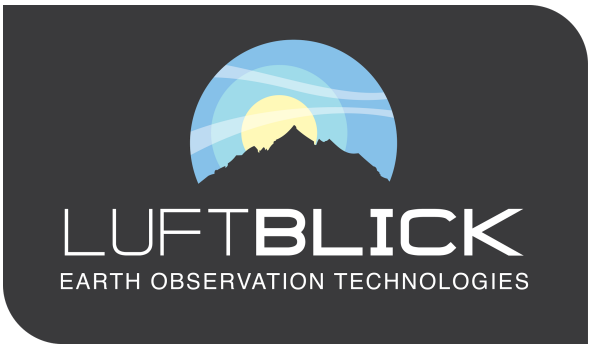


LuftBlick report 2022004

Quality Assurance for Earth Observation - QA4EO

WP2322 - TN on Pandora Lunar data quality assessment methods

	Name	Company	Date
prepared by	Manuel Gebetsberger	LuftBlick	13 Apr 2022
	Martin Tiefengraber	LuftBlick	13 Apr 2022
checked by	Alexander Cede	LuftBlick	13 Apr 2022

Table of Content

Document Change Record

Issue	Page	Date	Observations
1	All	2022-04-11	First version

1 Summary and Conclusions

The D-7 “Pandora Lunar data data quality assessment methods” of WP2322 of the ESA project “Quality Assurance for Earth Observation” (QA4EO) [AD1] has the goal to investigate the data quality for selected trace gas products retrieved from direct lunar observations. The method used to assess the data quality is called SABAT, which minimizes the comparison noise by compiling a close-to-truth reference dataset for the comparison. By this, systematic differences between datasets (bias) can be quantified and the remaining random error (noise) can be approximated. The retrieved biases between the evaluated instruments with 1-sigma uncertainty and the noise values are reported in Table 1.1. The bias is also reported in percentage relative to the median slant column amount for measurements below air mass factor 1.5.

Overall, the analysis highlights the ability to retrieve trace gas columns from Pandora direct lunar observation. By far, the H_2O product shows the best performance with less than 1% systematic difference between two instruments, and 4.4% random residuals. The largest difference was observed for the NO_2 product, which can be attributed to the calibration process. It has to be noted that the processed data are using a synthetic reference spectrum made from direct sun observations. Therefore, the reference uses a different filter setting as when taking the lunar measurements, which points out the need of improving on L1 characterization steps to close this gap.

Results of SABAT could in principle also be used in further studies to validate parts of the reported uncertainty in the data products. However, first results suggested that a yet unaccounted error source is distorting the results. This missing part reflects uncertainties related to differences between the model and the measurements can be

denoted as “discrepancy” uncertainty. This uncertainty source is foreseen to be included in an updated version of the Blick Software Suite.

Table 1.1: Data product quality summary for direct lunar observations of O_3, NO_2, H_2O and NO_3. Units are given in box brackets. The numbers in curved brackets report bias (left column) or noise (right column) in percent relative to the median slant column amount below air mass factor 1.5.

	Bias	**Noise**
O_3 [mmol/m^2]	5.406 +/- 4.8 (3.3 +/-2.8 %)	4.607 (2.8%)
NO_2 [mmol/m^2]	0.052 +/-9e-3 (23.7 +/- 4.3%)	0.022 (9.9%)
H_2O [mol/m^2]	13.68 +/- 65.2 (0.7 +/- 3.5%)	81.56 (4.4%)
NO_3 [mmol/m^2]	6.2e-4 +/- 5e-4 (13.9 +/- 12.5%)	4e-4 (9%)

2 Document logistics

2.1 Acronyms and Abbreviations

AD	Applicable Document
BrO	Bromide Oxide
D	Deliverable
I2	Molecular Iodine
H2O	Water Vapor
HCHO	Formaldehyde
HONO	Nitrous Acid
NO2	Nitrogen Dioxide
NO3	Nitrate Radical
O2	Molecular Oxygen
O2O2	Oxygen Dimer
O3	Ozone
OFFS	Offset Polynomial
OIO	Iodine Dioxide
PGN	Pandonia Global Network
PIT	Probability Integral Transform
QA4EO	Quality Assurance Framework for Earth Observation
RD	Reference Document
Rome-SAP	Sapienza University, Rome
SABAT	Smooth Approximation of a BAseline Truth
SMO	Smoothing Polynomial
WLC	Wavelength Change Polynomial

2.2 Applicable Documents

[AD1] Quality Assurance for Earth Observation project [Annex B, Statement of Work for LuftBlick], SERCO Contract QA4EO/SER/SUB/04, 2020.

2.3 Reference Documents

[RD1] A. Cede, M. Tiefengraber, M. Gebetsberger, and E. Spinei Lind. Pandonia Global Network Data Products README Document, 2020.
[RD2] H.H. Kieffer, T. C. Stone, The spectral irradiance of the moon. Astron.J. 129, 2887–2901, 2005.
[RD3] Tiefengraber, M., Cede, A., Gebetsberger, M., and Mueller, M.., LuftBlick Report 2019005: ESA Fiducial Reference Measurements for Air Quality, New Algorithm & Product Development Plan, Jan 2022.
[RD4] Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecast from ensemble model integration. *J. Climate, **9**, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2*
[RD5] Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, doi:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2
[RD6] Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability,* Shinfield Park, Reading, Berkshire, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25
[RD7] Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, doi:10.1175/WAF993.1
[RD8] Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x
[RD9] Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, Estimation Methods for Non-Homogenous Regression Models: Minimum Continuous Ranked Probability Score versus Maximum Likelihood, Monthly Weather Review, **146** (12), 4323-4338, doi: 10.1175/MWR-D-17-0364.1, 2018
[RD10] Tiefengraber, M., Cede, A., LuftBlick Report 2021002: TN on Pandora Lunar data processing set-up, Dec 2021.

3 Quality assessment

3.1 Method

The quality assessment method for lunar trace gas retrievals has the goal to provide independent estimates of the quality of the data products by addressing two major aspects:

a) which systematic differences can be expected between multiple instruments (referred to as **'bias'**)
b) what discrepancies can be expected if multiple instruments would have no bias (referred to as **'noise'**).

Addressing these two points would typically require a known truth that can be used for comparison to the retrieved data product. However, this universal truth is unknown, and due to the lack of potential comparison (reference) datasets for trace gas retrievals using direct lunar observations, an internal validation is performed with Pandora data only. Internal in the sense that two Pandoras were measuring at the same location during the same period of time. Since both instruments have been calibrated independently, they should therefore deliver, in theory, equal outputs if they would have been measured at exactly the same time.

This assumption allows us to derive a noise-free, and therefore smoothed, baseline truth using an approach called SABAT (**S**mooth **A**pproximation of a **BA**seline **T**ruth). This approach creates a bridge between multiple measurements to calculate a best guess of the truth in order to address a) and b).

