

Scoring: A Novel Approach Towards Automated and Reliable Certification of Pipeline Products

Reinhard W. Hanuschik¹, Wolfgang Hummel, Mark Neeser, Burkhard Wolff
Data Processing and Quality Control Group, European Southern Observatory,
Karl-Schwarzschild-Str. 2
D-85478 Garching, Germany

ABSTRACT

By 2010, the Paranal Observatory will host at least 15 instruments. The continuous increase in both the complexity and quantity of detectors has required the implementation of novel methods for the quality control of the resulting stream of data. We present the new and powerful concept of scoring which is used both for the certification process and the Health Check monitor. Scoring can reliably and automatically measure and assess the quality of arbitrarily amounts of data.

Keywords: Data processing, quality control, pipelines, QC parameters, databases, health checks

1. CALIBRATIONS: THE GOAL

At the Very Large Telescope (VLT), like at any other ground-based observatory, calibrations are regularly acquired with the purpose of measuring the instrumental and atmospheric signature. If done in the proper way and under valid conditions, this information can then be used to remove that signature from science observations; a process commonly called reduction.

The European Southern Observatory (ESO) has implemented a calibration plan which foresees a regular, predictable calibration scheme. It consists of daytime calibrations (bias, darks, flats, arcs); twilight calibrations (twilight flats); and night-time calibrations (flux, telluric, and RV standard stars; calibrators). This calibration plan is set up such that it includes both science-driven calibrations and calibrations used for maintenance and monitoring. The latter ones are also called Health Check (HC) calibrations. They provide a continuous record of instrument performance which can be evaluated for maintenance (even preventive), long-term trending, stabilization, and performance improvements.

At ESO's headquarters, the processing of calibration data and the reduction of science data is performed at the Data Processing and Quality Control Group (shorter: QC Garching) by instrument pipelines (Ballester et al. 2006). An important feature of these pipelines is that they not only *extract* the instrument and atmosphere signature and remove it from the science observations, but that they also *measure* that signature. Thereby they provide metrics for quality assessment, quality control and trending. At ESO, these numbers are called QC1 parameters. They are defined individually for each instrument, each mode and each data type. These numbers are stored in the QC1 database. From there they can be retrieved using a browser and a plotter interface (Figure 1).

2. QC1 PARAMETERS: THE CHALLENGE

At the time of writing (May 2008), the VLT's host 11 instruments and the VLTI (the integrated VLT interferometer) adds another two instruments. By 2010, two new telescopes with survey cameras will be in operation on Paranal. Most of these instruments have multiple modes, each of which usually requires several different types of calibration data. The QC process needs to control, on the order of, 150 different data types, each of which is characterized by 3-10 different QC1 parameters. From this one can estimate that a total of 500-1000 QC1 parameters need to be measured and monitored, all of them at different frequencies ranging from once per day to once per semester. This is the *complexity challenge*.

The current VLT/VLTI instruments have a maximum of four detectors simultaneously on the sky. With the operation of the survey instruments, this number will soon inflate to 16 (for the infra-red survey camera VIRCAM) and 32 (for the

