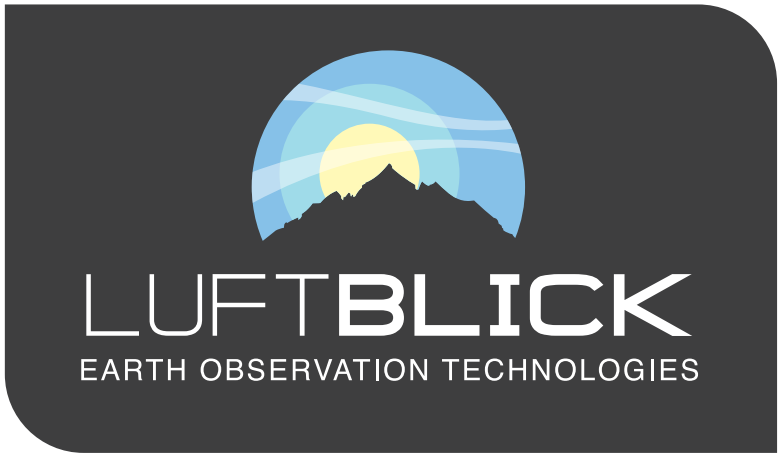

LuftBlick Report 2019008

Fiducial Reference Measurements for Air Quality

TN on Data Quality Flagging Generic Procedure Evolution

Version 7, 2022-12-31

	Name	Company
prepared by	Manuel Gebetsberger	LuftBlick
	Martin Tiefengraber	LuftBlick
	Alexander Cede	LuftBlick

Contents

Document Change Record

Issue	Date	Section	Observations
0.1	15th Dec 2019	All	First draft version
1	31th Dec 2019	All	Proof reading
2	27th Jun 2020	All	Moved procedures from testing to current. New testing procedures described in QP1/QP2. Major changes from Sec. 4.2 to 5. Proof reading
3	27th Dec 2020	2,4,5,6	Moving TVR+warning system (testing QP2) to current procedures.Proof reading / typographical corrections for conclusio subsections
4	20th June 2021	All	Outlook strategies QP3
5	31th December 2021	2,5,6	Testing, outlook strategies
6	28th June 2022	2,4,5,6	Atmophseric variability moved to current; new testing and outlook procedures
7	31th December 2022	2,4,5,6	Head sensor reading to current; Quality codes to testing;

Acronyms and Abbreviations

AV	Atmospheric Variability
NO_2	Nitrogen dioxide
O_3	Ozone
AMF	Air Mass Factor
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
DQ	Data Quality
FAR	False Alarm Rate
FRM4AQ	Fiducial Reference Measurements for Air Quality
GMM	Gaussian Mixture Regression Model
IQR	Interquartile Range
ML	Maximum Likelihood
P1	Probability to be in clear sky cluster
P2	Probability to be in overcast/cloud cluster
PDF	Probability Density Function
PGN	Pandonia Global Network
QA	Quality assurance
QC	Quality control
QI	Quality Indicator
QIT1	Quality Indicator Threshold 1
QIT2	Quality Indicator Threshold 2
QP1	QAQC pillar 1
QP2	QAQC pillar 2
QP3	QAQC pillar 3
RI	Representativeness Index
RSS	Residual Sum of Squares
TVR	Typical value range
VCU	Vertical column uncertainty based on measured uncertainty
wrms	Weighted root mean squared error based on measured uncertainty

1 Introduction

This report is deliverable D5 of the FRM4AQ project [1, 2]. It describes the quality assurance and quality control (QAQC) strategies for the PGN.

This document is structured as follows. Section 2 gives a summary of the current QAQC situation and highlights the main developments of methods under testing. Section 3 defines the PGN philosophy for QAQC, which is represented by the three QAQC pillars (QP1, QP2, QP3). Sections 4, 5, and 6 report the current, testing, and planned procedures for the individual pillars, where each pillar stage consists of a methodological and analysis part. This implies that procedures reported under the 'testing' section can be part of the 'current' section in the next version of the report. The same applies for methods in the 'planned' section, which gives an overview of untested methods and ideas, which would be part of the 'testing' section in a next report, and once the testing is done, such methods become part of the 'current' procedures.

1.1 Applicable Documents

[1] Fiducial Reference Measurements for Air Quality [Proposal], LuftBlick Proposal 201805DEV, Issue 1, 2018.

[2] Fiducial Reference Measurements for Air Quality [Contract and Statement of Work], ESA Contract No. 4000125841/18/I-NS, 2018.

[3] A. Cede. *Manual for Blick Software Suite Version 1.7*, 2019. URL `https://www.pandonia-global-network.org/wp-content/uploads/2019/11/BlickSoftwareSuite_Manual_v1-7.pdf`.

[4] A. Cede. *Manual for Blick Software Suite Version 1.8*, 2021. URL `https://www.pandonia-global-network.org/wp-content/uploads/2021/09/BlickSoftwareSuite_Manual_v1-8-4.pdf`.

[5] M. Tiefengraber, A. Cede, and M. Gebetsberger. Fiducial Reference Measurements for Air Quality, LuftBlick Report 2019005: New Algorithm and Product Development Plan, 2019.

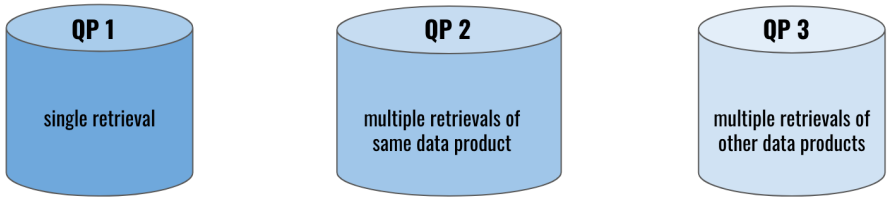

Figure 1: QAQC pillars

2 Summary

The current quality limits for quality indicators use a proper QF usable for end-users of L2 data, and provide detailed information about different sources of uncertainty. On L1 basis, the atmospheric variability is used as a decision basis for picking reference days during the calibration process. Each QI can be monitored and changes are detected and highlighted by a statistical timeseries analysis in combination with a warning system. Moreover, head sensor readings serve as a proxy about the sealing performance.

As part of testing procedures, daily shapes and AMF dependencies of parameters can highlight a potential calibration error, while a second approach employs quality codes using an extraterrestrial reference, and sets the focus on spectral residuals.

As part of QP3, two new quality indicators are under development in order to decrease the time spent on QC tasks, and to increase the reaction time if instruments run out of calibration. Both approaches use the output of gas retrievals using the synthetic reference from different wavelength regions.

3 QAQC pillars

The PGN strategy for QAQC is defined by three pillars as illustrated in Figure 1 and described in detail in the following subsections. An estimation about the readiness of each pillar can be made by 95% (QP1), 70% (QP2), 10% (QP3).

3.1 QP1

This pillar is based on one measured spectra of one data product, where the data quality (DQ) is assessed on the values of quality indicators (QI) which are determined during data processing from L0 data (raw measured counts) to the final L2 product (e.g. NO_2 direct sun total column). For instance, if the retrieved slant column uncertainty of NO_2 from the spectral fitting is used as a QI, and exceeds its QI threshold 1 (QIT1), it raises the data quality from high (DQ0) to medium data quality (DQ1). If it exceeds QI threshold 2 (QIT2), it raises the data quality from DQ1 to data quality low (DQ2).

All available QI and its QIT1 and QIT2 are given in the BlickProcessingSetups file under the so-called quality codes (q-codes). Since the processing goes from L0-L1, L1-L2Fit, L2Fit-L2, each processing level has its own quality codes for the individual QI's:

- L0 to L1: qs-codes
- L1 to L2Fit: qf-codes
- L2Fit to L2: qr-codes

The QF is based on these QIT's, and if the QIT1 of already one QI is exceeded, it raises DQ from high to medium. This implies that if an L1 error triggers DQ1, L2Fit and L2 processing could obtain QI's to be of DQ0, but L2Fit data and L2 data are overruled by this L1 error and can have at best DQ1. A detailed description of the different q-code tables is given in *Cede* [3].

3.2 QP2

This pillar uses multiple retrievals of the same data product. E.g., the retrieved NO_2 slant column uncertainty and the wavelength shift is used as QI for 3000 spectra measured over a month, rather than only 1 measured spectra. Herein, each QI is given as a timeseries with its own QIT1 and QIT2. The role of QP2 is more on QC as timeseries of QI's are used to control its stability over time to detect any instrumental changes or if an instrument runs out of calibration. Moreover, QP2 is used to characterize typical values of QI's under clearsky or cloudy conditions.

3.3 QP3

This pillar uses multiple retrievals of different data products, e.g. 3 QIs of direct sun NO_2 (vertical column uncertainty, wavelength shift, weighted rms), and 1 QI of direct sun O_3 (weighted rms). Similar to QP2, the role of QP3 is also towards QC to detect changes in calibration or an instruments' operation.

4 Current QAQC procedures

This section describes the current QAQC procedures for pillars in use, where each explains subsection the used methods, followed by an analysis. The end of this section gives concluding remarks.

4.1 QP1

4.1.1 Quality Indicators - Thresholds

Among the list of available QIs, which are written in each processed level file, the current procedures for QC in particular focus on following QI's:

a Weighted root mean squared error based on measured uncertainty (wrms): QF is evaluated at the processing level L2Fit.

b Vertical column uncertainty based on measured uncertainty (VCU): QF is evaluated at the processing level L2.

c Wavelength shift: QF is evaluated at the processing level L2Fit

d Integration time: no QIT defined.

e Mean of measured counts inside the fitting window: no QIT defined

Parameters a-b are the main drivers leading to medium/low quality data, and are based on average PGN instrument characteristics. Its respective QIT is determined via a Gaussian mixture regressison model (GMM), which is described in the following subsection.

Parameters c-e are of less importance regarding their absolute value, but are used for monitoring an instruments sensitivity and operativity in terms of sudden relative changes or drifts as part of QP2.

4.1.2 QIT determination based on Gaussian Mixture regression models

This model approach makes use of multiple retrievals of QIs to objectively derive its thresholds for QF based on an unsupervised learning approach. Unsupervised in this context means that there is no observed variable available, which could be used to train the statistical model.

As a starting point, we use a time series of a QI for a period, where the calibrated instrument was measuring without hardware issues. Consequently, differences in the value of a certain QI would result from atmospheric differences between clear sky and overcast. E.g., clear sky days are supposed to have small integration times, small vertical column uncertainty and small wrms. Clouds require the instrument to use smaller attenuation filters or to increase the integration time in order to have the same counts as if there where no clouds. However, larger integration times lead to larger uncertainties, which also affect the wrms.