SABAT, as described in more detail in RD3 (appendix section D), is performed on slant column amounts for a given set of instruments, and aims in estimating a joint baseline variation for each day to avoid analysis noise, in contrast to classical time interpolation strategies. Within this baseline variation, changes due to statistical noise are removed, but intrinsic systematic errors are considered. SABAT achieves this by optimizing the degrees of

freedom that are used to follow the measured structures based on the Bayesian information criteria (RD3). Hereby, overfitting the data is avoided by penalizing with the number of data and parameters used. The resulting baseline variation can be interpreted as a noise-free effect, which is shifted into each measured time series due to a potential calibration error or intercept. This calibration error can result from the field calibration where the slant column amount in the synthetic reference spectra is determined, and is referred to as common uncertainty in the data products (RD1).

SABAT will utilize spectrometer 2 data based on the same algorithms as described in RD10, from Pandora 117 and Pandora 138 located at the Sapienza university in Rome. A comparison using only two instruments has to be treated with care in terms of interpretation of differences, due to the possibility of one instrument having an issue, and the truth is unknown.

While a) can be estimated with the overall intercept between the slant column time series, b) results from residuals of the intercept-corrected baseline variation.

3.2 Extensions and current limitations

With the assumptions made in SABAT, an extended validation for parts of the reported uncertainties in the PGN data products (RD1) can be performed, namely

A) the common and
B) the independent uncertainty.

While A) can be already captured by the range of the obtained intercepts, B) could be validated by the remaining residuals to the intercept-corrected baseline variation. This would allow us to quantify the statistical consistency of B) which should be noise only, per definition.

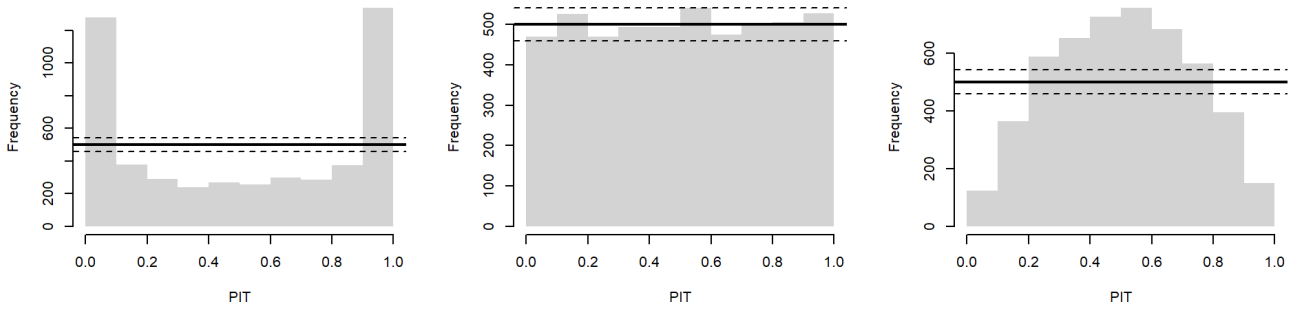
First results for B) were counterintuitively reporting a too small uncertainty, which indicates a potentially missing error source. This error source, which we call “discrepancy”, cannot be explained by noise or systematic errors solely. It can be caused by a mismatch between the measurements and the model, and might vary from instrument to instrument. It is foreseen to include this additional error source in an updated Blick Software Suite, which then allows SABAT to properly evaluate these uncertainty parts.

A useful tool to quantify the uncertainty of a probabilistic data product is defined by PIT histograms (RD4-6). Hereby, each individual Pandora measurement is taken as a Gaussian distribution with the reported expectation value and the reported independent uncertainty as distribution parameters, and the calculated baseline amount is taken as the best estimate of the true value. By binning the cumulative distribution function into probability bins, the observed frequency for each bin can be compared to the expected frequency.
E.g. an equal binning of 0.1 would evaluate probability bins for 10% which would mean that the baseline amount is ideally observed in 10% of the cases in each bin. Consequently, a histogram of the observed frequencies should be uniformly distributed, if the uncertainty reporting is appropriate. Conversely, if the reported uncertainties are too small, the observed baseline amount would be observed on the tails of the distribution function, which would lead to an U-shape of the PIT histogram. Examples are illustrated in Figure 1, where 5000 synthetic Gaussian data are drawn from an expectation value = 0 and a standard deviation of 1. If those data are validated as if they would have been drawn from a Gaussian distribution with standard deviation = 0.5, the resulting PIT would show the typical U-shape (left), and an inverse-U shape if the standard deviation would be 1.5 (right). But if the proper standard deviation is used, which is 1, the PIT illustrates the expected uniform distribution.

The concept of a probabilistic reporting should therefore follow the principles of sharpness and calibration (RD8), where sharpness refers to the reported uncertainty to be as small as possible, and calibration, which refers to the statistical consistency as expressed by the PIT.

To further increase the reliability of the PIT histograms, a consistency sampling, as suggested by RD7 and RD9, is applied. Based on the number of observations that are evaluated, and the number of bins used, a confidence interval can be drawn around the expected frequency that is based on a Binomial distribution of size *n* and the probability of success in each trial, which is given by the bin-probability.

Figure 1: Examples of PIT histograms from left to right using synthetic data from a Gaussian normal distribution with mean=0, sigma=1, but evaluated with a reported sigma that is 0.5 (left), 1 (middle), 1.5 (right).