¹ rhanusch@eso.org

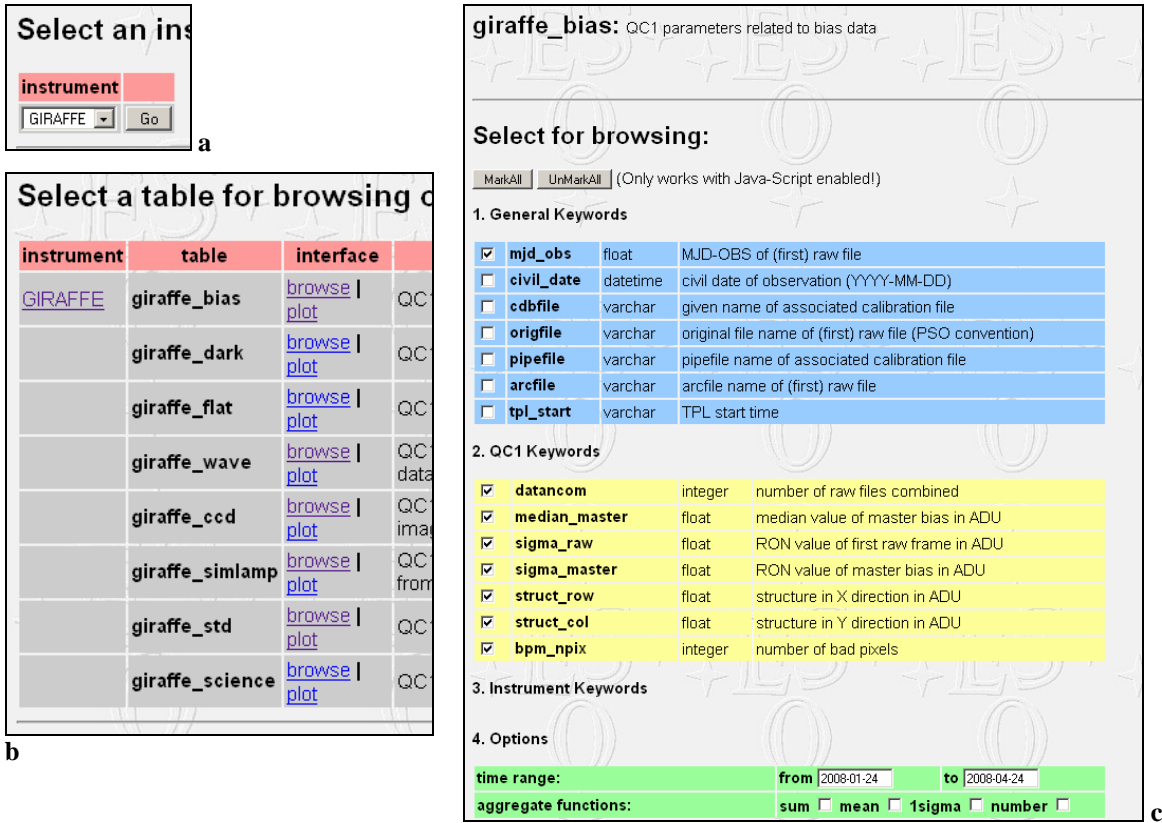


Figure 1. This is a screenshot of one of the interfaces to the QC1 database (c). You can find it from the URL <http://archive.eso.org/bin/qc1.cgi>, selecting “GIRAFFE” (a) and “giraffe_bias: browse” (b). You could select here a set of QC1 parameters for a given time range and include statistics like sum, mean, and scatter.

optical survey camera OmegaCAM). Since most QC1 parameters need to be measured and monitored per detector, a single VIRCAM night will easily produce thousands of QC1 parameter values, as much as all current VLT/VLTI instruments combined. This we call the *volume challenge*.

3. SCORING: THE SOLUTION

3.1 Principles of scoring

Our strategy to manage the two challenges of “complexity” and “volume” is implemented in three stages:

- measure quality (→ QC report)
- compare quality (→ trending)
- assess quality (→ scoring).

If all three steps can be implemented in automatic procedures, the whole quality control process can be automatically managed up to the level that the system selects all outliers and presents them for human intervention.

The *measurement* of the QC1 parameters is mostly done by the pipelines, but may also be supported by specific QC scripts. It depends on the proper parameter selection and on proper algorithms. Both required a learning process, and still require continuous adaption and improvement, but the process exists (since long) and is stable.

Putting the QC1 parameters into the context of other, similarly derived data points is *trending*. Only through this process will it be evident if a certain value exhibits nominal behavior or represents an outlier. In many cases, the exact value of a

parameter is less relevant than its evolution over time. For instance, the exact level of a bias frame is less relevant for QC purposes than detecting a sudden jump or gradual change, which may indicate a problem with the electronics. Such behavior can only be detected in the trending, not in the assessment of a single file. Trending is, like the QC1 parameters, an established process at ESO.