Based on this assumption there are two underlying regimes which show distinct QI characteristics for clear sky and overcast/cloudy conditions. Figure 2 illustrates a typical QI change from clear-sky (first half of first day) to cloudy conditions (second 2), and further to conditions with thick clouds (third day). By knowing these distinct clusters, one can derive probabilities for being within the clear sky or overcast cluster. In between, there is a transition period which between the two extremes which would correspond to a cluster with thin clouds.

An example of such wrms clusters is shown in Figure 3 for two Pandoras, where two peaks are visible for both Pandoras, and each peak can be assigned to the assumed clusters for clear sky and overcast conditions, respectively. In order to obtain the conditional distribution as illustrated in Figure 3, the Gaussian mixture regression model (GMM) is applied.

Equation 1 defines the conditional probability density function (PDF) h, which is expressed as the sum over K individual Gaussian PDF's $f(y|x, \theta_k)$, multiplied by their prior probabilities π_k. In our case, the number of clusters K equals two, since we assume 2 regimes. y defines the response variable, which is the QI (e.g., wrms), and x a vector of independent variables. θ_k defines the Gaussian distribution parameters μ and σ, and ψ the vector of all parameters $(\pi_1, ..., \pi_K, \theta'_1, ..., \theta'_K)'$.

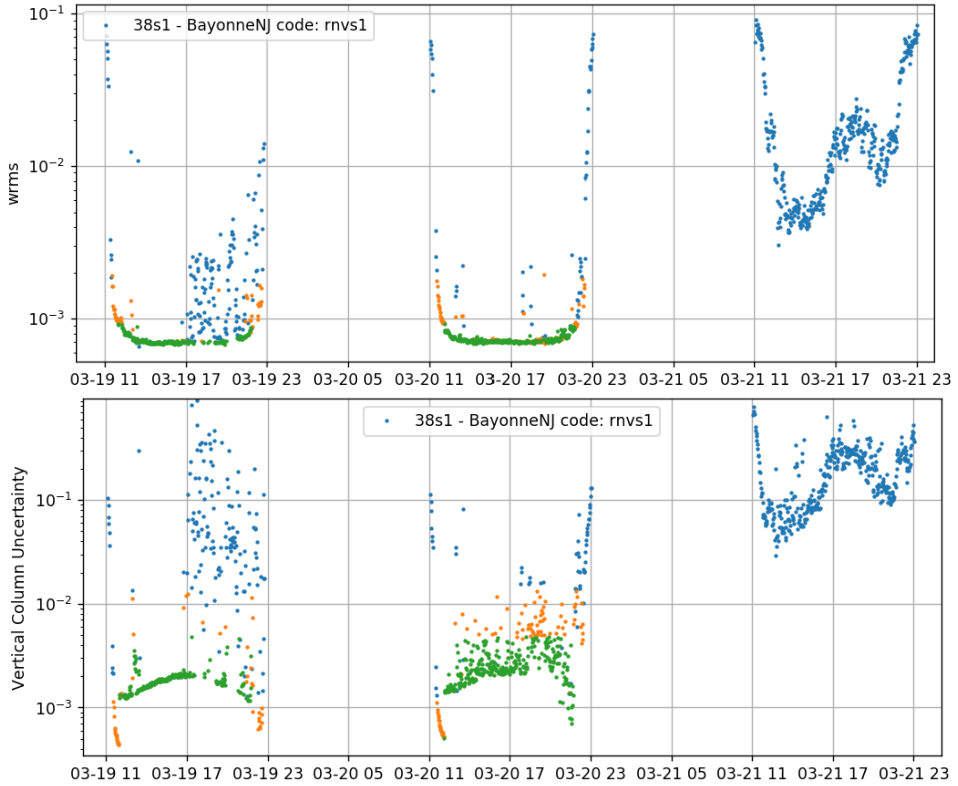

Figure 2: wrms (top figure) and VCU (bottom figure) of NO2 for Pandora38 (Bayonne, New Jersey), showing the quality flagging for DQ0 (green), DQ1 (orange), DQ2 (blue) data.

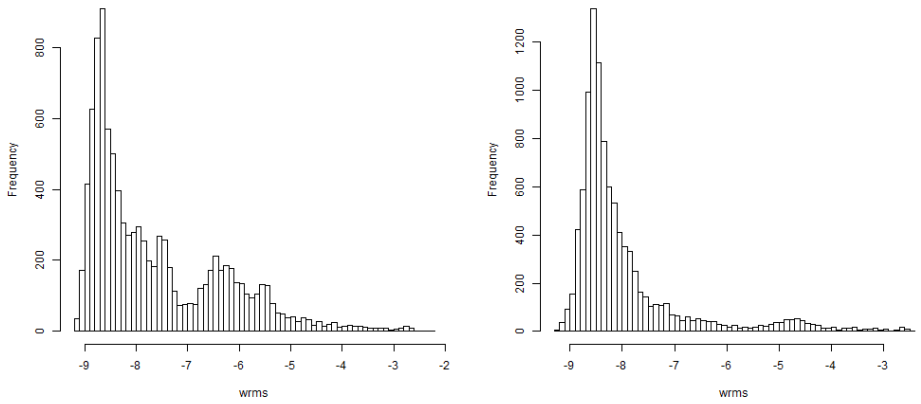

Figure 3: Logarithmic value of wrms for Pandora65 (Altzomoni, Mexico) and Pandora128 (Alice-Springs, Australia) for a subsample of 10 000 data points

$$h(y|x,\psi) = \sum_{k=1}^{K} \pi_k \cdot f(y|x,\theta_k) \quad (1)$$

The priors can be seen as a climatological value of occurrences for the regimes assumed, and expected to differ between different sites. Parameter estimation is based on maximum likelihood maximization (ML), using an iterative EM algorithm, as implemented in the R-package flexmix.

After estimation, the parameters are used to derive probabilities of being in cluster 'good' (clear sky) or 'bad' (overcast) by using the individual cumulative distribution functions (CDF). Since data points in the 'good' cluster are expected to have smallest values for certain QI, 1-CDF is taken as the probability for the marginal 'good' probability (p1), and the CDF for the marginal 'bad' probability (p2). The probability to be in the 'good' (P1) cluster is then derived by $P1 = \frac{p1}{p1+p2}$. Conversely, the probability to be in the 'bad' cluster P2 is given by $P2 = 1 - P1$.

4.1.3 QIT analysis

Gaussian mixture regression, QI and dataset The GMM previously described has been tested for official and non-official PGN instruments to derive generic

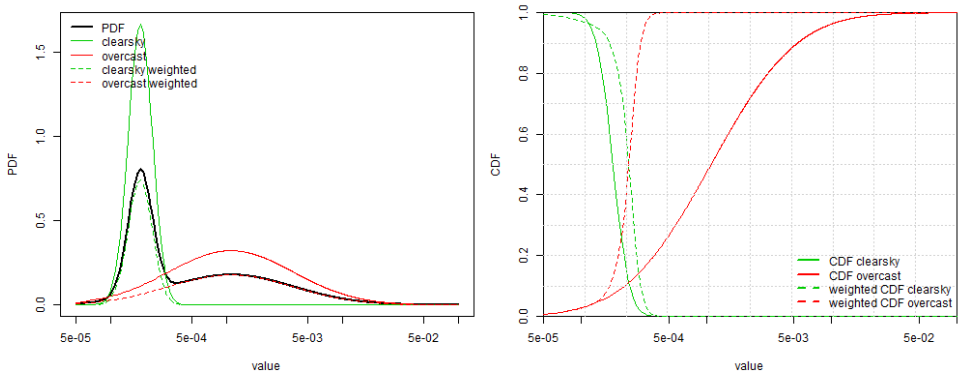

Figure 4: GMM result for wrms at Altzomoni, Pandora65. Left Figure shows the PDF, and right figure the CDF. Red lines illustrate the PDF/CDF for the 'overcast' cluster, green lines PDF/CDF for the 'clearsky' cluster. Black lines shows the PDF of the overall PDF. Dashed lines illustrate the individual PDF's weighted with π, and weighted CDF's P1 and P2.

thresholds for two parameters based on the official NO_2 L2 product for direct sun (nvs0): wrms, VCU.

In total 19 sites with measurements up to December 2019 have been tested, where only quality assured periods have been used. Furthermore, a subset of randomly selected 10000 retrievals is used for parameter estimation to decrease computational time. The GMM is specified without including any independent variables (x); neither for the Gaussian expectation value, nor for π. Model estimation is done using the R-package 'flexmix'.

GMM Altzomoni A typcial result of the GMM is shown in Figure 4 for the wrms GMM. The clearsky cluster (solid green) shows a very small variance of 0.23 on a linear scale, compared to the overcast cluster (solid red) variance of 1.23. The expectation value of clearsky wrms is 0.00018, and for the overcast wrms 0.001. This very low number is explained by the remote site of Altzomoni which typically shows smaller wrms values than urban sites.

GMM for multiple instruments and new thresholds Based on the obtained probabilities of being in the clearsky cluster, as exemplarily shown in the previous paragraph, the GMM results vary from site to site, as illustrated in the boxplot

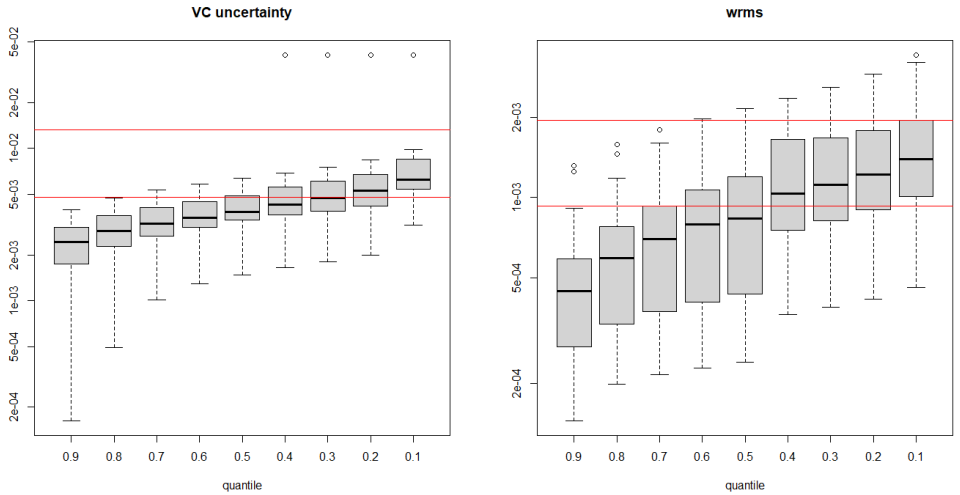

Figure 5: GMM results for 19 instruments for QI VCU (left) and wrms (right). x-axis shows the weighted clearsky quantiles P1. Red lines indicate QIT1 and QIT2. Each boxplot contains of 19 datapoints.

of Figure 5. These probabilities allow to obtain generic PGN limits for QIT1 and QIT2 as part of QP1 for wrms and VCU. For wrms, we use the 0.7 and 0.1 quantile, and take the 0.75 percentile value over the 19 instruments. For VCU, we use the 0.7 and 0.2 quantile, and take the 0.95 percentile value over the 19 instruments. Those new thresholds are summarized in Table 1 for the new retrieval code nvs1 for NO_2.