4 Data Product Quality Assessment

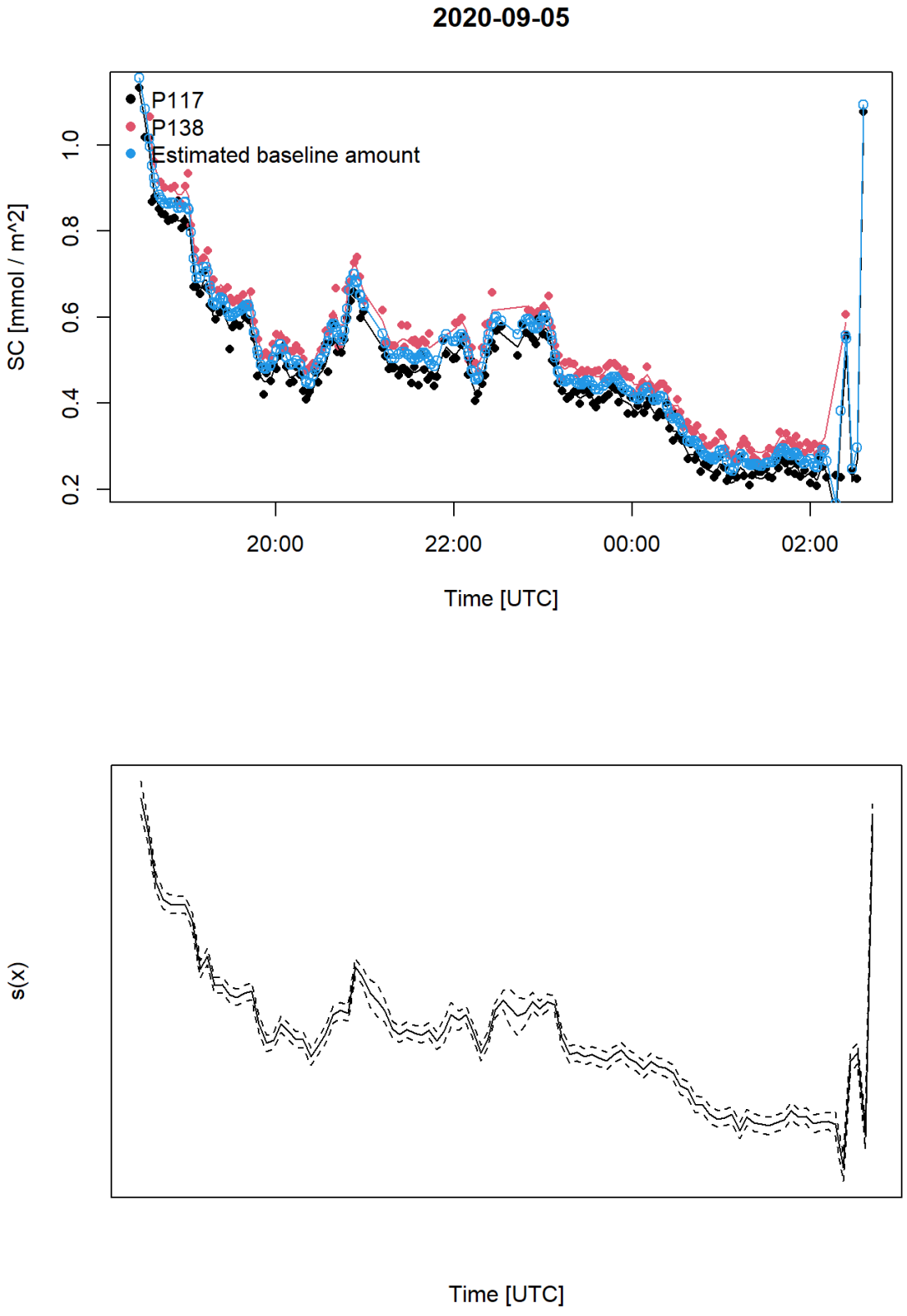
In the following subsections the lunar retrievals for NO_2, H_2O, NO_3, and O_3 are illustrated by:

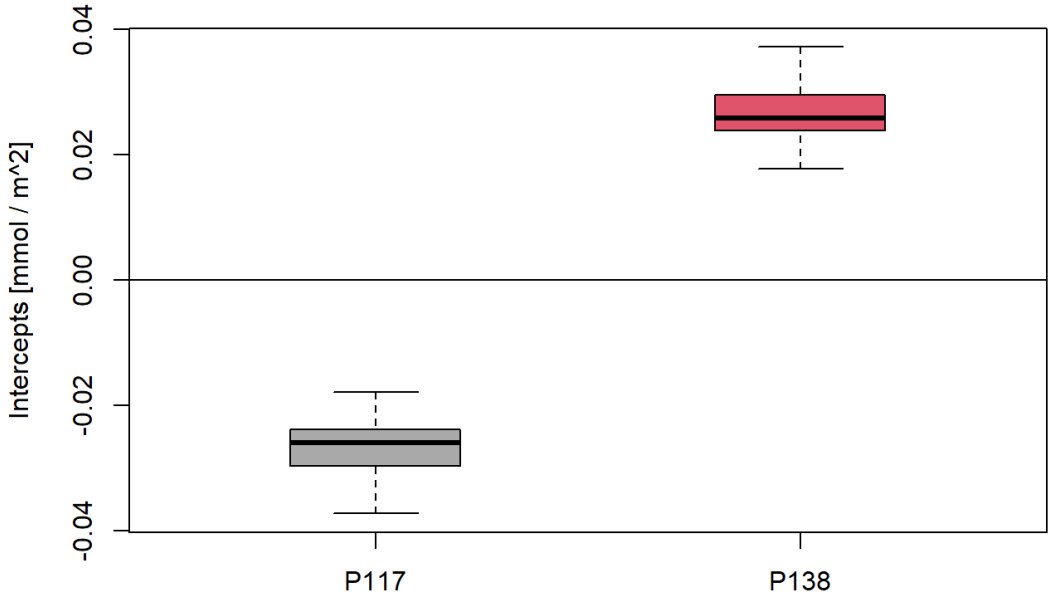
- First plot: An example day illustrating the slant column amounts for P117 (red dots) and P138 (black dots) and the estimated baseline amount (blue dots), where the red and blue lines are the respective instrument specific baseline amounts.
- Second plot: The estimated daily variation effect s(x) where x is the time of the day.
- Third plot: The evaluated intercepts for P117 and P138 compared to the baseline amount as boxplots showing the interquartile range (0.25-0.75) in boxes, and whiskers for +/- 1.5 times the interquartile range.
- Fourth plot (left, right): The residual distribution compared to the baseline amount for AMF groups 1-3 and 3-5.

