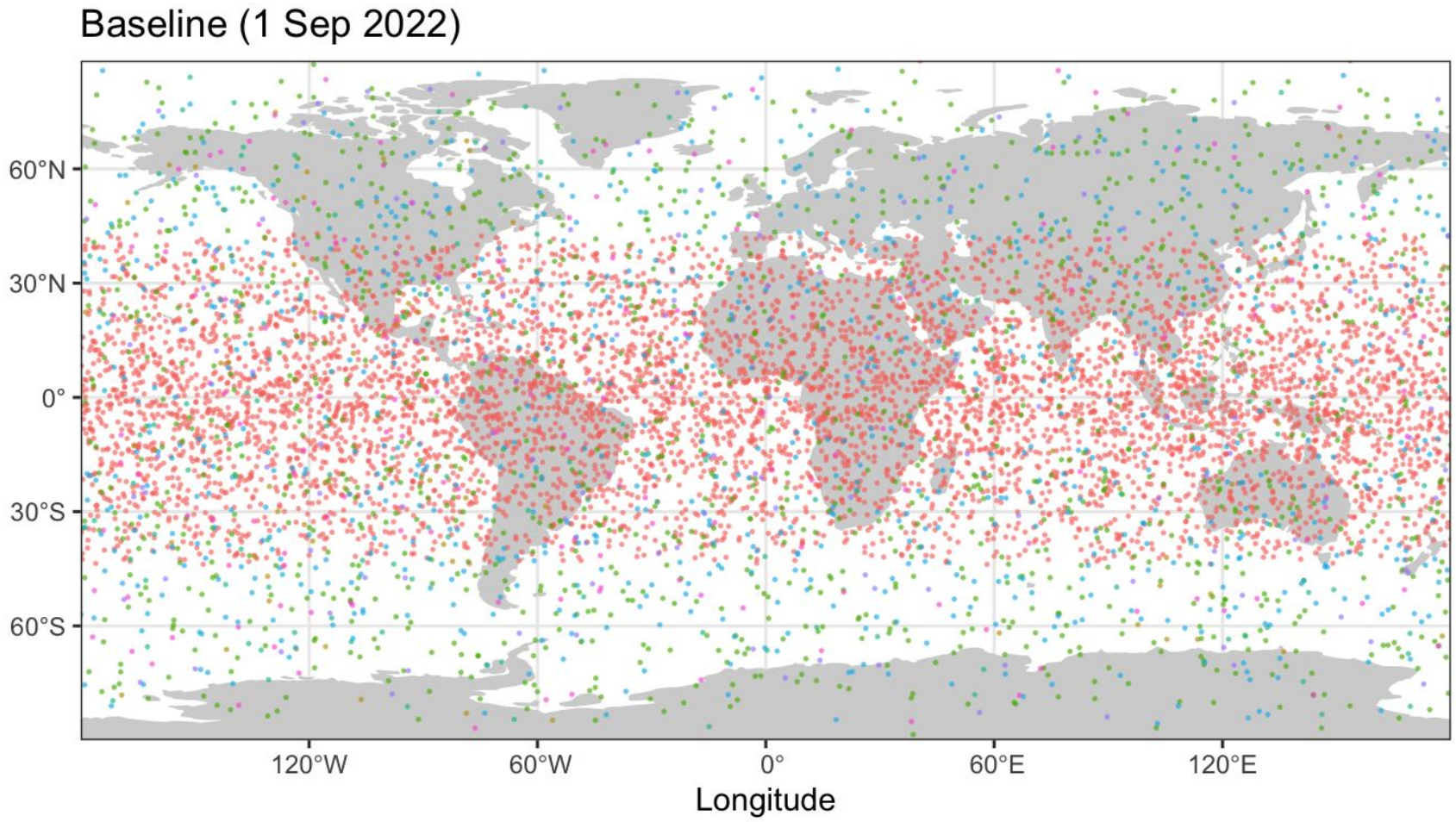
- For other data sets, more work was required:
  - CMA, Tianmu: Only high-resolution bending angle data: converted to BUFR by EUMETSAT
  - PlanetIQ: inconsistent number of level1a (excess phase) and level1b/2 (bending angle) data in high res and bufr: reprocessed by EUMETSAT from excess phases (PlanetIQ agreed)
  - Spire: Operational data from three sources: COSMIC, EUMETSAT, and Spire for those profiles not sold to either NOAA or EUMETSAT: Reprocessed by EUMETSAT from excess phases (Spire agreed)
  - Yunyao: Issues in data quality: reprocessed by EUMETSAT from excess phases (Yunyao agreed)

 In retrospect, we should have spent more time on CMA (FY-3) and also Tianmu data to improve data quality.