Table 1: Changed quality limits for wrms and VCU

Parameter/limit	nvs0	nvs1
wrms / QIT1	2e-3	9.3e-4
wrms / QIT2	5e-3	1.95e-3
VCU / QIT1	3e-2	4.8e-3
VCU / QIT2	5e-2	1.33e-2

Impact of new QF thresholds The formerly operational direct sun NO_2 product (nvs0) has been replaced by nvs1 in January 2020, which only differs in the limits

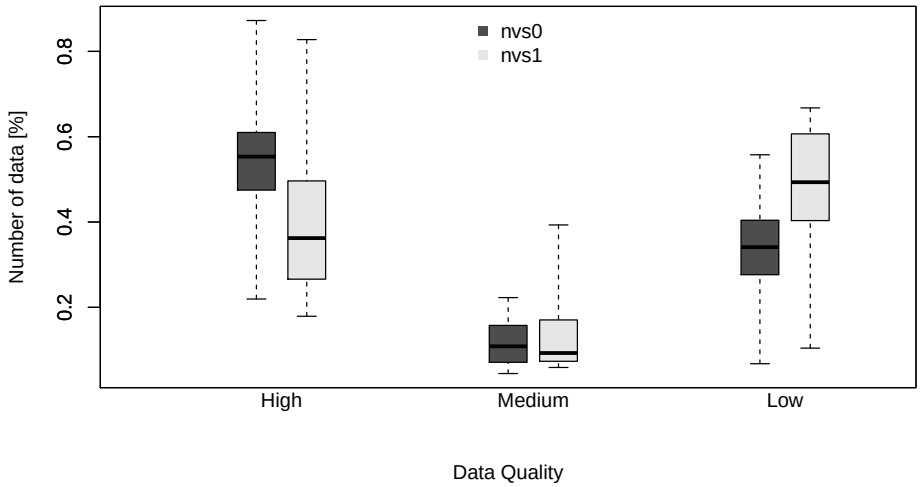

Figure 6: Data quality (high, medium, low) for the PGN , evaluated for each individual site seperately using nvs0 (dark grey) and nvs1 (light grey). Whiskers extend to the lowest/highest value. The lower hinge corresponds to the 25% percentile, the upper hinge to the 75% percentile, and the horizontal black line illustrates the median.

for quality flagging. Regarding traceability, this version increase was necessary since the quality limits of two retrieval parameters changed from nvs0 to nvs1.

The quality limits used in nvs1 cause a stricter filtering to ensure that only high quality data end up in the high quality cluster. Figure 6 summarizes the percentage of data to be of high, medium, or low data quality for the selected sites for both nvs0 and nvs1. Due to the stricter filtering, there is a clear drop in the median for high quality data. For nvs0, 50% of the evaluated sites have more than 55% of high quality data. For nvs1, 50% of the evaluated sites have more than 36% of high quality data.

This causes that a fraction of 'old' (nvs0) high quality data are now of medium quality, and/or already of low quality. This pattern ist strongly station dependent where the reduction in the amount of high quality is smaller for remote sites (e.g. Izana) as for urban sites (e.g. MexicoCity). Figure 7 shows the transitions of data

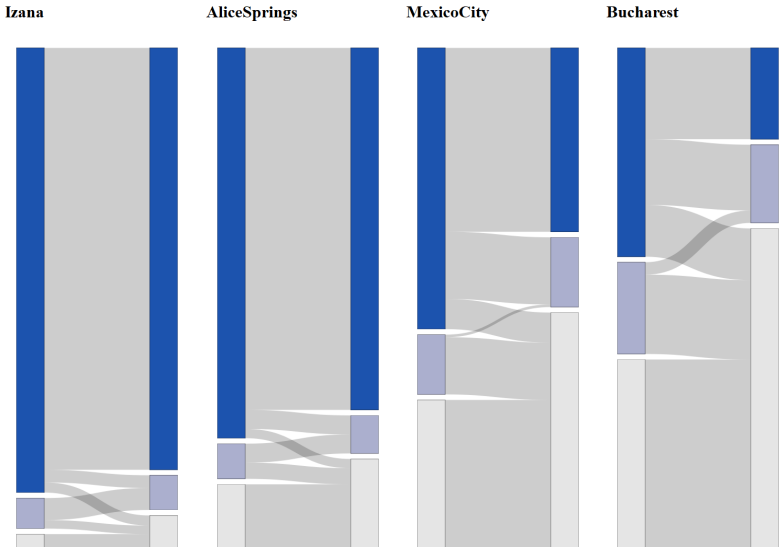

Figure 7: Transitions of old quality flagged data (left, nvs0) to new quality flagged data (right, nvs1). Data quality is shown as high (dark blue), medium (light blue), and low (grey) for Izana (Pandora 101), AliceSprings (Pandora128), MexicoCity (Pandora142), and Bucharest (Pandora111).

quality for data points from nvs0 to nvs1 for 4 selected sites.

The stricter filtering is important since we observed negative vertical columns which should not be of DQ0. Figure 8 shows the vertical column for AliceSprings comparing nvs0 QF (blue) and nvs1 using the GMM obtained thresholds (orange) between Januarry 2018 and November 2019. Clearly, the new limits decrease the negative outliers which are typically caused by wrong pointing or clouds which further lead to a wrong air mass factor calculation for direct sun retrievals. However, there are still outliers left which can be also filtered by stronger wrms or VCU limits for DQ0. This would be the case if the individual GMM result of AliceSprings is applied instead of the PGN limits defined in Table 1.

4.1.4 Atmospheric Variability Parameter

Each Pandora undergoes a comprehensive laboratory calibration in order to characterize and correct for instrumental features. This laboratory calibration is performed under ideal and stable conditions which allows to estimate an instrumental uncertainty σ_{instr}. A detailed description can be found in *Cede* [3], Chapter 6.2.

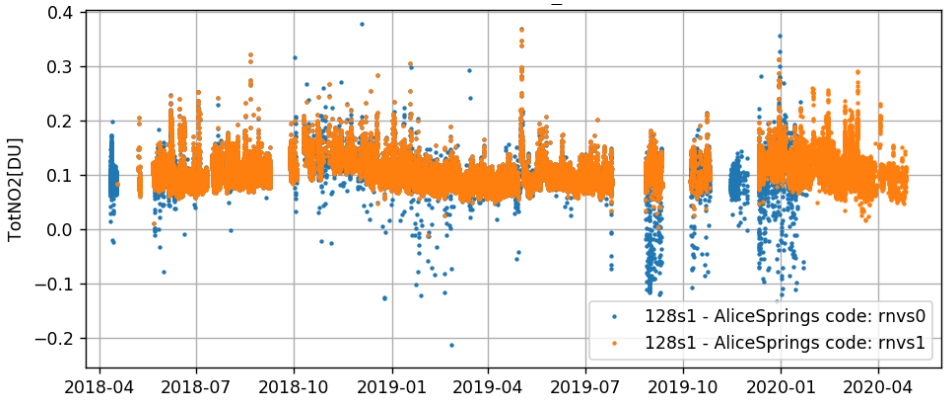

Figure 8: DQ0 data of NO_2 total column amount in DU for AliceSprings from 2018 to mid April 2020: nvs0 (blue) and nvs1 (orange).

Contrary, the measured uncertainty, σ_{meas} which is basically the standard deviation during a field measurement, consists of σ_{instr} and the atmospheric uncertainty σ_{atmos}. σ_{meas} increases for instance if clouds move in the FOV during a measurement period. Figure 2 showed already the QI's wrms and VCU, where the first half March 19th is characterized by almost clear-sky conditions, and the second half by cloudy conditions. The cloudy conditions increased both QI's and led to DQ1 and DQ2 data. It seems that the VCU, which is based on σ_{meas}, is more sensitive to changing conditions due to σ_{atmos}. This can be seen in the first half of March 19th where VCU illustrates a peak up to 2e-1. However, QIT1 for VCU is not perfectly filtering this peak.

Therefore, we suggest a new QI called 'atmospheric variability' (AV, Eq. 2), which accounts for σ_{atmos} and offers a quantitative filtering already on L1 basis. Although we cannot measure σ_{atmos} directly, we can put σ_{instr} and σ_{meas} into a ratio, where values larger than 0 are associated with σ_{atmos}.

$$AV[\%] = (1 - (\frac{\sigma_{instr}}{\sigma_{meas}})^2) \cdot 100 \quad (2)$$

In the limit, AV goes towards zero if σ_{meas} goes towards σ_{instr}. Top graphic of Figure 9 illustrates σ_{meas} and σ_{instr},where σ_{meas} is always slightly larger than σ_{instr}. Applying Eq. 2 leads to the values shown in the middle graphic, where a baseline

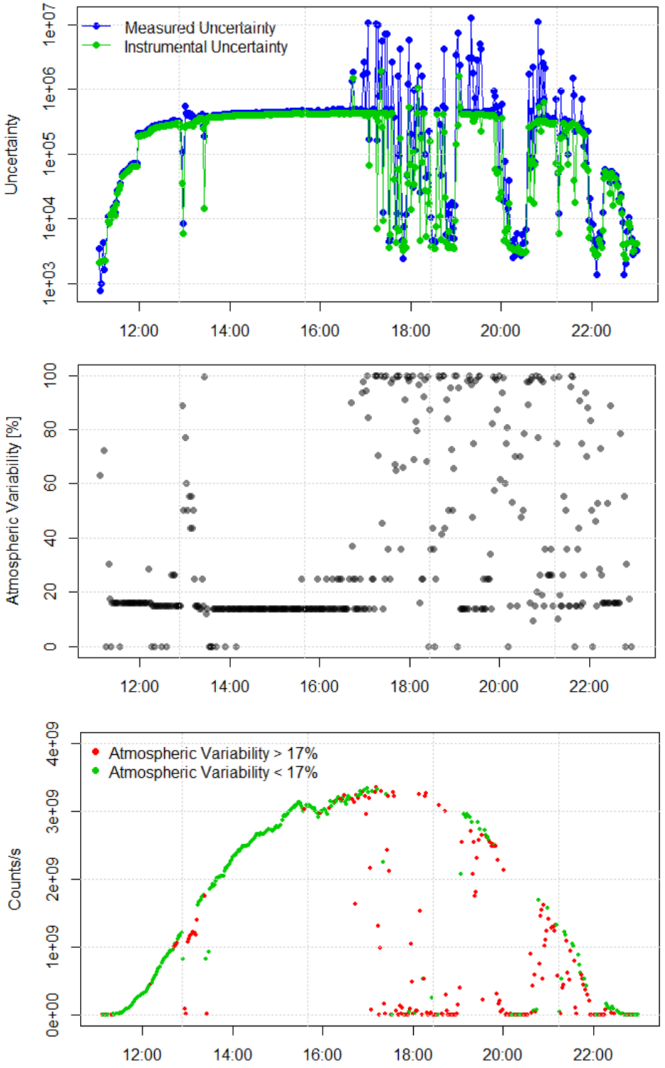

Figure 9: L1 data of Pixel 800 for P38 at Bayonne, New Jersey on March 19th 2019, where x-axis denotes time in [UTC]. Top: measured and instrumental uncertainty. Middle: atmospheric variability. Bottom: L1 corrected count rate.