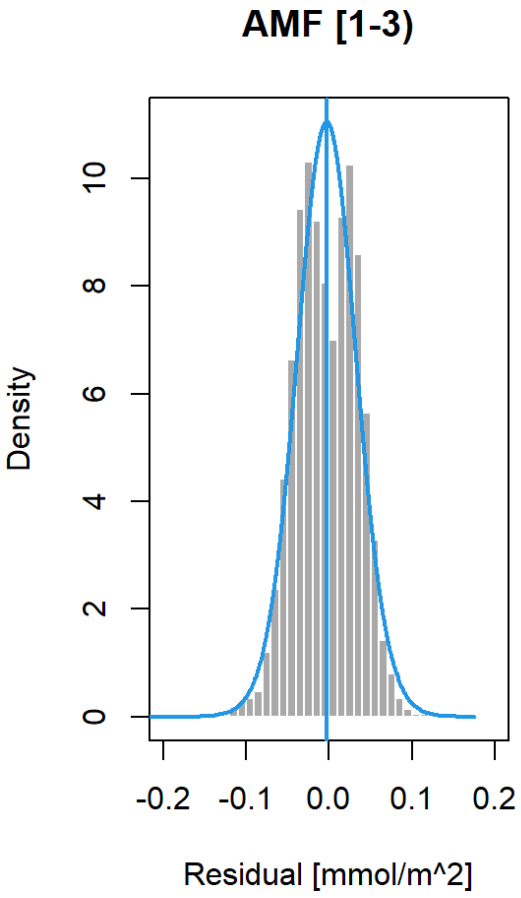
4.1 Lunar total column NO_2

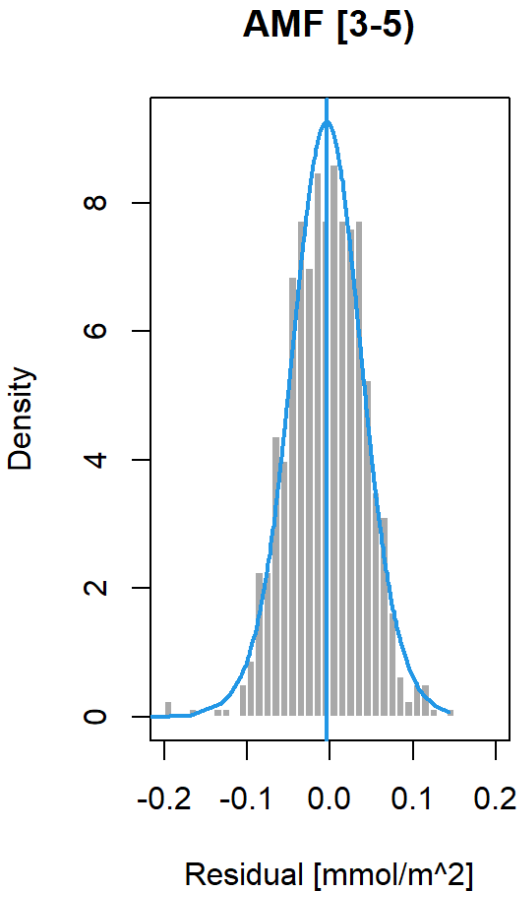
The instruments illustrate an agreement in the daily variation which can be captured by a daily common effect. However, there is a systematic calibration offset in the median slant columns of P117 of 0.0518 mmol/m^2, which is illustrated by the intercepts. This range is much larger than the reported common uncertainty as obtained by the calibration process, which is 0.007 mmol/m^2 (P117) and 0.017 mmol/m^2 (P138). This can most likely be linked to uncertainties in the L1 optical filter calibration, since the reference spectrum is compiled out of a different set of filters (diffusers and NDs) as used during the lunar observations (no filter in place).

This leads to an overlap of two the two error distributions at AMF group 1-3, with the 68% error interval from +/- 0.036 mmol/m^2 . This overlap is not present for AMF group 3-5 (+/- 0.043 mmol/m^2), which implies that the baseline amount is less certain about a common effect for this second group and obtains an average baseline amount.



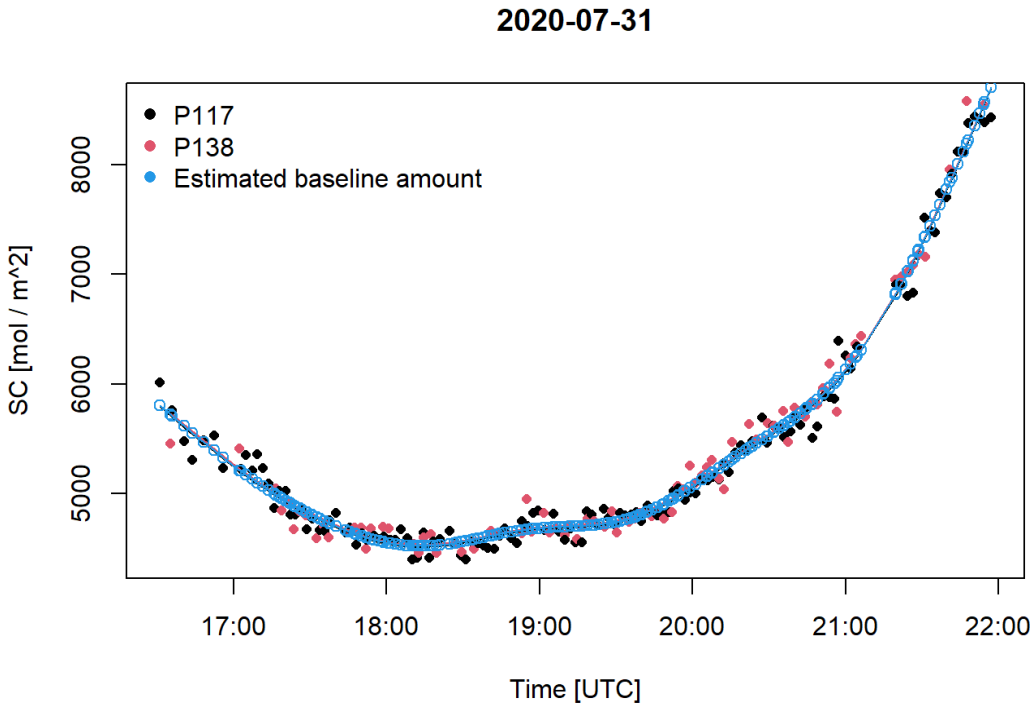

Intercepts [mmol / m^2]
0.04
0.02
0.00
-0.02
-0.04
P117
P138

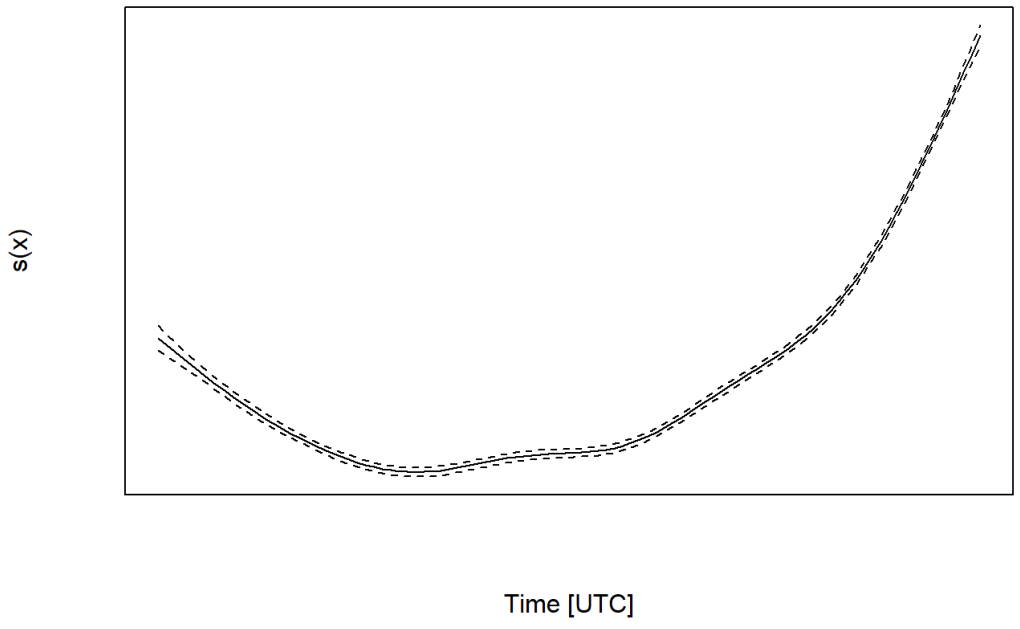

AMF [1-3)
Density
0
2
4
6
8
10
-0.2
-0.1
0.0
0.1
0.2
Residual [mmol/m^2]