Closely related to trending is the *assessment* or evaluation; the critical third component in our QC process. Our solution employs scoring and is a novel approach. We have started implementing it in the file-by-file certification, and in the trending process.

Scoring essentially means finding outliers and, if found, raising a warning flag. We define an outlier by its non-compliance to pre-defined lower and upper thresholds.

3.2 Thresholds

Within our QC process, we define thresholds in two different ways: through statistics and through specifications. *Statistical* evaluation has no preconception about the “right” value. It simply derives lower and upper thresholds from an average and a scatter parameter (e.g. mean \pm 3 sigma). This usually works well for homogeneous data sets that are well-sampled. Statistical evaluation (at least in the simple-minded implementation used in ESO QC) is quite sensitive to clustered outliers; while a single outlier can be clipped to not skew the mean, a certain number of them will ultimately widen the threshold values and make any tests less sensitive.

The other important evaluation concept is based on *specifications*. No matter what the actual trend in the data is, they should remain within a specified and fixed range. For the process to work properly, it does not really matter whether that range comes from a specification document or from experience. As an example, a flat-field is acceptable as long as its signal level is above a certain exposure level (under-exposure if dominated by photon noise) and below another critical level (over-exposure). The exact flux value, so long that it is between these two extremes, does not matter.

The specification concept has the advantage that it works independently of the detailed behavior of the data points and reliably detects outliers, irrespective of how rare or frequent they are. In contrast to the statistical evaluation this method is applicable to inhomogeneous data sets. A practical drawback, however, is that specifications require active configuration; in other words knowledge about the underlying instrument component.

Selecting one or the other threshold strategy depends on the exact QC1 parameter to be checked. In the example mentioned before, the measured flat-field exposure level (the total counts) is scored against relaxed thresholds as long as its fitness for science reduction is assessed. If the calibration lamp performance is going to be monitored, the flux (as counts per second) needs to be scored against narrow thresholds for any degradations to be readily visible.

3.3 Certification of products

When it comes to certification, i.e. acceptance or rejection of a calibration product file, its quality needs to be assessed. As long as it is based on the above principles of measurement and trending, the scoring can be done in an automatic way. Once configured correctly and inclusively, scoring can auto-certify (or auto-reject) arbitrarily large and complex data sets, technically limited only by computational performance.

The scoring process is now being used for data from all VLT/VLTI instruments. As we have implemented it, scoring per QC1 parameter is very simple, it returns either 0 (ok) or 1 (nok: outlier). Thereby the complexity of calibration data quality assessment is dramatically reduced to just counting outliers².

The two main modes of assessment propagate to scoring: we score QC1 values either against *specified thresholds*, or against comparison to a history of data, thresholds then being defined by their average and sigma values. The former scoring mode can also be called “static”, the latter “dynamic”. In the flat-field example from above, the exposure level will be scored statically (against specifications), the flux level dynamically (against statistics).

² More precisely, the complexity has moved to the configuration of our scoring tools.