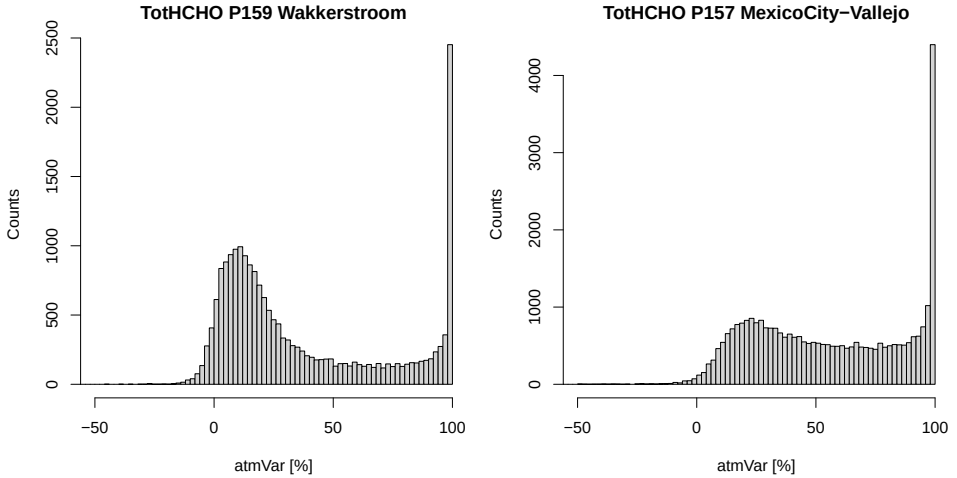

Figure 10: AV for two HCHO timeseries using the 1.8 processor: clean remote (left graphic), polluted urban (right graphic).

AV of 14-18% is visible. If a threshold of 17% AV is used to filter L1 data, the cloudy afternoon period, and in particular the peaks before noon are properly taken into account.

While σ_{instr},σ_{meas} show a strong daily cycle and jumps in the magnitude due to the input signal, AV is more robust against changes in the integration time, filter-wheel setting, and shows a more stable baseline over the day. Moreover, AV would provide a useful QI not only to improve the general QF up to L2 data, but would also deal as a decision basis for an end-user who only works with L1 data.

Examples and QIT determination AV is implemented in the 1.8 processor version and part of the L2 output. Therefore,it is also used as for QF where thresholds for QIT1 and QIT2 must be evaluated. Similary to wrms, we make use of the GMM as described in Section 4.1.2, with the assumption of having a 'good' and 'bad'- day clustering. A characteristic AV is illustrated in Figure 10 for a remote site (left), and a urban polluted site (right). Clearly, there is a peak at lower AV values which is found at higher AV value for the polluted site than for the remote site. However, both locations illustrate a very intense peak on the very right end, exactly at 100%, which defines the upper boundary of the AV.

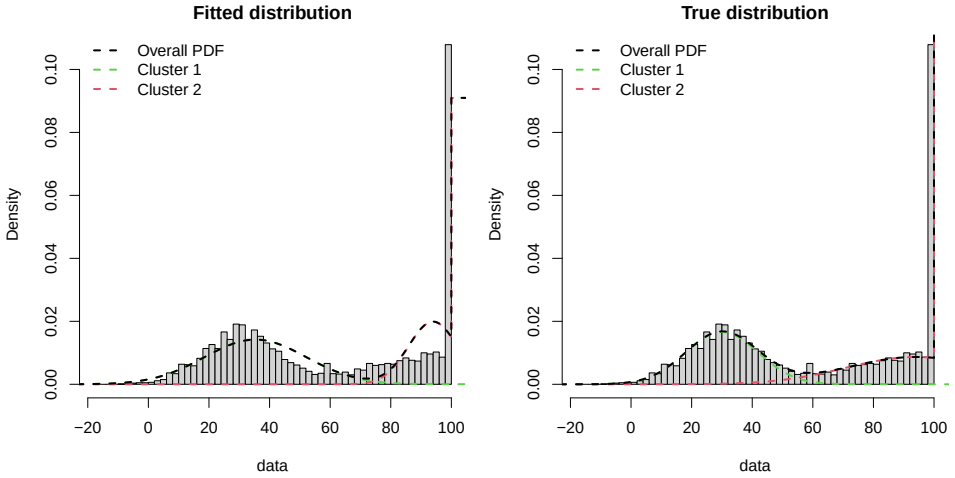

Figure 11: Simulated data of AV with a fitted GMM (left graphic), and the censored Gaussian approach (right graphic).

This unusual point mass at 100% can cause a bias in the estimated GMM fits for the two clusters. A simulated example is illustrating this in Figure 11, showing the overall PDF of the obtained GMM. The two peaks are not ideally fitted, in particular the right peak at 100%, since the GMM wants to obtain a large value for logarithmic densities of the two Gaussian distributions. As a consequence, the obtained standard deviation is smaller than the truth would be, with a biased expectation value towards lower values. However, the standard GMM gives already a good first guess QIT1 and QIT2.

Nevertheless, an appropriate distribution for these kind a data is suggested by the censored Gaussian distribution, which could properly account for the point mass at 100%. Hereby, we assume a latent Gaussian process $y^* \sim N(\mu, \sigma)$. All values $y^* \geq \tau$, where τ defines the censoring level of 100% in our case, are unknown and therefore censored to τ. The likelihood which needs to be maximized, would be splitted into an uncensored and censored part .

This approch is currently being implemented and tested for various sites to improve the GMM approach for QIT determination of AV.

4.1.5 L2 Uncertainty Information

With the processor version 1.8, there are more uncertainty information given in the output of the L2 data. Blick 1.7 only provided the measured vertical column uncertainty in the output format.

Blick 1.8 data products report different sources of uncertainty [4]:

- **Common uncertainty**: Fully correlated to other variables. E.g., a calibration error in the reference slant column amount
- **Structured uncertainty**: Partially correlated to other variables. E.g., the uncertainty of straylight correction for different pixels.
- **Independent uncertainty**: Uncorrelated to other variables. E.g., read noise or photon noise in a pixel.

This granular information can give insight about the different sources of errors, and could also lead to a future data filtering by setting dedicated thresholds. At the moment, these uncertainties are not used in a quantitative way by any QP.

4.2 QP2

4.2.1 Daily Aggregates and Air Mass Factor Binning

Each Pandora has it's own characteristic value ranges, and day to day variations due to weather. But based on the thresholds for QP1, the ranges are expected to be at a certain value level for clear sky or cloudy days. This also applies, if an instrument has obstacles in the FOV (e.g., water, bees, spiders, buildings, towers), or is not pointing on the sun properly.

Instead of evaluating each individual measurement, daily aggregates using the 10%, 50% (median), 90% percentile of a QI's value range, condense the available information of multiple measurements and provide a quick overview about an instruments daily performance. However, each day typcially consists of DQ0, DQ1, DQ3 data, which also have their own typcial value range, where a median over the whole day is not representing a characteristic aggregation value. Therefore, the aggregation is performed only on the best available DQ. This means that if there is a cloudy day, where no high quality data (DQ0) can be obtained, but at least three DQ1 data are available, the aggregation is done on the DQ1 data.

To overcome possible air mass factor (AMF) dependecies of QI's, such aggregates are calculated for 3 AMF bins: [1,3), [3,5), [5,7)

Case study Figure 12 illustrates the aggregates of three QI's for a time window of approximately three months, and binned for AMF groups [1-3) and [3-5). All three parameters show a typical range which is almost constant over time for the integration time and the wrms. The counts in the fitting window show an increase towards spring. The AMF binning highlights the different value level, in particular for the wrms.

However, in mid of March 2020 a sudden change occurs in the aggregates for all QI's, affecting both AMF groups in a way that no DQ0 data are available. This sudden change corresponded to a damaged tracker, which led to an inproper pointing.

4.2.2 Breakpoint Analysis

Breakpoint analysis is a method to detect jumps or structural changes in a time series of QI's. The concept is a linear regression model, where the approach searches for m breakpoints in a time series where the regression coefficients change from one stable regression to another one.

As implemented in the R-package 'strucchange', the approach searches for an undefined number of breakpoints by minimizing the residual sum of squares (RSS) and accounts for the number of parameters to be estimated via the Bayesian information criterion (BIC).

Clearly, the more breakpoints are used, the smaller the RSS becomes. Hence, those m breakpoints where the BIC is not improving anymore are returned to identify changes in the timeseries.

Case study The breakpoint analysis allows any kind of statistical model assumption, where the simplest one is a constant value over time. This is also the assumption being tested for different L2 QI's of direct sun NO_2. An example is illustrated in Figure 13 which shows the time series of aggregated daily median integration times for measurements between AMF values [1,3) for Boulder in 2019, illustrating the logarithmic integration time on the y-axis. A clear baseline of 2.5 is visible until end of November 2019. The breakpoint analsis returns November 22 as illus-

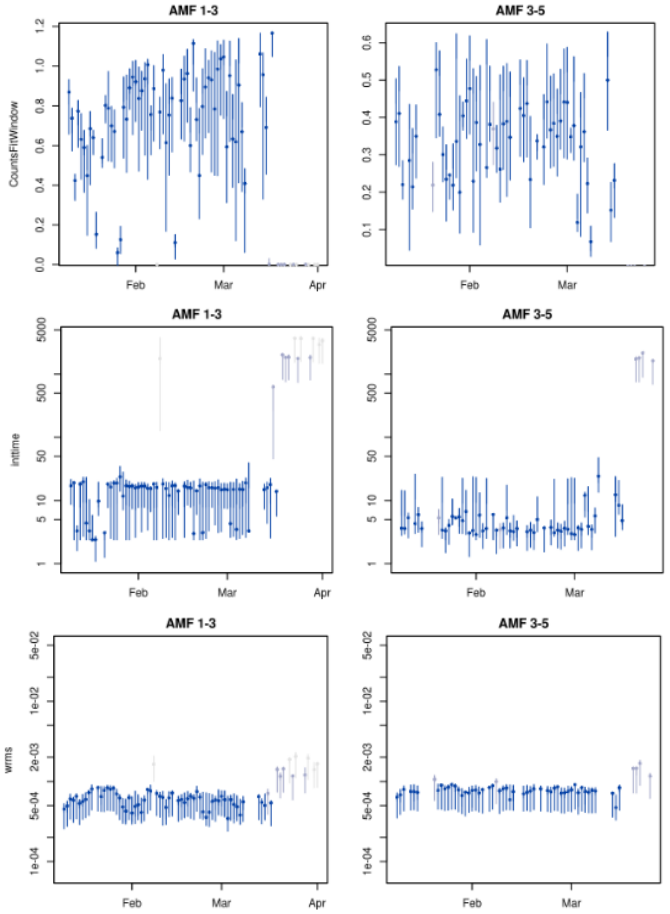

Figure 12: Aggregates of NO2 at P119 (Athens-NOA) in 2020 for counts inside the fitting window $[W/m^2/nm]$ (top panel), integration time $[ms]$ (middle panel), and wrms (bottom panel). Dots represent the daily median value, and vertical bars the range between the 10% and 90% percentile. Color coding refers to the lowest (best) data quality for each day: DQ0 (dark blue), DQ1 (light blue), DQ2 (grey).