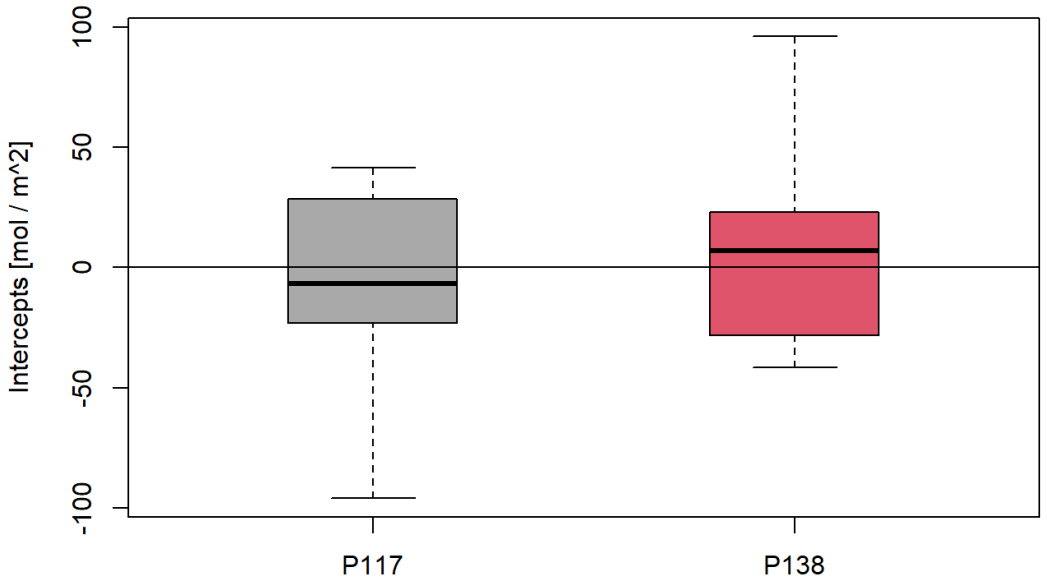

AMF [3-5)
Density
0
2
4
6
8
-0.2
-0.1
0.0
0.1
0.2
Residual [mmol/m^2]

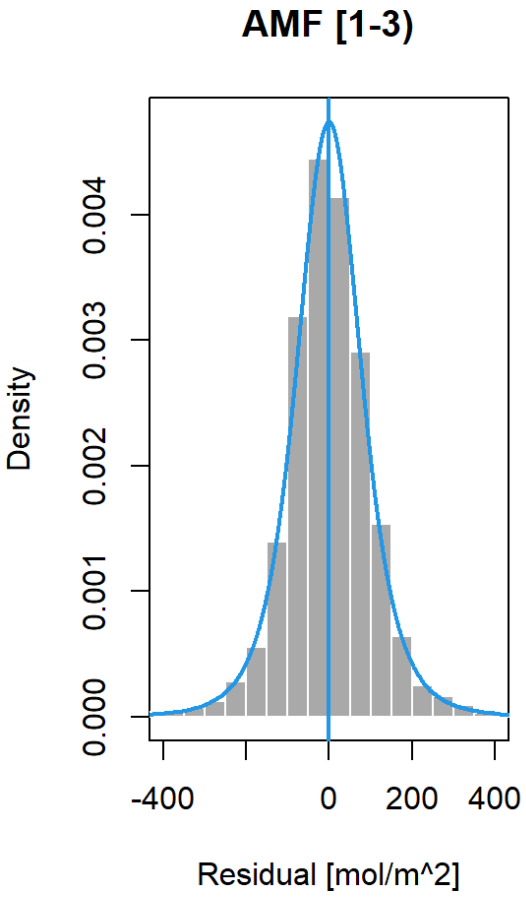
4.2 Lunar total column H_2O

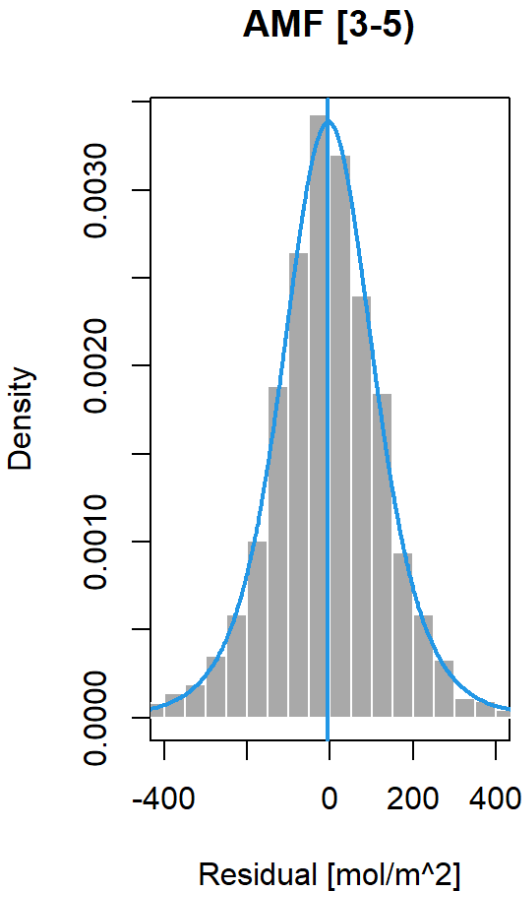
The two instruments show a good agreement in terms of calibration where the systematic offset in the slant columns as described by the median intercepts is 13.7 mol/m^2. This is even smaller than the estimated common uncertainty during the calibration process, which is 59.5 mol/m^2 (P117) and 88.4 mol/m^2 (P138).
The 68% error residuals range from 88.8 to 122.7 mol/m^2 for AMF groups 1-3 and 3-5, respectively.





