Automated outlier detection with machine learning in GRACE and GRACE-FO post-fit residuals

Jonas Zbinden, Martin Lasser, Ulrich Meyer, Brandon Panos, Daniel Arnold, Adrian Jäggi

Problem

GRACE/GRACE-FO inter-satellite K-band range-rates are the main observable for the determination of monthly solutions of the Earth's gravity field. The range-rates are sensitive not only to the mass distribution of the Earth, and as a consequence, the relative motion of both GRACE and GRACE-FO satellites respectively, but also to the relative orientation of the satellites and consequently to the attitude handling. Therefore, an efficient screening of the range-rate observations is not trivial. In this contribution, we apply machine learning to flag outliers in an unsupervised and fully automated way.

Isolation forest for outlier detection

For outlier or anomaly detection isolation forests offer a fast and easy way to detect outliers within a high-dimensional dataset.
The data is split along each dimension randomly. The easier it is to isolate a point, the higher its anomaly score.

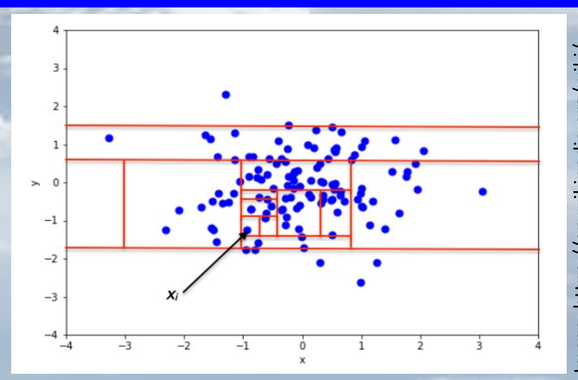


Image: https://en.wikipedia.org/wiki/Isolation_forest

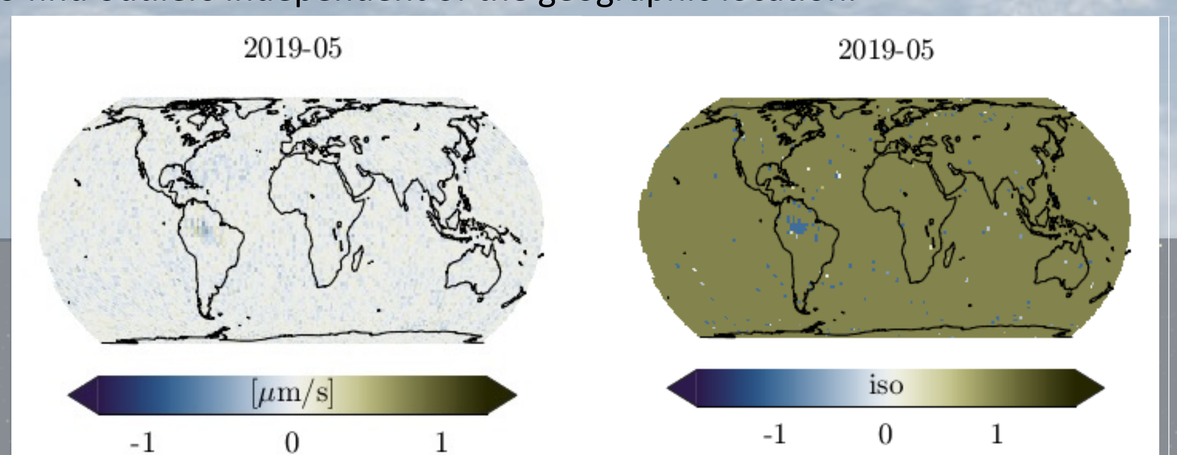
The anomaly score $s(x)$ of a point x_i is measured with the number of random cuts $h(x_i)$ to isolate it. The score $s(x)$ is given by

$$s(x) \propto -2^{-\mathrm{E}(h(x))},$$

where $\mathrm{E}(h(x))$ is the expectation value of the path length of all trees to isolate the point x. A score $s(x)$ close to -1. The minus is a convention from the Python library scikit-learn.