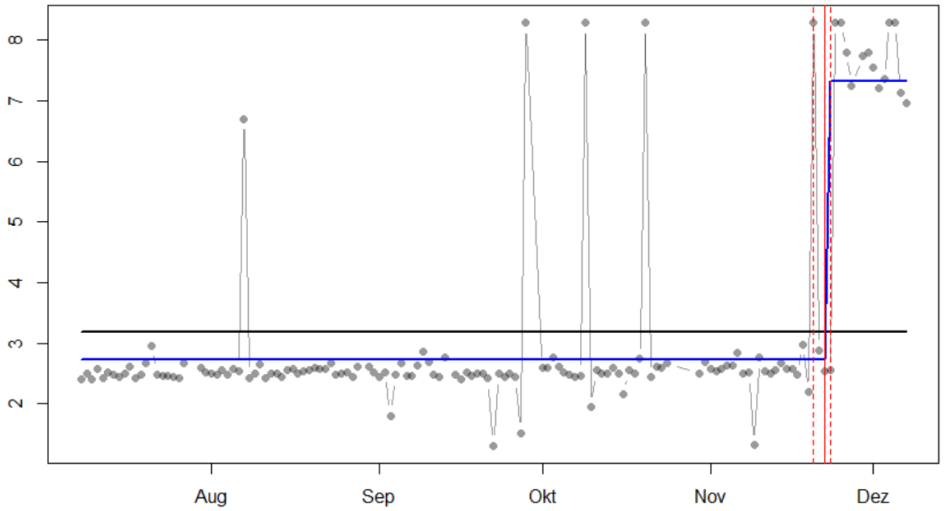

Figure 13: Timeseries of daily aggregates of the logarithmic integration time in ms for Pandora57 in Boulder: daily median integration time (grey dots), constant linear model (black solid line), breakpoint (vertical solid red line), breakpoint 90% confidence interval (vertical dashed red line), breakpoint models (solid blue lines.)

trated by the red vertical bars when a structural change in the timeseries occured. Approximately at that time, Boulder had bad weather conditions with a heavy storm and the instrument lost the alignment, which could successfully be aligned again.

A second example for different QI's is shown in Figure 14 for P119 at Athens, using the operational 30-day-window. Similar to the Boulder example, a tracker related issue is visible in all three QI's and both AMF groups, where the breakpoint analysis properly returns the same dates where the structural change occured.

This statistical method proves to be a powerful tool for routine QC to detect jumps in the QI characteristic of an instrument. The examples for Boulder and Athens are ideal cases, where the tool warns PGN operators that an instrument action might be needed. However, it is also possible that that there is one week of overcast conditions which would also trigger a breakpoint. Such false alarms should be avoided in principle, but this false alarm rate (FAR) has not been quantified so far in order to optimize the algorithm. The sensitivity to detect a sudden change is currently implemented for 5 days, which means the baseline must change for at least 5 days. A smaller sensitivity would raise too much false alarms for locations

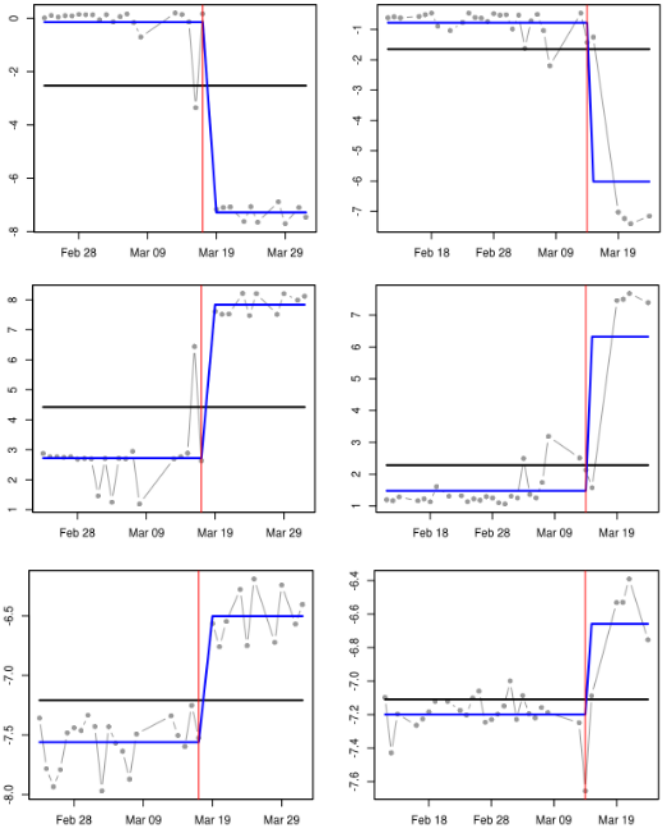

Figure 14: Breakpoint analysis for aggregates of NO2 at P119 (Athens-NOA) for a 30 days window 2020. Shown are the QI: counts inside the fitting window $[W/m^2/nm]$ (top panel), logarithmic integration time $[ms]$ (middle panel), and logarithmic wrms (bottom panel). Each plot illustrates the daily median (grey dots), constant linear model (black solid line), breakpoint (vertical solid red line), and the breakpoint models (solid blue lines.)

with variable weather conditions. However, in order to maintain an instruments operativity, a higher FAR is better than critical events are missed.

4.2.3 Typical Value Range Determination - Automated Warning System

The concept of daily aggregates in combination with a breakpoint analysis is already used to automatically detect jumps in a timeseries. The following approach also aims in detecting an instrumental change, by deriving a typical value range (TVR) based on a running window of the previous 30 days. The TVR here is defined as the QI value range under best possible (clear-sky) conditions. The procedure is as followed:

1 Extract aggregated percentiles (10,50,90) over the last 30 days.

2 Subset days which only have DQ0.

3 Calculate the TVR:

- 3a Take the 10% percentile of the daily 10% percentile values.
- 3b Take the 90% percentile of the daily 90% percentile values.
- 3c Take the interquartile range IQR of the daily median values: 25% and 75% quartile.

4 Compare today's aggregated median value with the TVR

An example of this procedure is shown in Figure 15 for direct sun NO2 wrms of P164 at SeoulSNU. The three plots correspond to a weekly check: initial check (top), 1 week later (middle), 2 weeks after the initial check (bottom). Each plot shows the available DQ0 aggregates chronologically ordered on the x-axis. For P164, there where 25 days with DQ0 data within the 30 days testing period during the initial check. The two vertical bars to the right show today's aggregates on index 31, and the TVR on index 32. Herein the thick black vertical area is the IQR (TVR step 3c) and the light grey range covers the respective percentiles (TVR step 3a-b). The TVR gives a good impression of the typcial DQ0 wrms during the initial check, although there is a jump during the testing period where the last 2 days show a new level . The TVR calculation is already affected by those two days, but not as strong

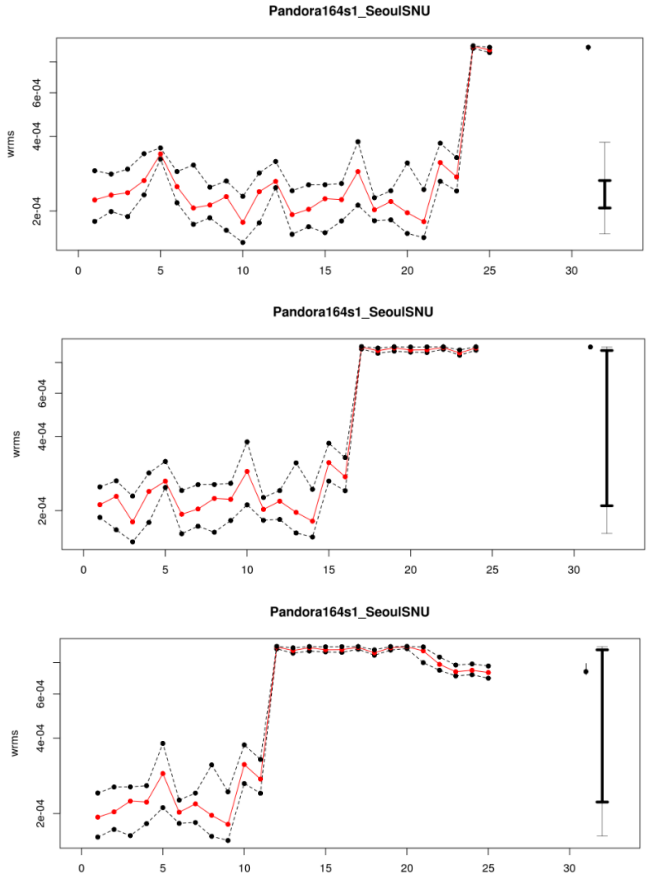

Figure 15: TVR determination for three time periods of NO2 wrms for P164 at SeoulSNU, for aggregates of AMF group [1,3): initial (top), 1 week later (middle), 2 weeks later (bottom). Dashed lines show the 10,90% percentile (black) and the median (red) of the daily aggregates of only DQ0 data. Index 31 shows the aggregates for the last available day, and index 32 the TVR.

as during the second check (middle graphic of Fig. 15). Herein, the IQR is strongly affected by this new wrms level due to the running 30 day window. Two weeks after the initial check, the IQR is still covering a large value range, and today's median value is within the TVR.

The TVR adjustment to new levels is expected due to the running window. But as visible during the initial check, the effect of 2 days is almost negligible. This means that it needs least more than 3 consecutive days with a new DQ0 level before TVR begins to adjust. Even after 1 week, the last evaluation day is outside the TVR. For weekly network checks, the TVR would be sufficiently stable to give warning If there would be a severe issue which result in no DQ0 data, such days would not be included in the TVR calculation.