Numbers after processing centre QC

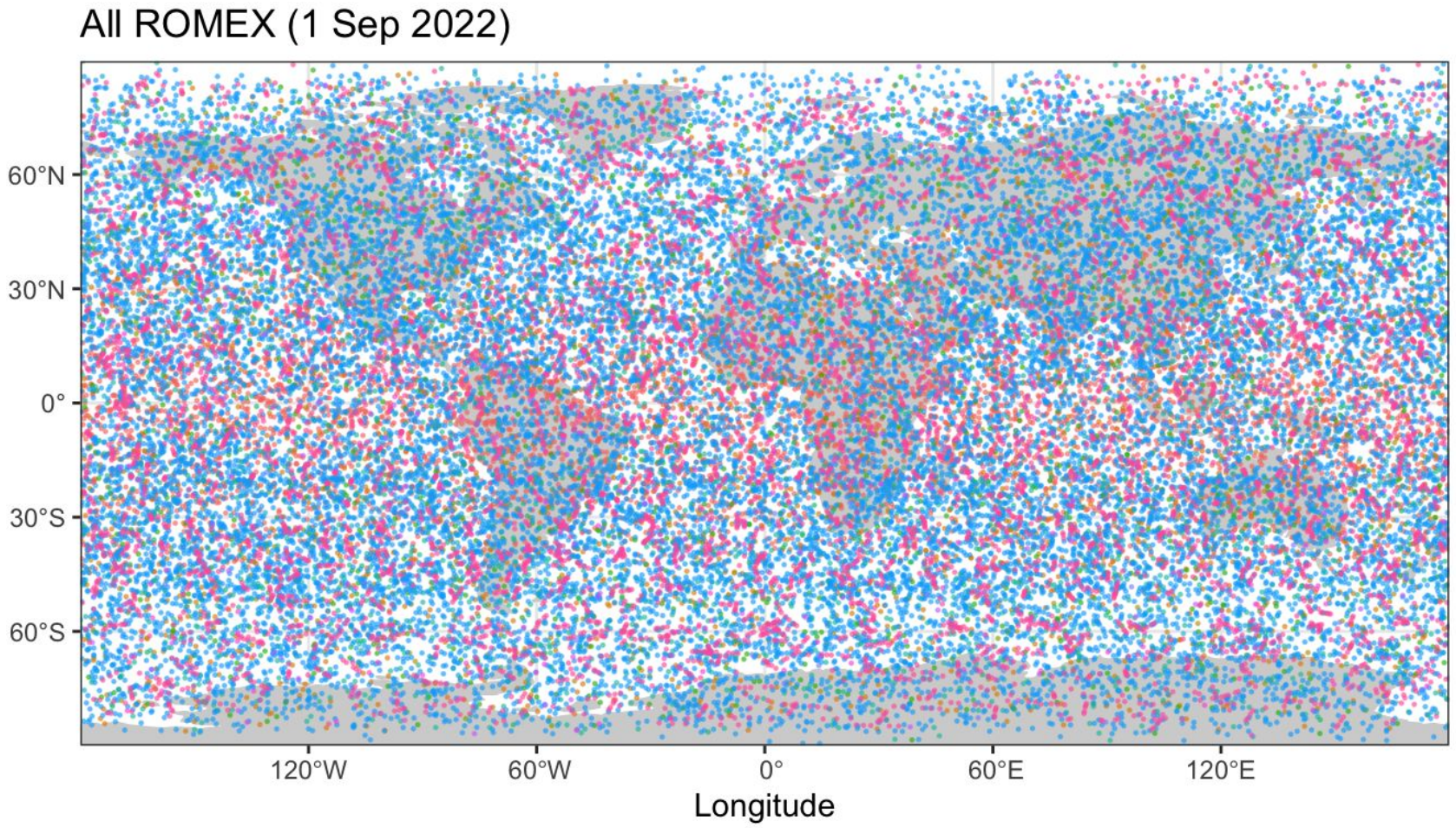
Mission	Occs/day	Baseline	Notes
Spire	16.746		Processed by EUM (excess phase by Spire)
Yunyao	5.370		Processed by EUM (excess phase by Yunyao)
COSMIC-2	4.883	Yes	Processed by UCAR (nrt)
PlanetIQ	2.767		Processed by EUM (excess phase by PlanetIQ)
Fengyun-3	1.960		BUFR-encoded by EUM (atmPrf by NSSC/CMA)
Metop	1.146	Yes	Processed by EUM (Operational)
Sentinel-6	849	Yes	Processed by EUM (NTC)
Tianmu	272		BUFR-encoded by EUM (atmPrf by NSSC/Tianmu)
Other	484	Yes (w/o GeoOptics)	Two or more weeks missing for some missions (TanDEM-X, TerraSAR-X, Kompsat-5; GeoOptics ends during ROMEX-1); processed by UCAR (nrt or postProc)
Total	**34.478**	7.295	Actual daily number between 28.000 and 40.000

