Oceanography Centre
Liverpool, UK
Joanne Williams and Andy Matthews

4th May 2020
EGU General Assembly

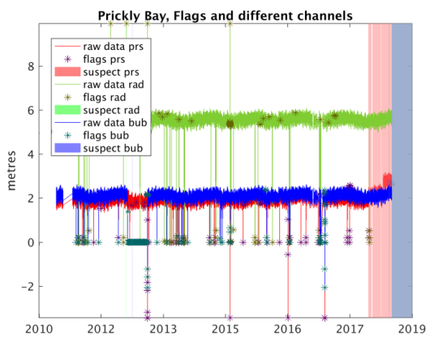
Lots of high-frequency data, no Quality Control

Tide gauges operated by BIG (Indonesian geospatial information centre) http://tides.big.go.id:8888

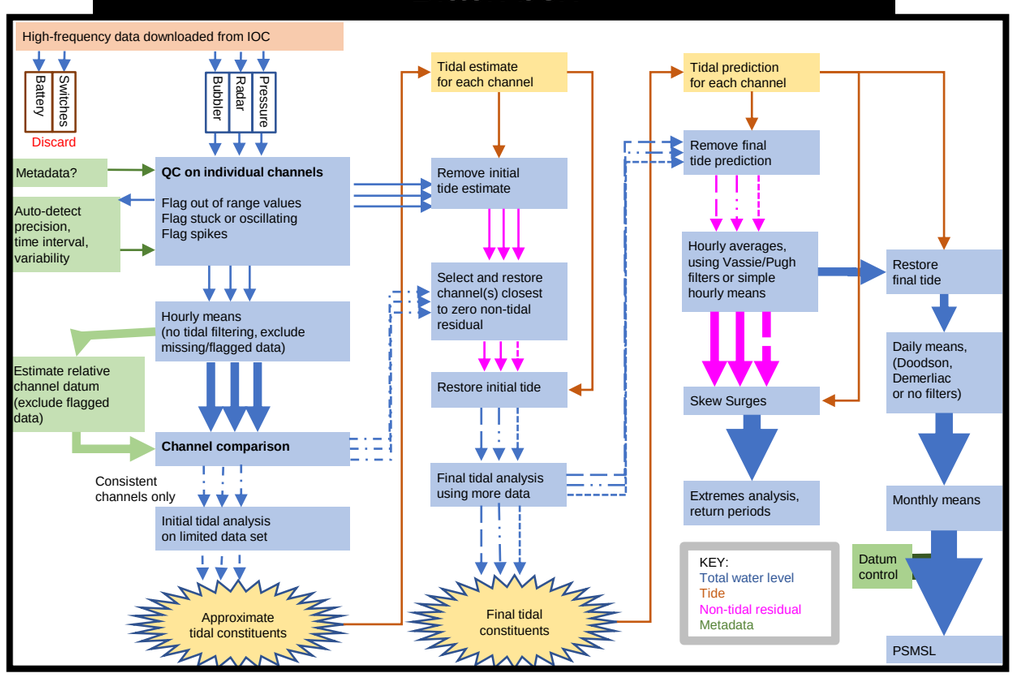
How can we implement Quality Control on all this data?

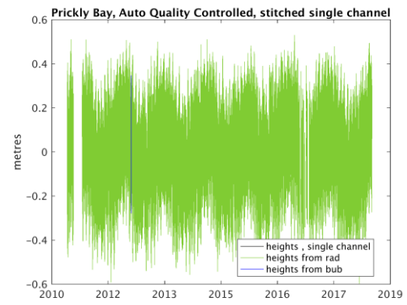
Much less in quality controlled global repositories