The idea is now to use the TVR as a reference range, and compare the last day with the previous 30 day's TVR. Finally, five different cases can be obtained:

-2 Median is outside the TVR and below the 10% percentile

-1 Median is between the IQR and the 10% percentile

0 Median is within the IQR of the TVR

1 Median is between the IQR and the 90% percentile

2 Median is outside the TVR and above the 90% percentile

The whole procedure can be done for all QI's that are of interest for QC, for all Pandora's providing processed data, which should provide an overview about changes in the instrument's DQ. Doing this kind of evaluation is planned to give a first overview in decision-making which instruments have changed and need a more detailed look.

An example for TVR determination and comparison, and a first warning system is shown in Figure 16 for 35 instruments, and 6 different QI's. The bottom row represents P164 at SeoulSNU and shows the initial comparision to TVR as illustrated in Figure 15 (top). The warning matrix shows that the last day's wrms exceeds the wrms TVR for all three AMF groups, while the TVR comparison of other QI's do not show a clear change. Only VCU is also leaving its TVR on the lower end. However, if 2-3 AMF groups are affected, a closer look for P164 is necessary and the jump visible in Figure 15 (top) could also be detected with the breakpoint analysis (not shown). In this particular case, the instrument was moved onto another mounting platform, which obsviously changed the spectral response that the wrms increased.

A more clear example is given by Pandora29 at Fairbanks, Alaska which lost the alignment at the end of April 2020. This change is visible in all AMF groups and all QI's. Such a severe issue does not leave any DQ0 data, which means there is no TVR adjustment happening with time. However, the number of DQ0 days will become smaller due to the sliding window approach, and after 1 month the rows would be colored bright yellow. This tells the operator that there are data in the testperiod, but no DQ0 to derive a TVR.

This system is currently being used for NO2, but applicable for other species as O3. The analysis highlights that even small changes within DQ0 can be detected, but lead to an adjustment of the TVR in the running window approach. However, this adjustment takes longer than the typical reaction time of the PGN QC operator is, since this QC check is supposed to happen once per week.

4.3 QP3

4.3.1 Head sensor readings

This idea rises from the fact that head sensors are supposed to be well-sealed and therefore represent an almost closed system regarding the ideal gas equation $pV = nRT$, where p(pressure),T(temperature) is measured in recent instruments. Since the temperature-dependent volume change is expected to be negligible, the only term that could change over time is the right site, which includes n (number of moles) and R (universal gas constant). nR is also expected to be constant over time if:

- There is no leakage
- The head sensor is not opened by purpose There is another sink/source in the system (e.g. evaporation / condensation)

Both cases would cause a mixing of the initially closed head sensor air with outer air.

Figure 17 shows two example days for Pandora 190 located at Bangkok, where the day on 2021-10-25 shows a high R2 value, while this correlation is gone on the

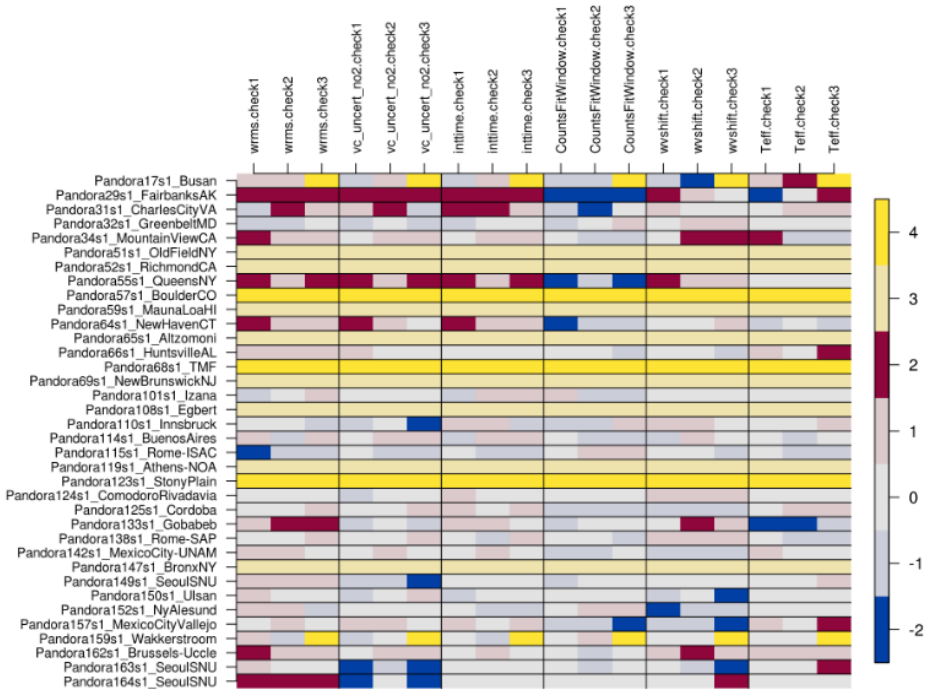

Figure 16: Warning matrix of 30 days TVR analysis for 35 Instruments made on 2020-05-29, where rows represent the instruments and columns 6 QI's with 3 AMF groups for each: wrms, VCU, integration time, counts in the fitting window, L2Fit wavelength shfit, wavelength effective temperature. Colorcoding refers to the 5 classes where the last evaluation can fall into the TVR (-2,-1,0,1,2) where blue colors are blue represent smaller values than the TVR IQR, and red colors larger values than the TVR IQR. The two yellow colors are special cases for instruments where the last available data are outside testperiod (light yellow), and data available data within the testperiod but without DQ0 (bright yellow)

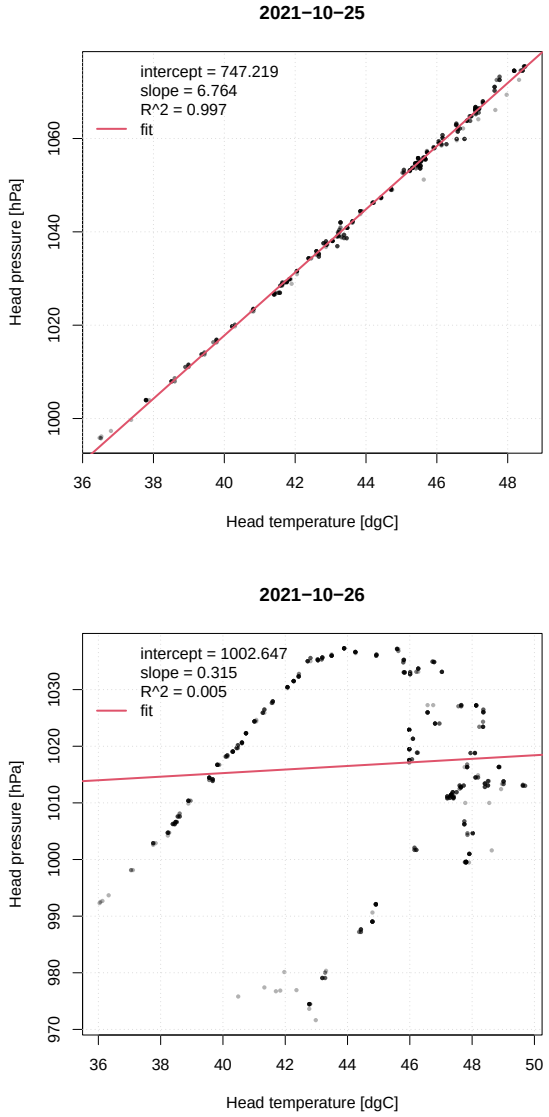

Figure 17: Head sensor readings for Pandora 190 (Bangkok).

following day. We expect that this indicates that the head is 'breathing' and internal air is mixed with outer air.

Thankfully to atmospheric measurements that are provided by Hugo De Backer from RMI, we could compare the internal head sensor measurements with atmospheric measurements taken nearby Pandora 162, located at Brussels-Uccle in Belgium. From the upper graphic in Figure 18 we see that the expected correlation does not exist. The corresponding lower graphic highlights the strong correlation of daily variations in temperatures (red) and pressures (black). With rising head temperatures we would expect also an increase in the head pressure, but instead we observe exactly the same variation as seen in the atmosphere.

Although it has to be investigated in detail how an open head is affecting a retrieval, an open head allows water to enter the system, which can lead to problems with the electronics. However, the R2 value as provided by the linear fit can be used to monitor the head sensor conditions.

4.4 Conclusion

The presented GMM approach is suitable to characterize instrument-specific quality limits for QP1, and leads to objectively defined QF limits for the PGN dataproducts. However, the statistical model can be improved in order to capture daily variations of QI's. Nevertheless, the objective quality flagging leads to a proper filtering of obvious outliers, and the approach is extendable to other QI's than presented.

Regarding routine QC, QP2 uses a statistical breakpoint analysis on daily aggregated median values to search for structural changes in a QI's characteristic. The approach is implemented operationally and applied on AMF groups to capture daily QI variations. In addition to the breakpoint analysis, a TVR determination is performed for each QI as part of a warning system to highlight an instrument where its' QI leaves the TVR.

From QP3, we make use of a derived variable from the headsensor readings, which allow to recognize sealing problems. Although this does not necessarily affect the retrievals, sealing problems can cause severe issue with the electronics.

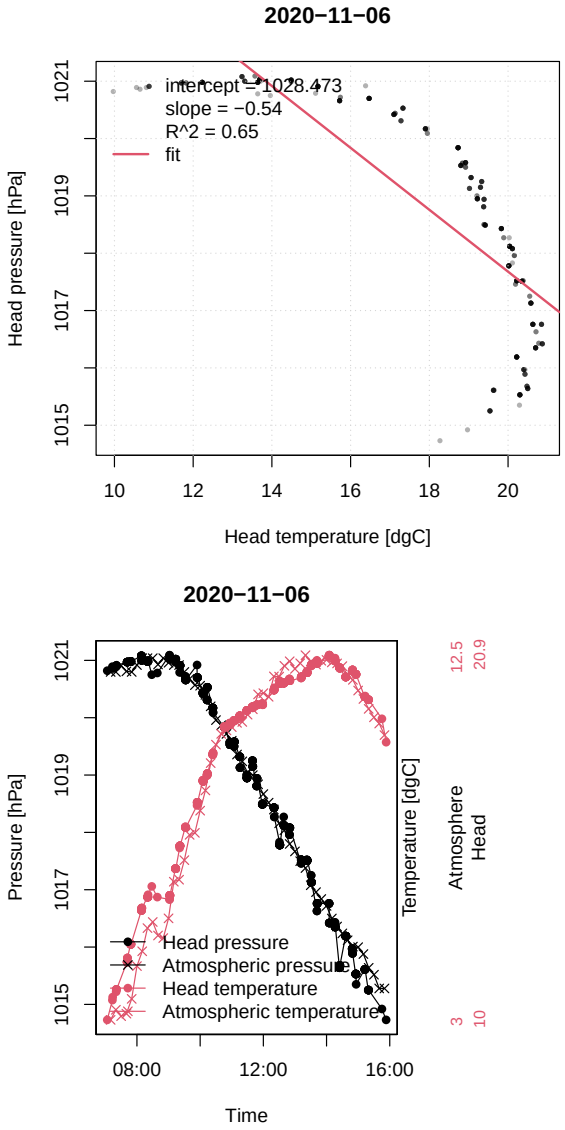

Figure 18: Head sensor readings for Pandora 162 (Brussels-Uccle).