ROMEX missions – Baseline



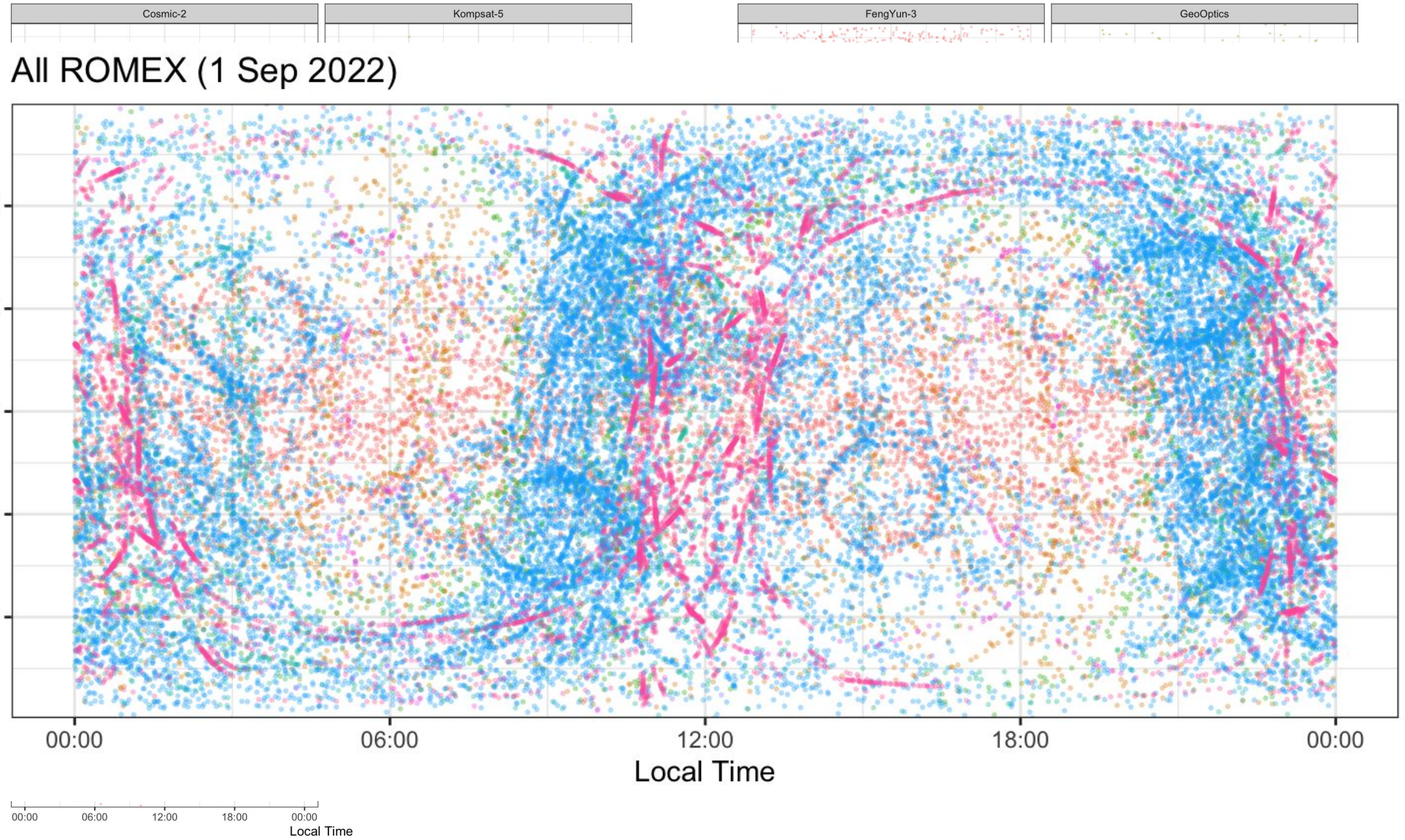
- Baseline missions – COSMIC-2, Metop, Sentinel-6, Paz, TanDEM-X & TerraSAR-X