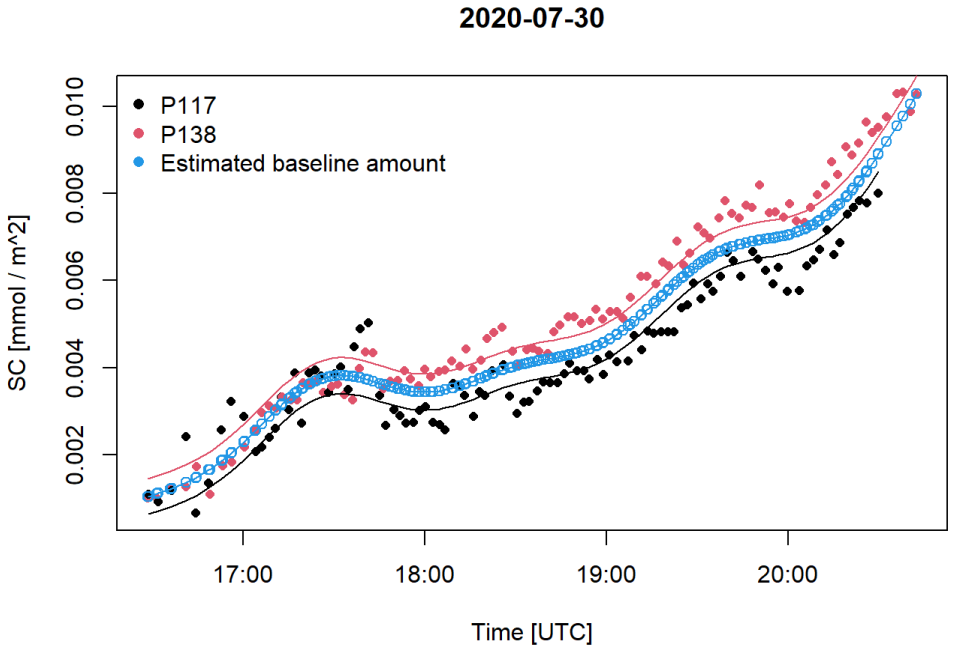

Intercepts [mol / m^2]
100
50
0
-50
-100
P117
P138

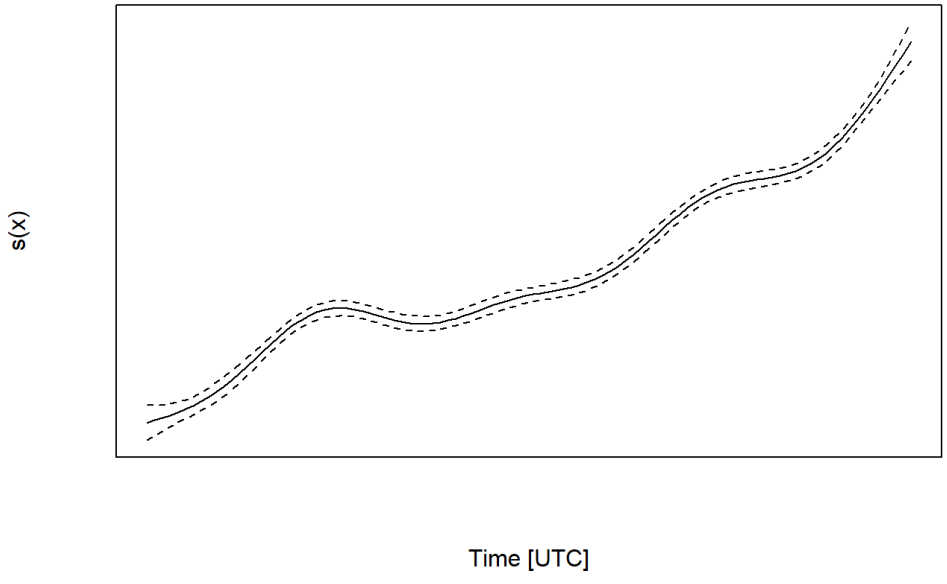

AMF [1-3)
Density
0.000
0.001
0.002
0.003
0.004
-400
0
200
400
Residual [mol/m^2]

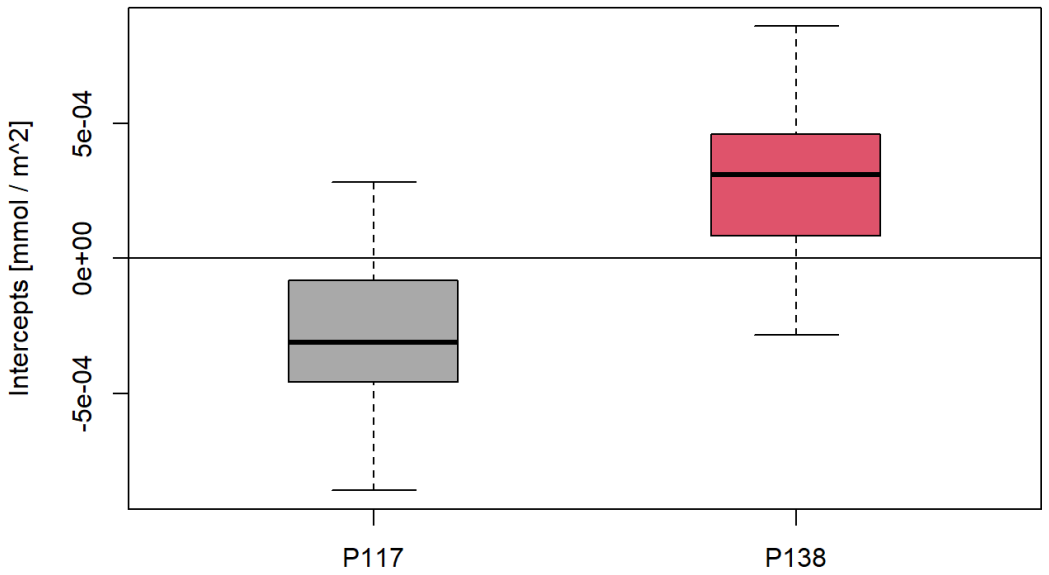

AMF [3-5)
Density
0.0000
0.0010
0.0020
0.0030
-400
0
200
400
Residual [mol/m^2]