https://www.gesla.org/



Automatic Quality Control Black box



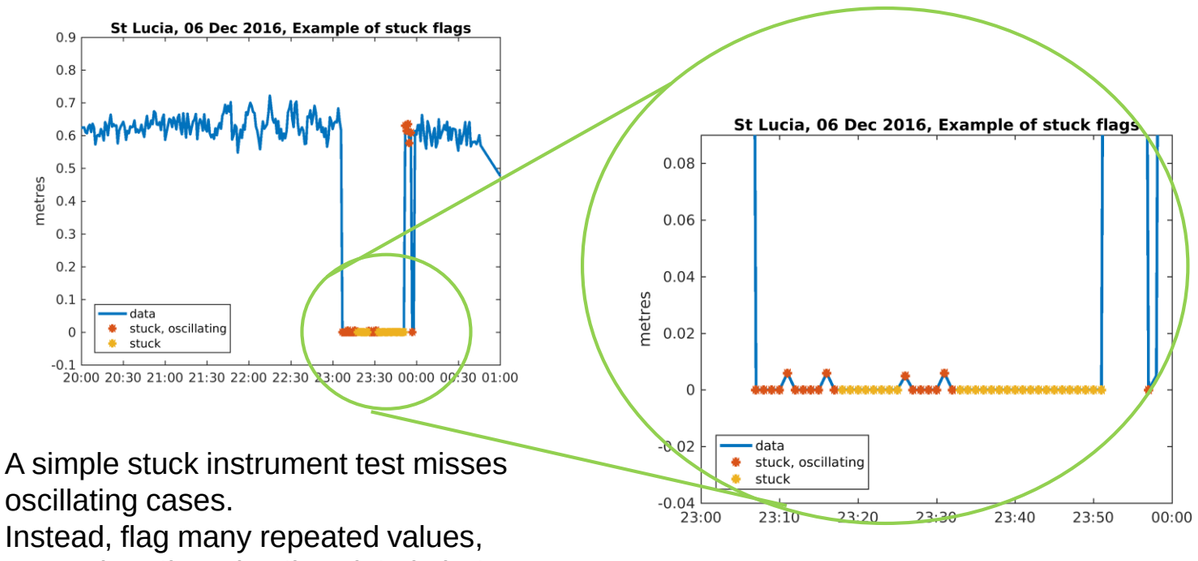


We have Matlab code, able to run directly on data downloaded from www.ioc-sealevelmonitoring.org .

Recent improvements have focused on the spike and stuck-instrument algorithms, particularly to handle low-resolution (1cm) data at high frequency (1 min) without any manual setting of parameters.

An early version of this work is described in detail at: www.psmsl.org/cme .
For the latest code please email us.
We are grateful for funding from ACCORD and the CME programme.

Stuck instrument detection



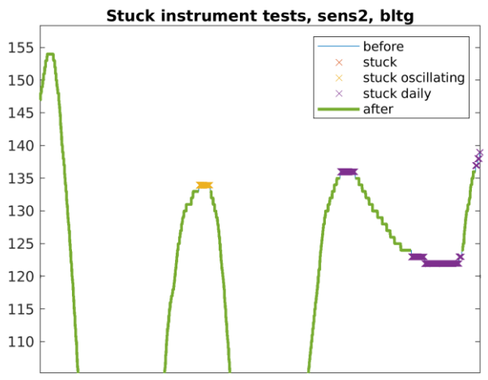
A simple stuck instrument test misses oscillating cases.
Instead, flag many repeated values, even when there is other data in between.

Stuck instruments: The problem of low precision

Very low precision recording with high frequency leads to genuinely constant data at slack tide.
This can be >100 consecutive data points!

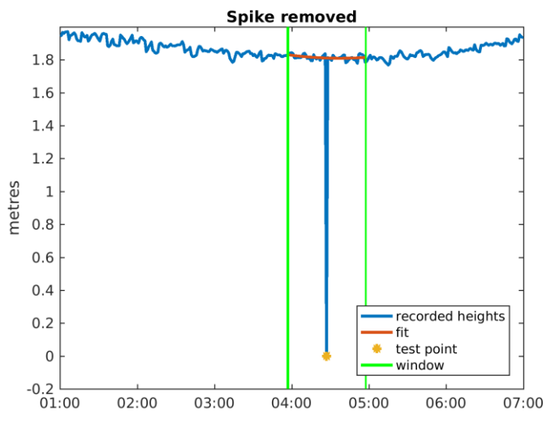
Parameters must be adjusted for different instruments.

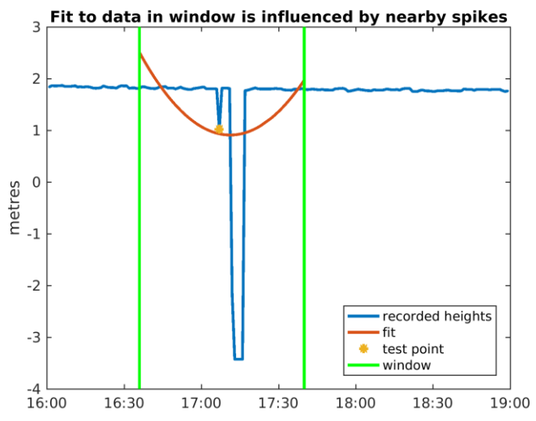
Our code now estimates timestep, precision, tidal range, noise levels and hence sets these parameters automatically.