ROMEX missions – All ROMEX missions



- Baseline missions plus Spire, Yunyao, PlanetIQ, FengYun-3, Tianmu & GeoOptics

Local time coverage

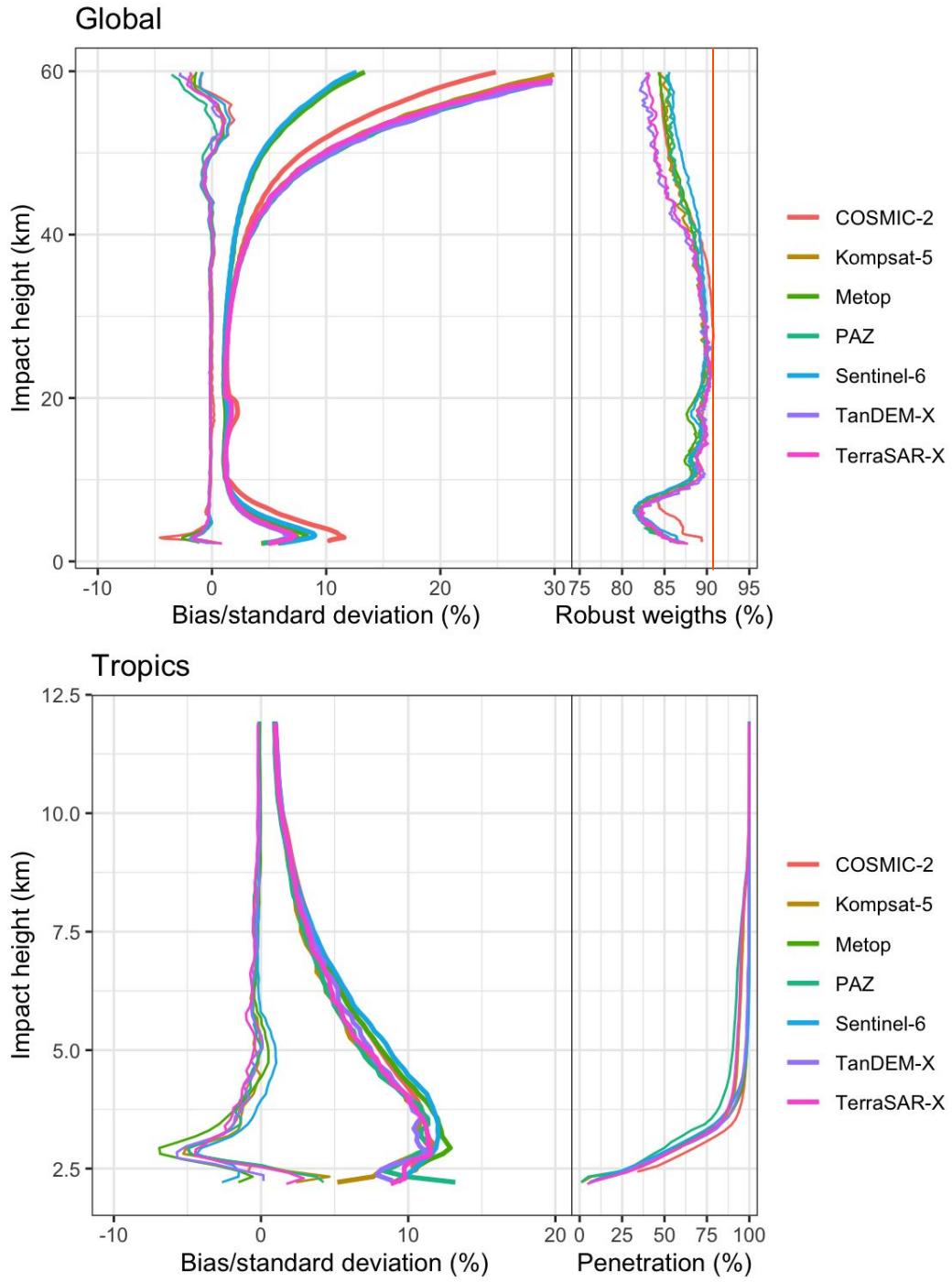


- *Despite the large number of occultations, local time coverage is not homogeneous.*

Baseline missions

- “Baseline” missions are either processed by UCAR or EUMETSAT and exhibit the well-known statistical properties known to all of us.