5 QAQC procedures under testing

This section describes QAQC procedures, which are in the testing phase. Each pillar section describes the used approach. The end of this section gives concluding remarks.

5.1 QP2

5.1.1 O3 temperature

With the upcoming 1.8 processor version, it is possible to retrieve the ozone effective temperature [5] as part of the ozone retrieval. Based on the assumption that the majority of O3 is located in the stratosphere, there should be little variation over the day (+/- 0.5K) with an almost flat temperature shape. This is not necessarily true for all locations, because if there would be a significant tropospheric contribution which would increase over the day, the ozone effective temperature is therefore also expected to increase over the day. This is expected to be visible for highly polluted areas as Mexico City.

However, the absolute value of O3 temperature is not of interest for this QC, rather than a sudden change in the diurnal behaviour.

O3 temperature is being calibrated with the AXC approach [5], which means an instrumental change can affect the calibration validity that the synthetic reference spectra is not valid anymore. Typically, this shows up in systematic daily shapes.

Figures 19 shows an example for Wakkerstroom, South Africa, for a valid period of 4 days around the reference, where the O3 temperature variation is almost within an interval of 1K without systematic daily shapes.

The idea is to fit a second order polynomial in the O3 temperature for each individual day:

$$O3T(t) = b0 + b1 \cdot t + b2 \cdot t^2 \tag{3}$$

where t denotes the time of the day. The obtained b2 parameter is assumed to indicate a systematic behaviour due to instrumental changes if b2 > 1. If the calibration is still valid, b2 should be close to zero, and this curvature term does not reveal any systematic daily shape.

However, for periods where this is not true, b2 seems to nicely highlight such periods as illustrated in Figure 20.

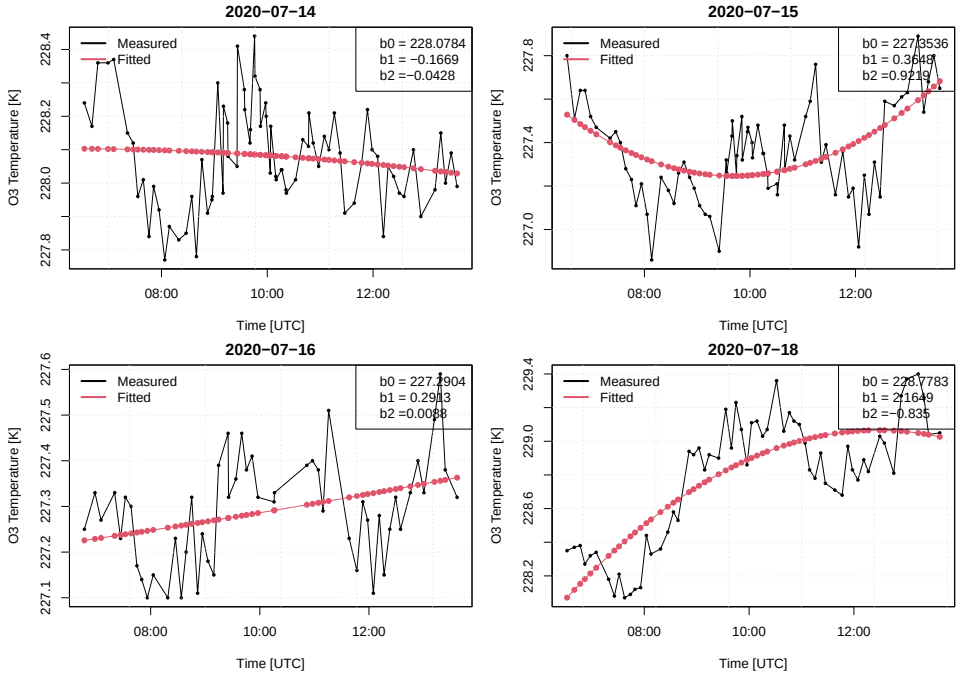

Figure 19: Ozone temperature for Wakkerstroom, South Africa for a period with a valid calibration

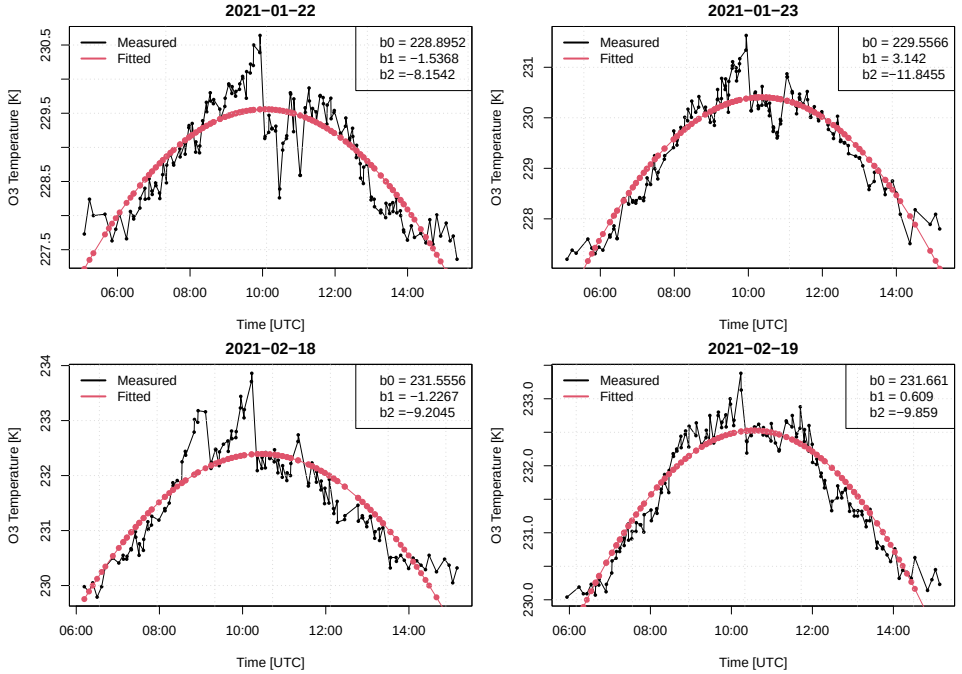

Figure 20: Ozone temperature for Wakkerstroom, South Africa for a period where a new calibration is needed.

With more and more instruments being calibrated with the 1.8 processor, O3 temperature will be evaluated for other sites with longer timeseries and known instrumental changes.

5.2 QP3

5.2.1 Quality codes using extraterrestrial reference spectra

Most of the gas retrievals use higher order closure polynomials, which are actually masking instrumental changes. Moreover, there might be a spectral dependence of change patterns, wherefore it is essential to investigate different fitting windows.

Therefore, we split the Pandora spectral range of interest into 4 distict fitting windows of length 70 nm from 300-520 nm, to also have a certain overlap. Although we do not apply any UV retrievals using the DIFF, the quality codes are also defined for the UV region for this filter. The retrievals use only a smoothing polynomial of order 1 to immediately see instrumental changes, including only the strongest absorbers using the extraterrestrial reference spectra. Table 2 gives an overview about the quality codes.

Table 2: Fitting setup for the quality codes. Resolution, wavelengthshift, and offset polynomial used is zero, while smoothing polynomial is one.

QC-code	Start [nm]	End [nm]	Fitted gases
w1	300	370	O3,NO2,O2O2,SO2
w2	350	420	O3,NO2,O2O2
w3	400	470	O3,NO2,O2O2
w4	450	520	O3,NO2,O2O2

The idea of those qcodes is not to look at the gas retrievals itself but to look at the spectral residuals, and if there might be a change in time. An example of this approach is presented by Figure 21. Using the extraterrestrial retrieval we expect certain spectral residuals since it is not a Pandora spectra, but convoluteted with the Pandora slit. Therefore, we want to focuse on the lower graphic of Figure 21 and evaluate the change in time of the spectral difference to a pre-defined reference period.

5.3 Conclusion

The Ozone effective temperature is being tested as a proxy for instrumental changes. Similarly, quality codes using small order closure polynomials apply extraterestrial references that highlight spectral residuals related to calibration analysis, but also due to changes in the field. Both approaches are being tested to detect instrumental changes.

6 Outlook - Strategies

This section gives an outlook of potential improvements for the three pillars.

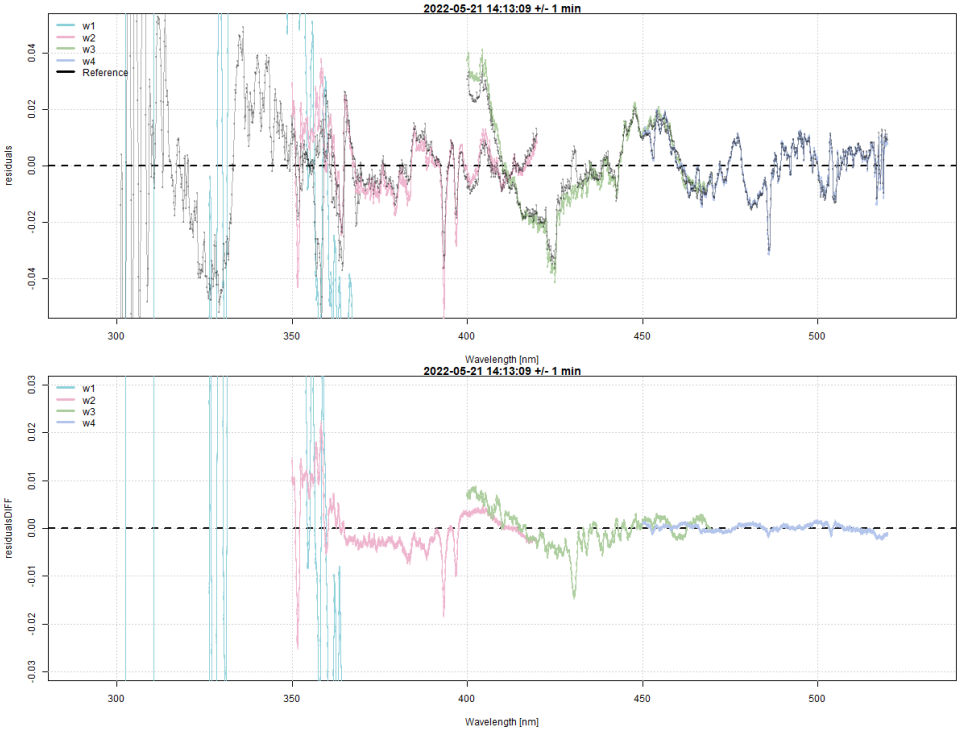

Figure 21: Spectral residuals (top) and the difference (bottom) to a picked reference measurement, shown for the four quality codes w1,w2,w3,w4 for Pandora 157 at MexicoCity-Vallejo