GRACE-FO outliers based on isolation forest

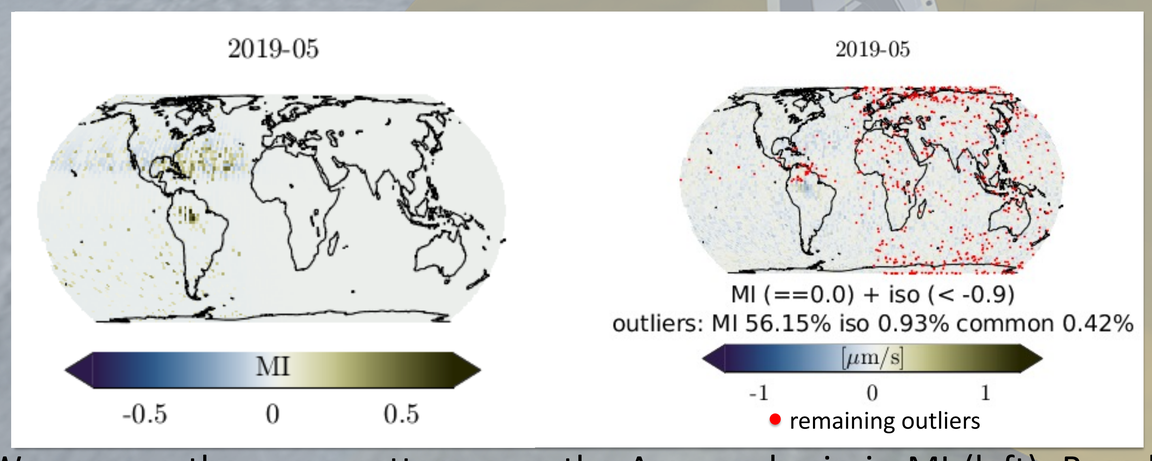
We applied the isolation forest algorithm to the combination of the post-fit residuals and the geographic location of the satellite at this time to find outliers independent of the geographic location.



As an example, we show the post-fit residuals (left) and the outliers found with isolation forest (right) from May 2019. Over the Amazon basin both maps exhibit some geographical pattern. To distinguish between true anomalies, and unfitted gravity signal we make use of mutual information to find any (cor)relations between the geographical position and the post-fit residuals.

Isolation forest + Mutual information

Mutual information MI can show more complex relations and dependencies between the geographic position and the post-fit residuals. If there is any geographically (cor)related signal in the post-fit residuals, then $MI \neq 0$. Therefore, we strive to find true outliers flagged by isolation forest which have $MI = 0$.



We can see the same pattern over the Amazon basin in MI (left). Based on the combination of isolation forest + MI we can isolate the true outliers (right), not affected by a bad model fit.

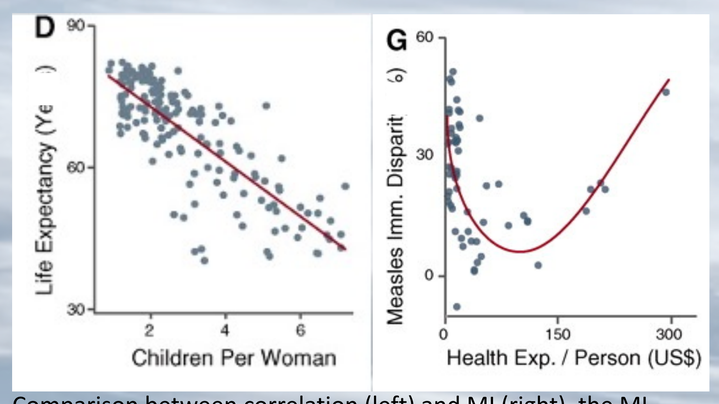
Mutual information

Mutual information MI captures the dependency between two random variables and their probability distributions. Two variables X, Y are independent if and only if

$$p(x, y) = p(x)p(y).$$

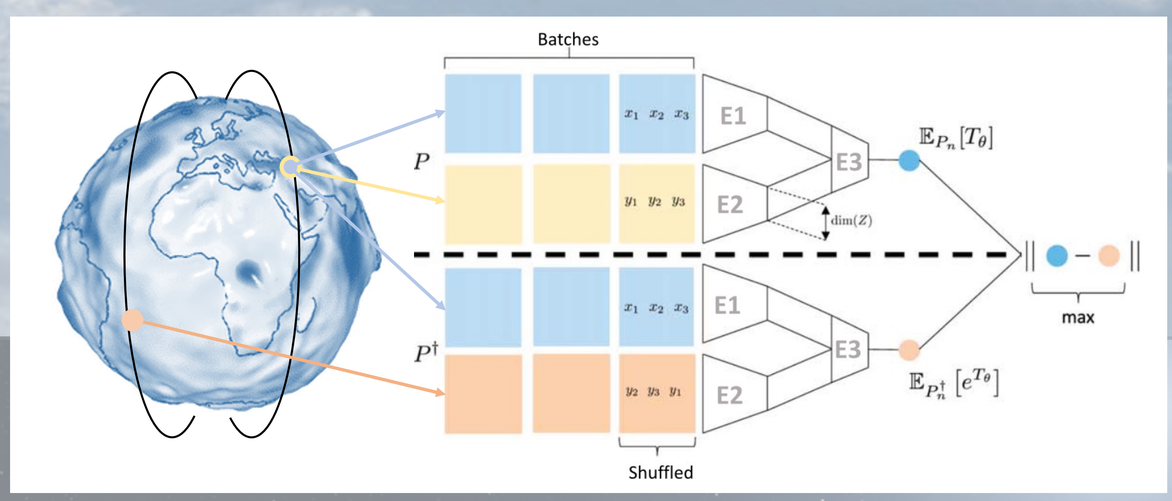
To capture the dependency between x and y we compute the pointwise-mutual information measure $PMI(x_i, y_i)$:

$$PMI(x_i, y_i) \triangleq \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)}.$$