- Noteworthy:
  - GRAS and Sentinel-6 have by far the lowest stddevs above 40 km; COSMIC-2 performs worse despite high SNRs
  - UCAR-processed data set exhibit increased stddevs below 20 km
  - GRAS rising occultations have issues in the lowest few kilometres (setting occultations are similar to COSMIC-2 and Sentinel-6 wrt bias)
  - COSMIC-2 data outperforms other missions in terms of penetration into the lowest few kilometres
  - Shown are robust statistics; robust weights < 91.4% indicate non-Gaussian behaviour / long tails / outliers



Background: Robust statistics

- Traditional statistical methods (mean, standard deviation, ordinary least squares regression) are very sensitive to outliers, i.e. deviations from a Normal (Gaussian) distribution.

- Robust statistics provide methods insensitive to outliers and deviations from a Normal distribution.
- Typical examples: median instead of mean, MAD (Median Absolute Deviation) or IQR (Inner-Quartile Range) instead of standard deviation, M-estimators of location and scale.
- Robust estimators provide estimates of the mean and spread of the bulk of data, assuming that part of the data is normally distributed. They ignore outliers/heavy tails in the distribution.
- We use an M-estimator exploiting Tukey's biweight for location and spread (e.g., Maronna et al., Robust Statistics, 2nd Ed., 2019, section 2.6).
- M-estimators are a weighted mean of the original data points; outlying data points are downweighted.

 When analysing multiple data sets, some of them being treated by an (unknown) QC, others not being QC'ed at all, the calculation of ordinary statistics to make judgements on their quality is, at best, misleading. More effort is needed to either implement consistent QC for all data sets or apply advanced statistical methods (like robust estimators).