Spike detection

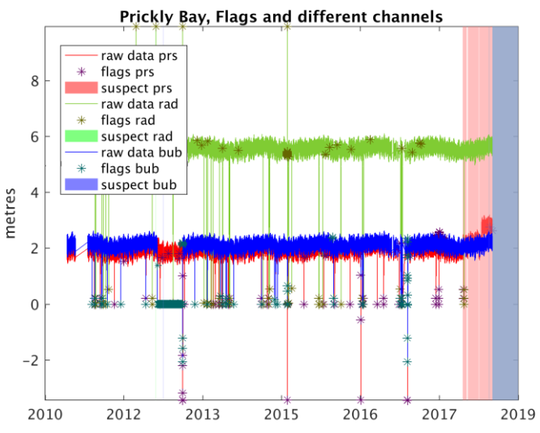
Spikes are detected using MEDIAN absolute deviation from a spline fitted to neighbouring points.
Clusters of spikes are challenging, but automatic parameter selection helps prevent false-positives and iteration helps eliminate all spikes.

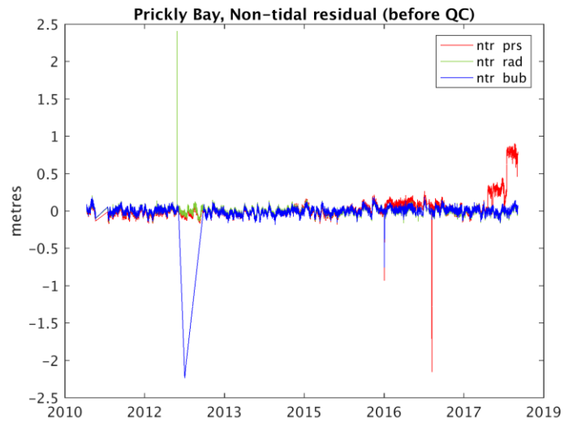




Channel comparison

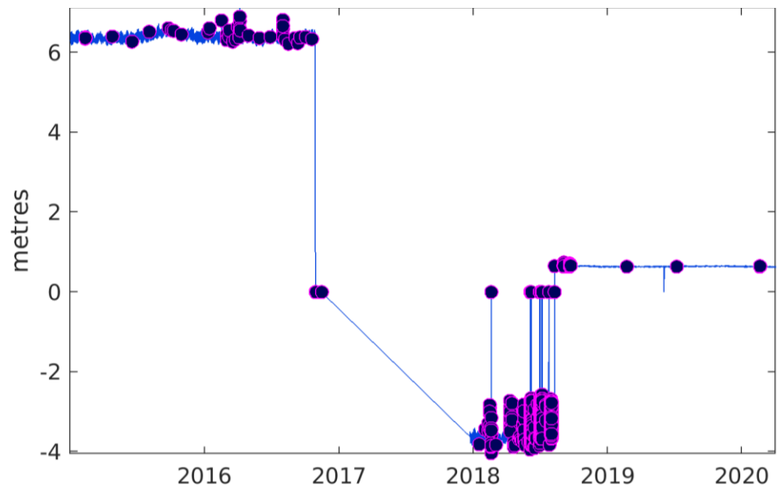
Hourly non-tidal residual is used for comparison between channels with different frequencies.
Constant offset is allowed, using median of unflagged data.
Channels are “dissimilar” for a time if they are more than 4 x median difference apart.
The best channel for a time is assumed to be the closest to the initial predicted tide.
Channel comparison is then applied to original-frequency data.

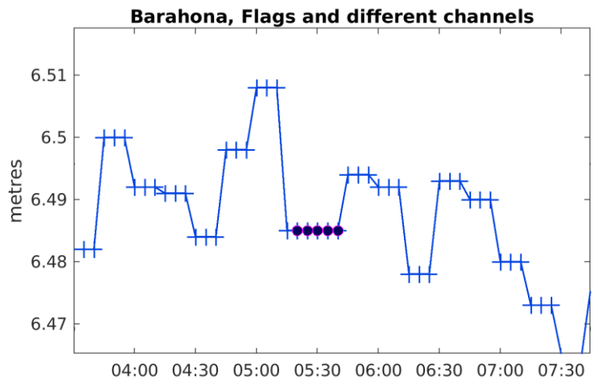




Some of the remaining challenges

Barahona (Dominican Republic) shows up some remaining problems.
There appear to have been several instrument deployments, that are treated as one record although the metadata only refers to the most recent.





Here the data also appears to have been filled from 15 min to 5 min intervals with nearest neighbour interpolation. This confuses both stuck and spike tests.
It requires specific intervention to fix.

Elsewhere at EGU: More about PSMSL in https://meetingorganizer.copernicus.org/EGU2020/EGU2020-3054.html (G2.1, Chat Tuesday 08:30am)
This work was first developed for a sea level project: https://meetingorganizer.copernicus.org/EGU2020/EGU2020-6495.html (CL4.5, Chat Monday 08:30)