Comparison between correlation (left) and MI (right), the MI also captures non-linear dependencies. Image: https://www.science.org/doi/10.1126/science.1205438

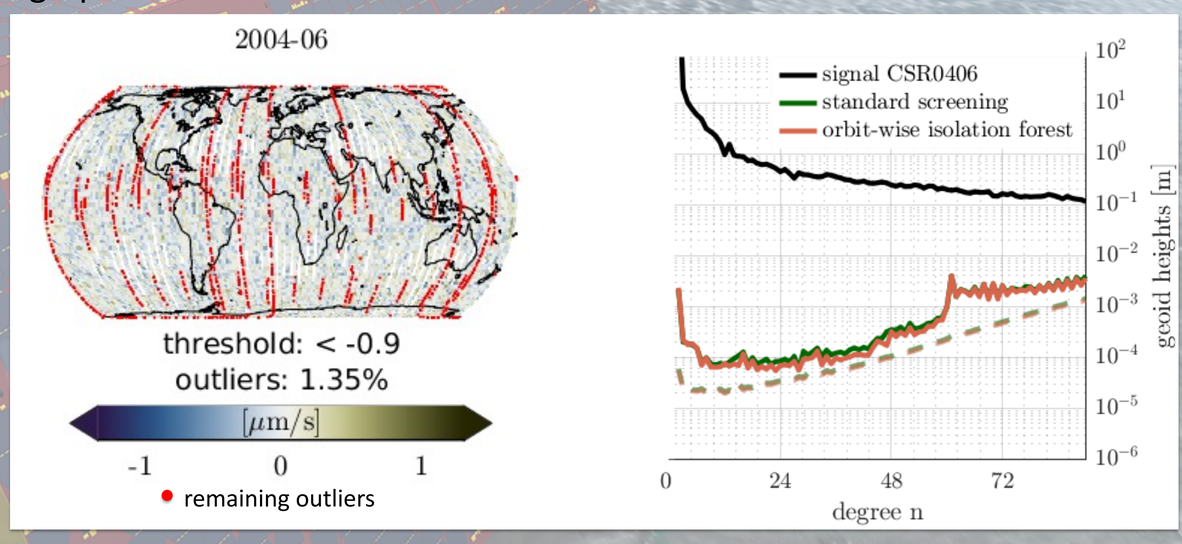
The PMI can be calculated by training a neural network T at capturing the dependencies between X and Y for instance through sampling the marginal probabilities $\mathrm{P}(X)$, $\mathrm{P}(Y)$ (●) by sampling signals from different pixels in an image and the joint probability distribution by sampling signals from the same pixel $\mathrm{P}(X, Y)$,(●) maximizing the difference || ● - ● || .



With our trained neural network T we can then estimate the $PMI(x_i, y_i)$ on e. g. the position on Earth and the post-fit residuals for a single timestep.

Outliers based on sequences of single orbits

Due to the lower global coverage by GRACE in 2004 (repeat cycles) compared to GRACE-FO, we cut the data into single orbits, each segment going from North pole - North pole. We applied the isolation forest algorithm to the residuals in each orbit to incorporate time and reduce the bias from the geographical distribution of the orbits.



Most but not all outliers coincide with what we would remove by eye as bad data ('standard screening'). Additionally, the gravity field uncertainty estimation improves compared to a manual removal of outliers from visual inspection of post-fit residuals time series.

Conclusions

We have successfully shown how anomalies in the post-fit residuals have been identified as true anomalies with **isolation forest** combined with **mutual information** MI to keep remaining anomalous signal. Additionally, we have shown that the post-fit uncertainty estimation **improves** for **GRACE**, when removing outliers on an **orbit-by-orbit** basis with **isolation forests**. Due to the inhomogeneous coverage **isolation forest + MI** was not applicable to the **GRACE** dataset from 2004. The model fit errors are compared to a (tedious) manual removal of outliers by visual inspection. While we can see an improvement for **GRACE**, removing the flagged outliers from the **GRACE-FO** did **not improve** the monthly Earth gravity field solution so far. In future we could investigate with a parameter grid search, if there is an optimal set of parameters for the outlier selection to improve the model fits for **GRACE-FO** as well.

Background Image courtesy: https://www.exploratorium.edu/climate/atmosphere, https://gracefo.jpl.nasa.gov/mission/overview/