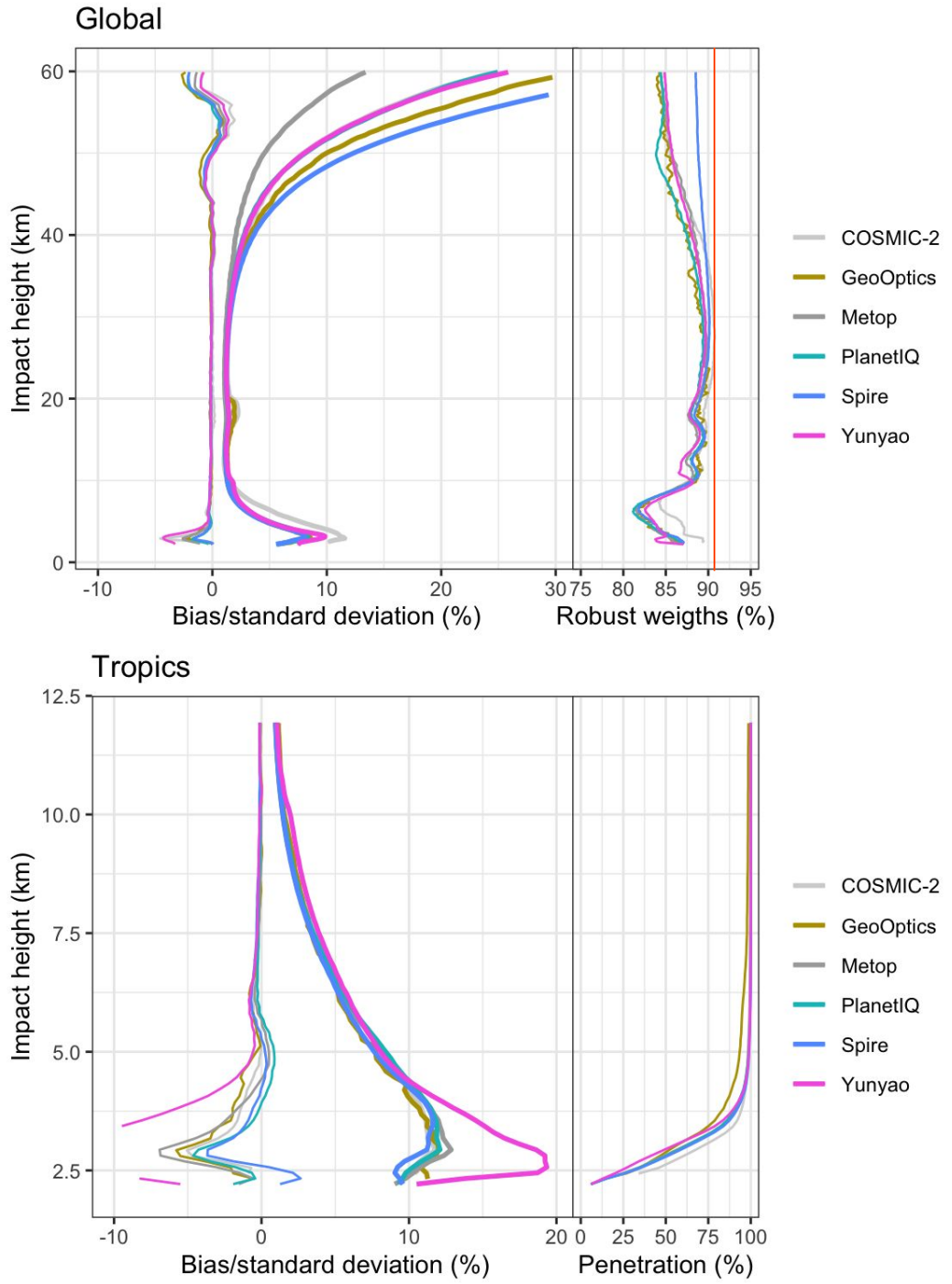
Commercial missions

- Spire, Yunyao, PlanetIQ and GeoOptics all perform very similar to the baseline missions.

- Exception:
  - Yunyao data (processed by EUM) exhibits large negative biases and large stddevs below 5 km impact height.
  - The problem is due to a very strict QC in Yunyao's lower-level processing, where carrier phase measurements are removed from the excess phase data.
  - We are working with Yunyao to improve the low-level data processing and improve the lower tropospheric retrievals.

  - The results of this work will be released as v2 of the Yunyao data.

 It is essential to have access to low-level data – level0 – to sort out issues or reprocess data.



Fengyun-3 & Tianmu

- A bit more complicated…

  - Overall, statistics of the bulk of FY3 and Tianmu GNOS data are in line with the other missions.
  - However, the number of outliers is higher than for baseline and other commercial missions.
    - Expect higher rejection rates in your QC.

  - FY3-C has a positive bias below 6+ km.
  - FY3-D setting has a bias issue above 40 km, and penetration problems in the troposphere.
  - FY3-E rising has large stddevs high up and exhibits very non-Gaussian behaviour.

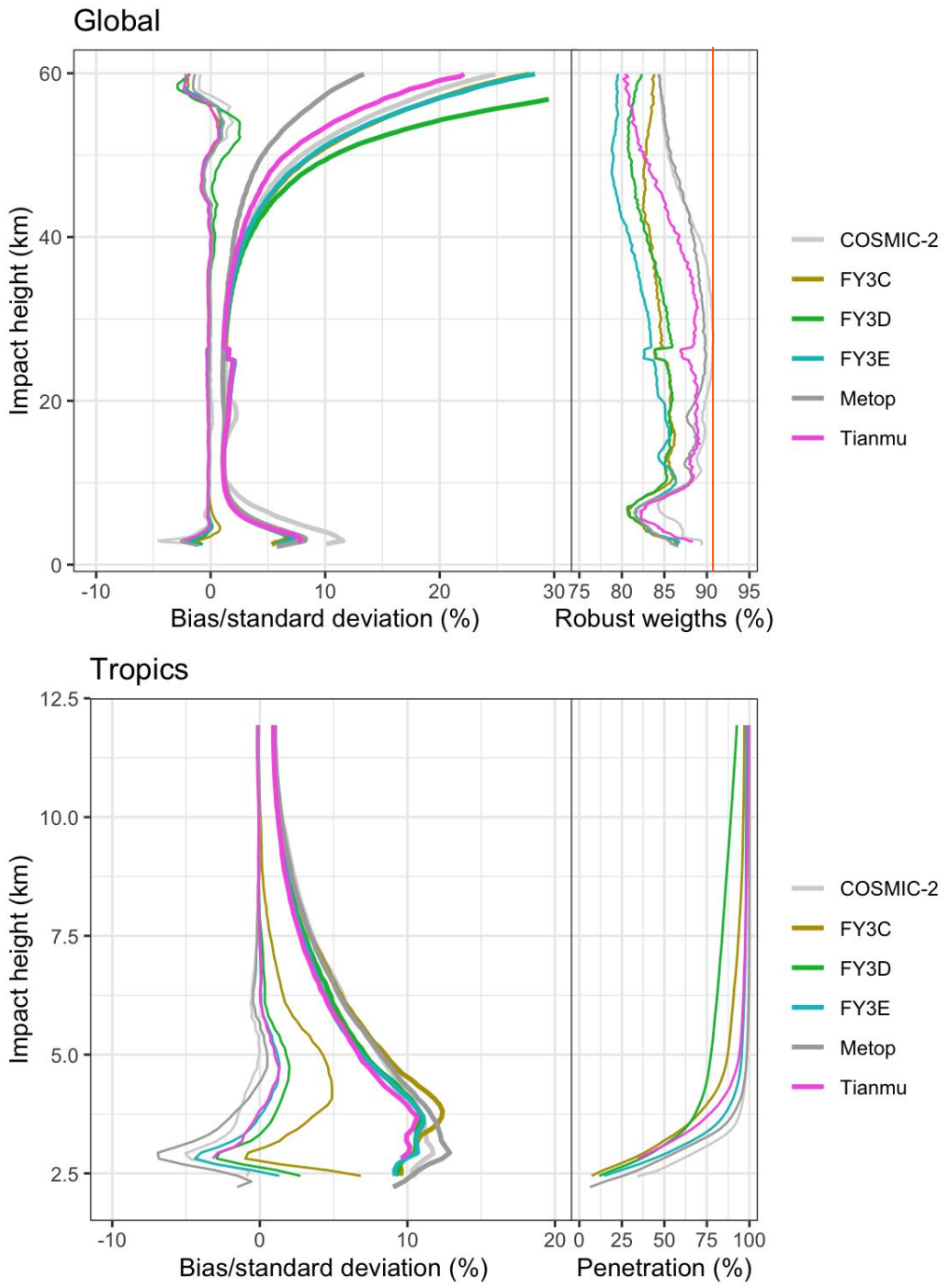
  - We are working with NSSC, CMA and Tianmu to improve their data quality and gave first promising results
  - Note that FengYun-3 data has been used since several years in European NWP, despite some known issues.