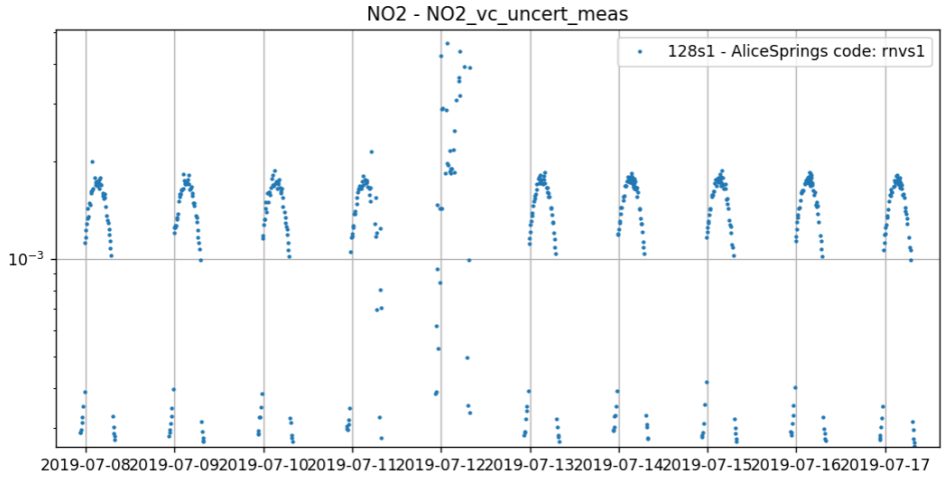

Figure 22: NO_2 vertical column uncertainty for AliceSprings from 2019-07-08 to 2019-07-17.

6.1 QP1

6.1.1 New/Modified QA Parameters

The VCU and wrms are typical QI used. However, both parameters show an air mass factor (AMF) dependence. As an example, Figure 22 illustrates the VCU for AliceSprings for a clearsky period, showing a clear inverse-U shape over the day.

This intra-daily variation would require AMF dependent thresholds for QF. A possible solution is an integration time weighting to reduce the daily variation towards an almost flat baseline for clearsky days.

6.2 QP2

6.2.1 Representativeness Index

The representativeness index (RI) is based on GMM results for individual QI as described in Section 5. Since the GMM obtains different results for different locations, it might me more suitable to use site-specific or instrument-specific thresholds for QF.

Figure 23 shows the vertical column uncertainty of four instruments in Boston. The difference results from physical differences in the filters used, which lead to

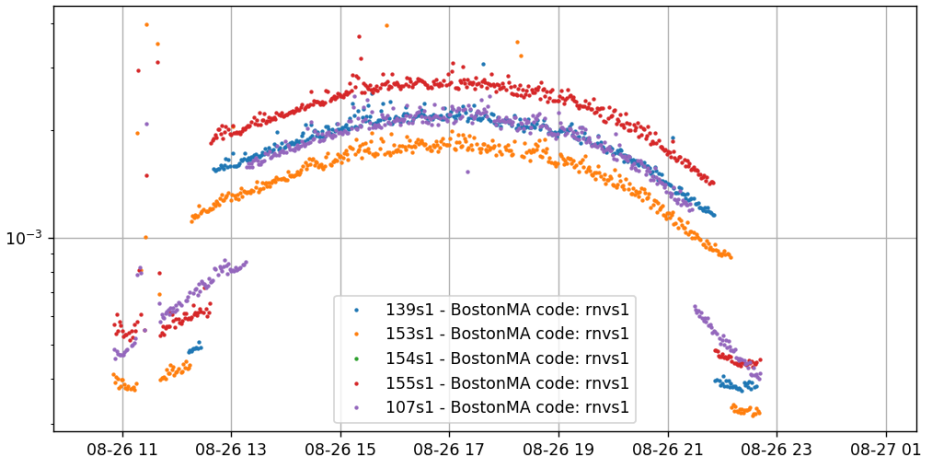

Figure 23: DQ0 data of NO_2 vertical column uncertainty for four instruments at Boston, USA on 2019-08-26.

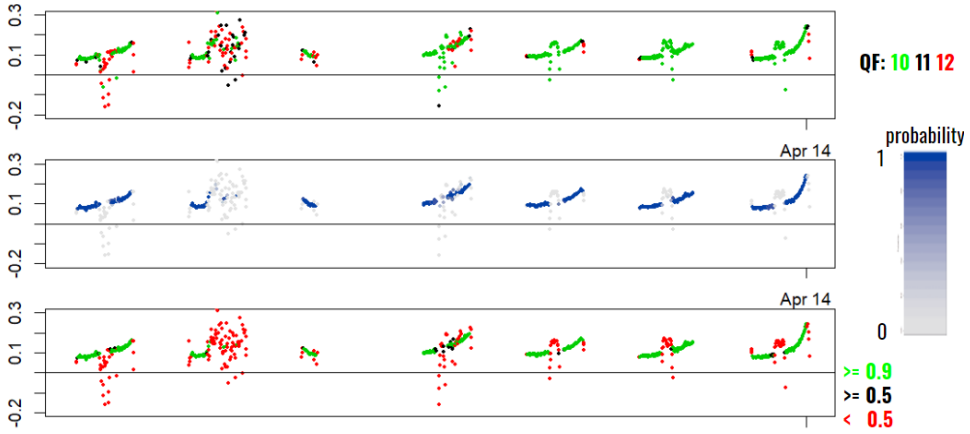

Figure 24: NO_2 vertical column amounts in DU using different QF: nvs0 (top figure), GMM P1 (middle), GMM P1 for selected quantiles. Green color refers to DQ0, black to DQ1, and red to DQ2 (top and bottom figure). Blue color code refers to the probability P1 (middle figure).

differences in the filter transmissions and therefore in the integration time used by the BSS. If QIT1 for this clearsky day is based on Pandora139 (blue) and would be used for QF, then Pandora155 (red) would exceed the quality limits and data would be of DQ1 although all instruments are calibrated and measuring fine.

Contrary, fixed thresholds for QF applied to the entire PGN can lead to DQ0 data which definitely should not be labeled as high quality data, as e.g., negative vertical column amounts for NO_2. An example is given for Altzomoni (2019-04-07 to 2019-04-14), where DQ0 data using nvs0 show negative 'outliers' (top figure, green data points).

By using site-specific GMM results for P1 using VCU as QI, there are no obvious outliers left showing high P1 values (Figure 24, middle). Note that these results differ to the GMM described in Section 5 as the AMF dependence is explicitly accounted by including the square of the solar zenith angle as an independent variable for the expectation value.

P1, which is called the RI, could be used by end-users to choose selected quantiles for their own quality flagging, as illustrated in Figure 24 (bottom). Additionally, GMM based distribution parameters for individual sites could serve as a proxy for changes in the calibration status. E.g., if the daily shape of VCU as illustrated in Figure 23 is suddenly biased, it is possible that the pointing of the instrument changed and higher integration times had to be used.

6.3 QP3

6.3.1 Direct sun total column O2O2

Similarly to O3 temperature, direct sun total column O2O2 is supposed to serve as a proxy to detect instrumental changes. O2O2 absorbtion peaks are relatively well-defined and located at different spectral regions. In theory, small order smoothing polynomials should be sufficient for valid calibration periods, but are supposed to be larger if instrumental changes affect the calibration. Moreover, such changes could also have an spectral effect, affecting certain wavelengths stronger than others.

Therefore, the idea is to retrieve direct sun O2O2 for different wavelenghts, e.g. for a fitting windows in the UV and VIS region, retrieved with small and larger smoothing polynomials [5].

An example of how O2O2 could be used is illustrated in Figure 25, showing daily averages of two different O2O2 fitting windows [5], retrieved with smoothing polynomials 1 and 3. The low order polynomials for both fitting windows (black and red) show a stronger day to day variation, and its difference is larger as comparing high order polynomials (blue and cyan). This might be already an indicator for

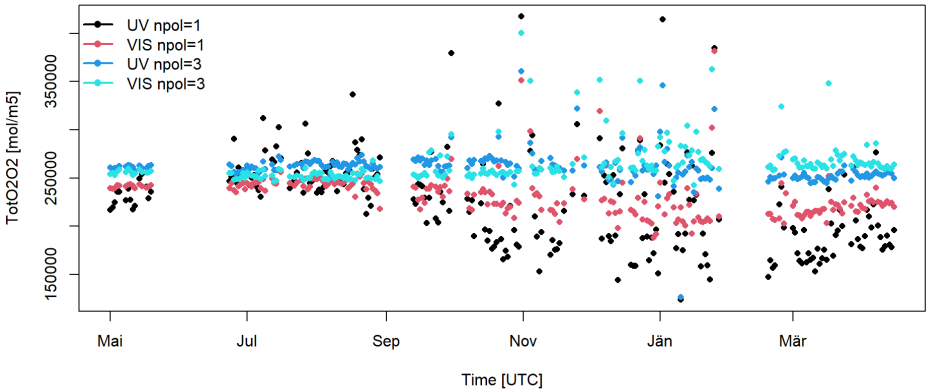

Figure 25: Daily averages of direct sun O2O2 for Wakkerstroom, South Africa from April 2020 to April 2021, fitted for the 360 nm (UV) and 480 nm (VIS) peak of O2O2, for smoothing polynomials order 1 and 3.

imperfections in the calibration.

However, in particular in July, which is the period where the reference has been taken, the difference is generally smallest. This is expected around the reference for a valid period. But this difference begins to increase towards winter. Most importantly, the difference between the high order polynomials (blue and cyan) reverse its sign in December 2020, which is also visible in the L1 wavelength change (not shown). This date served as a new validity period for operational usage already.

Although the wavelength change could already be sufficient to detect the change, the O2O2 fitted in different wavelength range can give additional information to highlight the spectral region might be affected most.

6.3.2 Direct sun total column O2

Similarly to direct sun total column O2O2, the total O2 column retrievals which are currently under development could be used for detecting instrumental changes. O2 has four well defined bands within the Pandora S2 spectral range. At 690 nm and 765 nm, the optical depth is already above 1e-2 with mainly H2O and O3 are interfering in the fitting windows.

O2 is expected to show less daily variations. This feature in combination with a solely pressure and temperature dependence could serve as an ideal candidate for a QC data product.