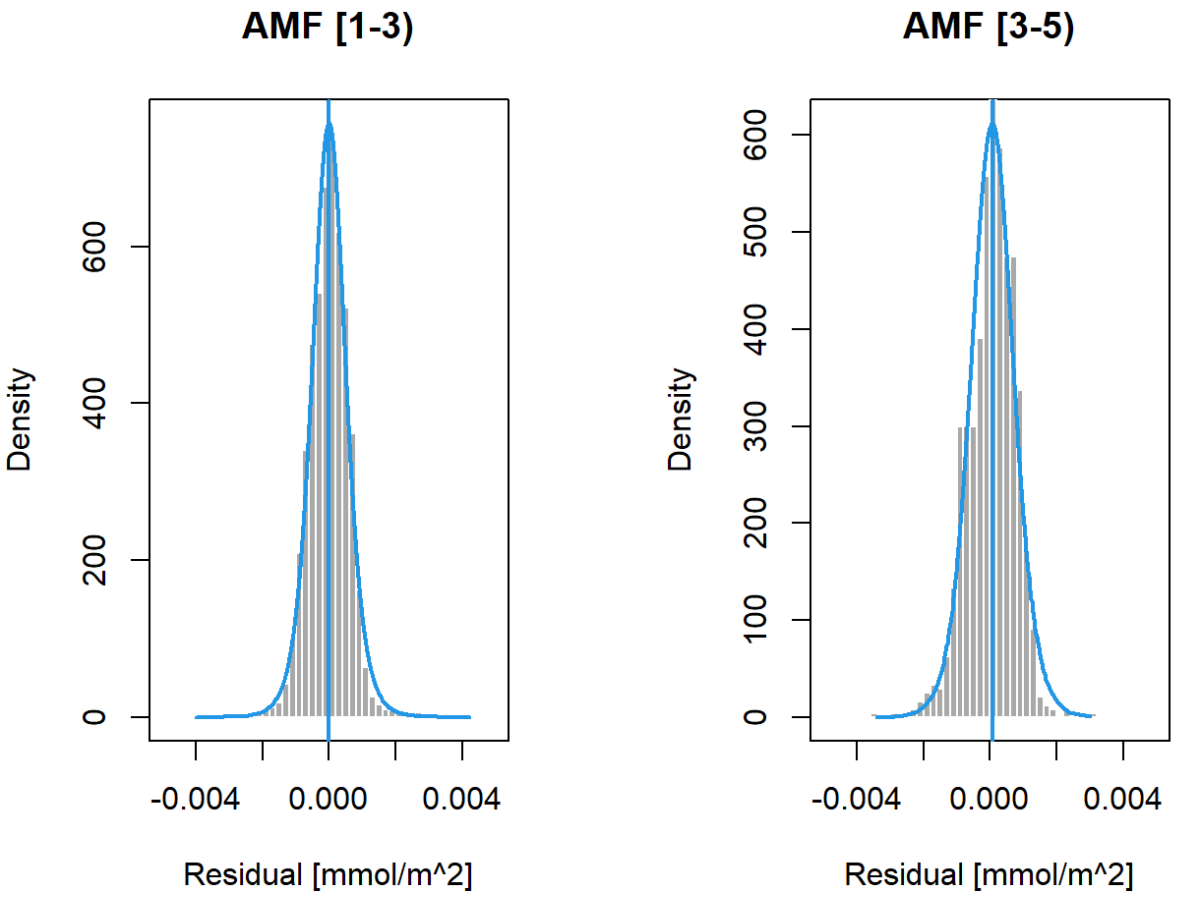
4.3 Lunar total column NO_3

The median difference in the intercepts between the two datasets is determined by 6.2e-4 mmol/m^2, where the reported common uncertainty during the calibration process is given by 3.1e-4 and 1.1e-4 mmol/m^2 for P117 and P138, respectively.
As illustrated in the example day, the two instruments show a varying difference during the day, which suggests already the presence of a distorting feature, which would be captured by the mentioned discrepancy uncertainty in the future. The resulting 68% residual distribution to the intercept-corrected baseline amount is given by 5.5e-4 and 6.6e-4 mmol/m^2 for P117 and P138, respectively.