  - Plots are available on the following slide and in the annex.

Fengyun-3 & Tianmu (cont'd)

- Note the biases of
  - FY3-C below 7.5 km in the tropical troposphere
  - FY3-D above 40 km
- Also:
  - Reduced robust weights at nearly all heights above 10 km compared to COSMIC-2 and Metop
    - Deviations from ECMWF are more non-Gaussian, i.e. have many very large values due to large fluctuations in bending angle profiles
    - Provided QC removes those profiles the remaining ones have similar statistics than baseline missions.
  - Bulb in stddev around 27 km due to GO/WO transition in NSSCs processing – similar to UCAR's feature around 20 km.
  - Penetration into the lowest few km is comparatively poor in the original Fengyun-3 and Tianmu data.

  - Reprocessing Tianmu data by EUMETSAT provided excellent results, with Tianmu ROMEX data characteristics fully consistent with baseline and other commercial missions.
  - We are also looking forward to starting working on FengYun3-E data soon.



Data status & upcoming updates

- v1 of the ROMEX core data set is available on the ROM SAF server.
  - Some data sets missing refractivity.
  - Yunyao: Use care for tropospheric data below 5 km impact height until v2
  - Fengyun-3 & Tianmu: Use care in QC, expect higher rejection rates and satellite-specific issues

- v1.1 – No changes in bending angle profiles
  - Refractivity for Sentinel-6, Fengyun-3, Spire, PlanetIQ & Tianmu are available since a while; the ROM SAF has now finished Yunyao refractivities which will appear on the server shortly will appear shortly
  - Other fixes:
    - Updated smaller missions after UCAR kindly reprocessed them to fix data gaps
    - New satellite IDs in for Yunyao (901 instead of 1001) and Tianmu (902 instead of 1002)
    - Missing rising/setting information for FY3-C fixed

- v2 – enhanced bending angles, might come for GRAS, Yunyao, Tianmu and maybe even FY-3, but this is work in progress.

Data status & upcoming updates

- The 20k data set will appear on the ROM SAF server after IROWG.
- Bot UCAR and NOAA/STAR have provided their version of the reprocessed data set. It will be on the ROM SAF server as well.
- Katrin is preparing an ECMWF data set (operational analyses) for verification, which will also appear on the ROM SAF server.

 I apologise for some things taking too long – we are doing this on the side (as everyone), and sometimes other things happen (operational issues, broken feet,…)

Summary (cont'd)

- We have an unprecedented amount of RO data available in ROMEX.
  - However, local time coverage remains to be a problem, also with commercial data.

- With proper processing applied, all ROMEX data sets exhibit excellent data quality (at least when compared with NWP data). We believe the data quality is sufficient to conduct the experiments foreseen in ROMEX.

- Differences seem to be larger between processing centres than between missions or instruments.
  - *Customers who buy commercial data must know about RO processing and quality control.*
  - *Both NOAA and EUMETSAT continue to ensure consistency in data quality and quality control across missions and instruments by processing the data internally instead of buying higher-level products. Users benefit, so that is a good idea.*
  - *Instrument characteristics such as SNR performance appear in upper-level noise levels but not anywhere else. So – is it really only about numbers, reliability and low cost? Can ROMEX answer this?*

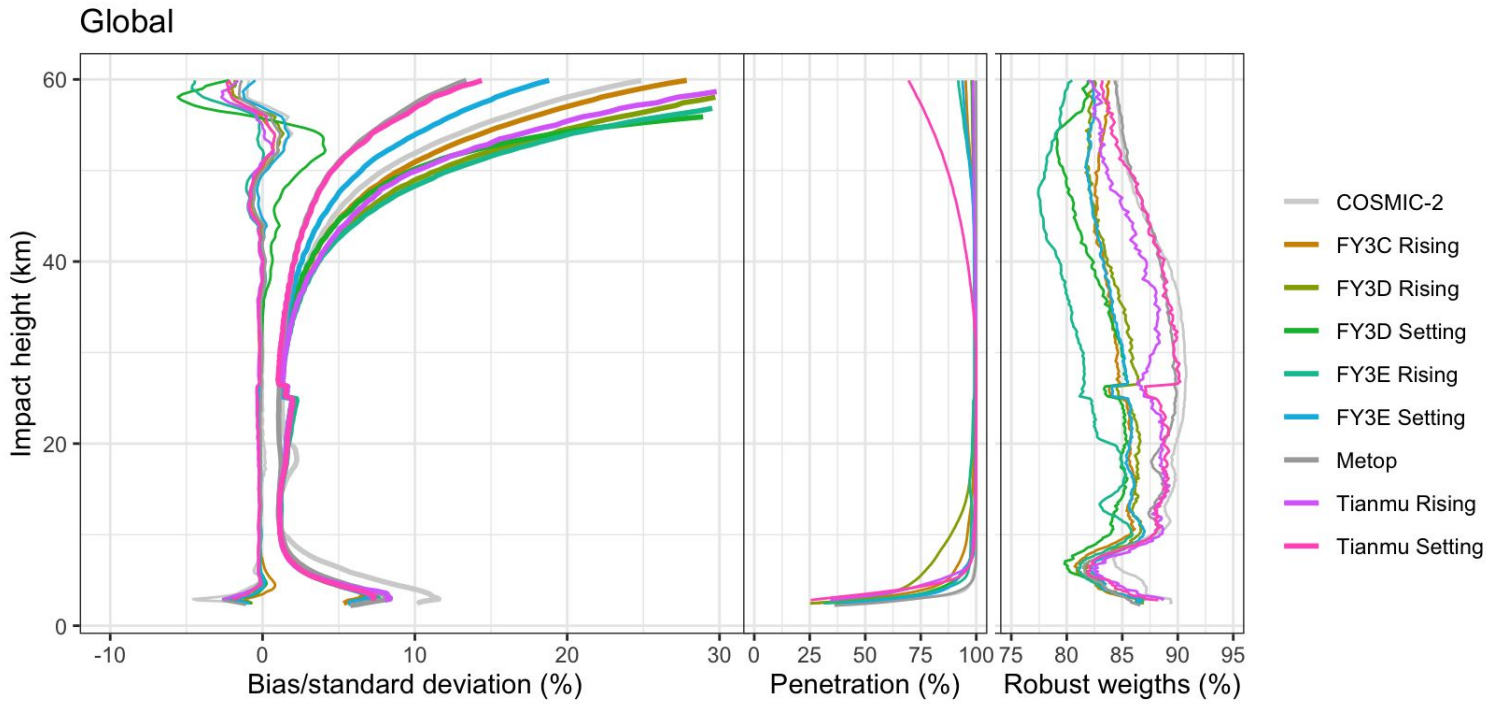
- That all said, these differences will be analysed and help all data providers to improve their data.

Thank you!

Questions are welcome.

Fengyun-3 & Tianmu – even more details

- Different FY3 satellites behave differently in their statistics, and between rising and setting
- FY3-D setting exhibits positive bias above 40 km
- FY3-E rising in particular exhibits very non-Gaussian statistics above 20 km, but the other ones – including Tianmu – suffer from many outliers as well.



- Tianmu performs slightly better than FY3's, but deviations from ECMWF are still more non-Gaussian than baseline (and other commercial) missions.

Reprocessing Tianmu data

- Reprocessing the ROMEX Tianmu data produced excellent results.