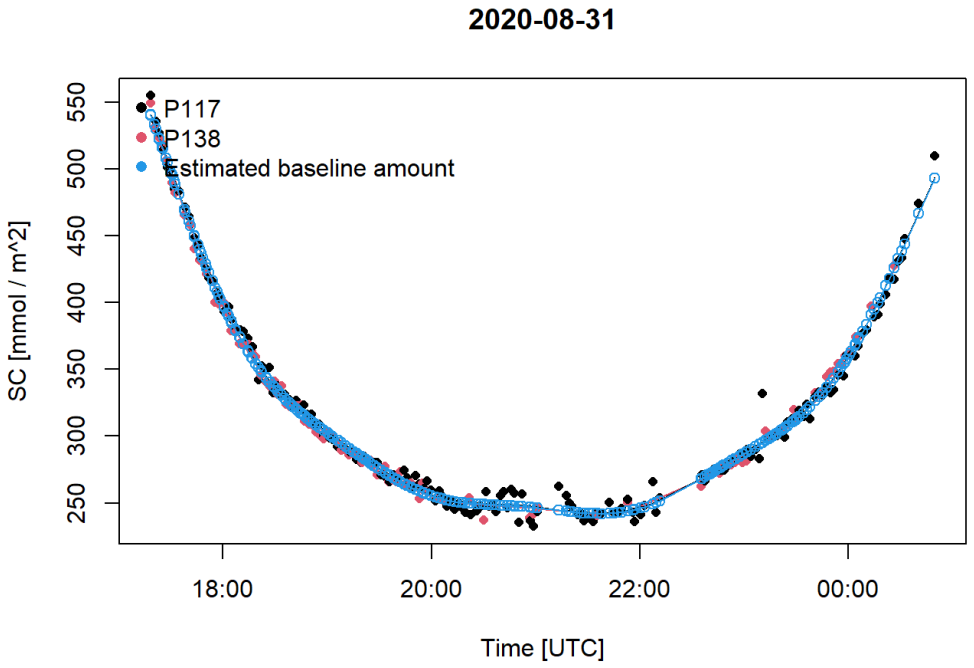


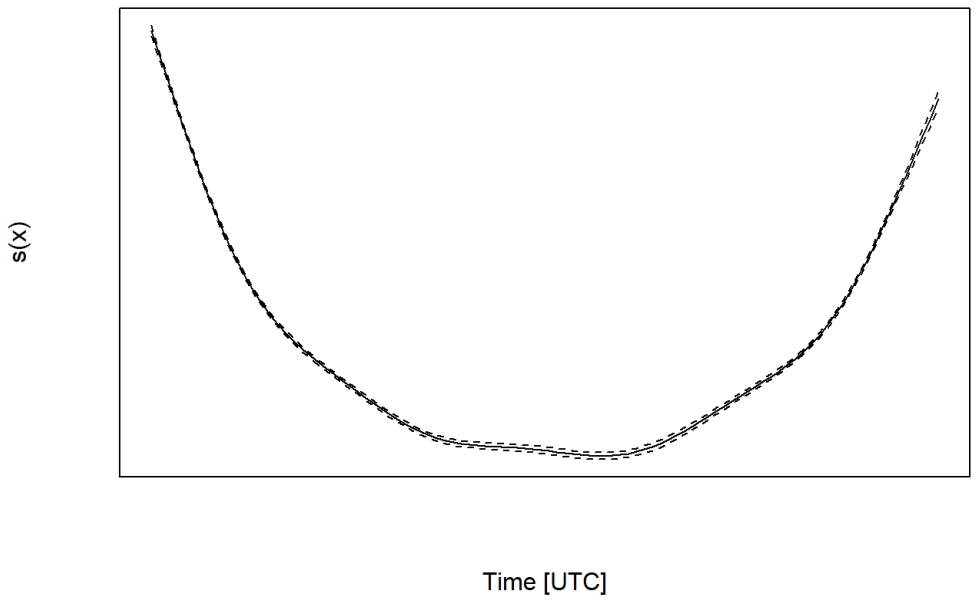

Intercepts [mmol / m^2]
-5e-04
0e+00
5e-04
P117
P138

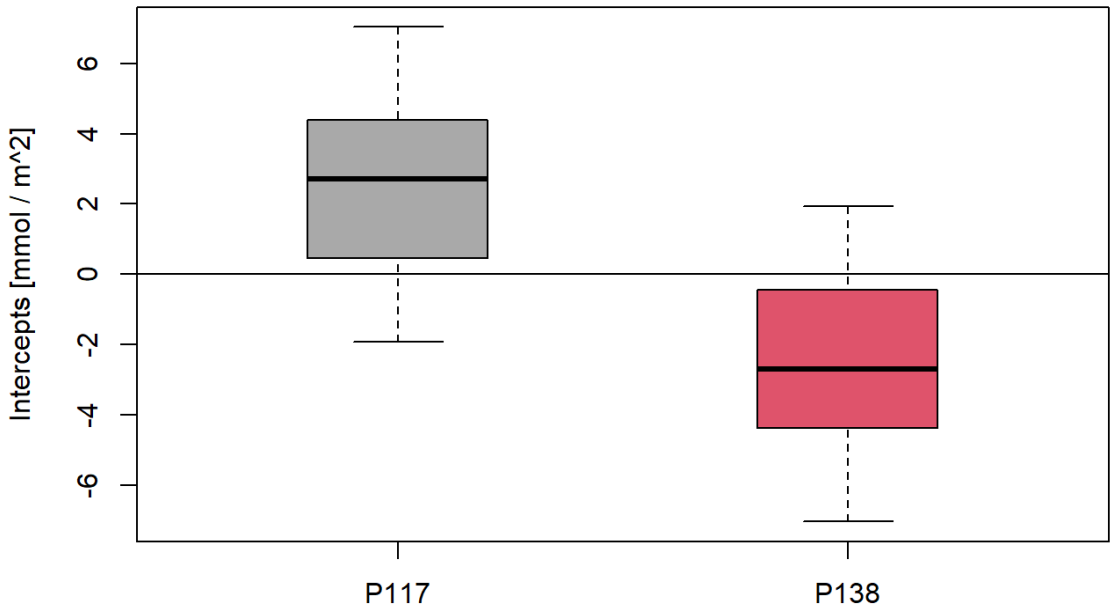

AMF [1-3]
Density
0
200
400
600
-0.004
0.000
0.004
Residual [mmol/m^2]
AMF [3-5]
Density
0
100
200
300
400
500
600
-0.004
0.000
0.004
Residual [mmol/m^2]

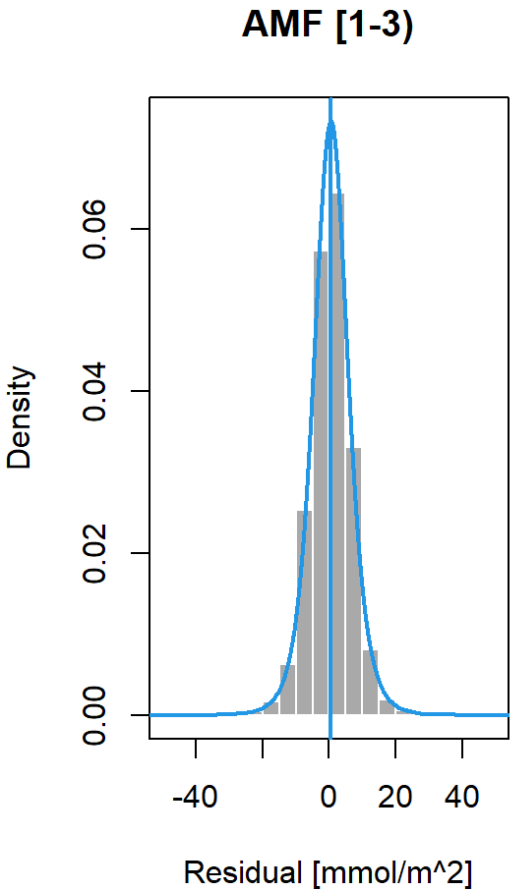
4.4 Lunar total column O_3

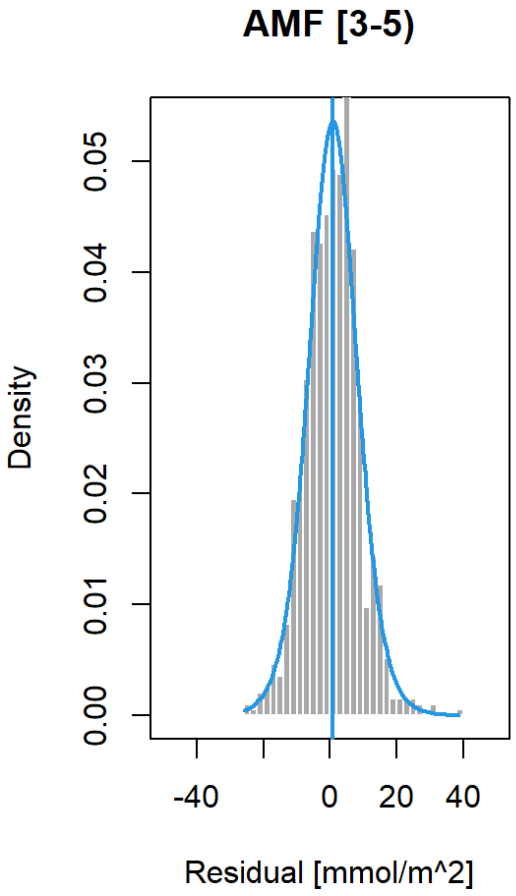
The two instruments report a systematic calibration difference of 5.4 mmol/m^2, with P117 being slightly higher than the baseline amount by 2.7 mmol/m^2. The corresponding calibration error during the calibration process has been determined by 1.3 and 0.7 mmol/m^2 for P117 and P138, respectively.
The resulting 68% residual distribution to the intercept-corrected baseline amount is determined by 5.672 to 7.585 mmol/m^2 for AMF groups 1-3 and 3-5, respectively.






Intercepts [mmol / m^2]
6
4
2
0
-2
-4
-6
P117
P138


AMF [1-3)
Density
0.00
0.02
0.04
0.06
-40
0
20
40
Residual [mmol/m^2]


AMF [3-5)
Density
0.00
0.01
0.02
0.03
0.04
0.05
-40
0
20
40
Residual [mmol/m^2]