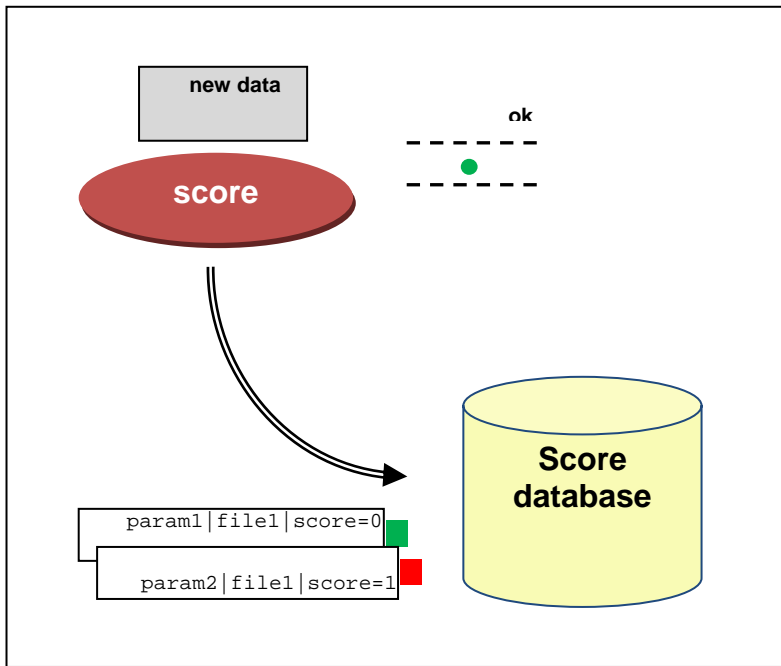


Figure 2. Sketch of the scoring process. Scores for each product consist of a number of records which are inserted into the scoring database. Scores are either 0 (ok, light grey in the figure) or 1 (nok, dark grey).

4. IMPLEMENTATION

4.1 Scoring of new products

Scoring is applied by the QC group at two distinct stages in the QC workflow:

- certification of new products
- monitoring of instrument health (“Health Check reports”)

The fundamental data set for scoring is a QC1 parameter, its value, its score, and the applied thresholds. As a result of the scoring process, a set of such records per product file is obtained, evaluated, and inserted into the scoring database (Figure 2). In addition to the scores per QC1 parameter, all scores for the product are added to the total score. Still, the total product file score is either 0 (ok), or >0 (nok).

4.2 Some examples

Figure 3 displays an example, with 6 QC1 parameters scored, all ok. The score result is symbolized with little squares, either green (grey in the figure) or red (dark-grey). Each score square is linked to a dynamic query to the QC1 database which could be followed to display a trending plot of that QC1 parameter (“information on demand”). The cumulative total score (last row) is usually evaluated for automatic certification, while the detailed scores are reviewed only when non-zero scores trigger the need for further investigation.

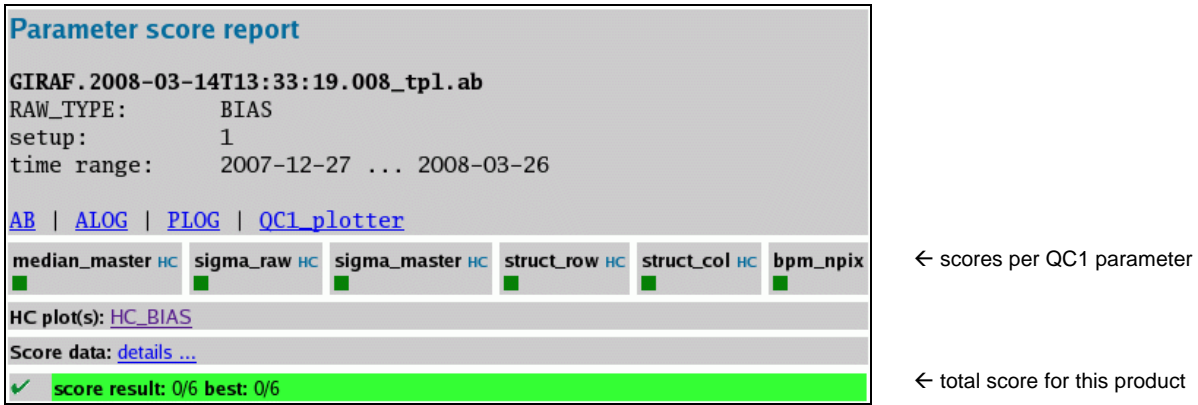


Figure 3. The score report for a set of GIRAFFE BIAS frames from 2008-03-14. The six QC1 parameters (median_master etc.) are all scored against configured thresholds. In this example, they all comply with the threshold ranges and are scored ok. The total score is also determined to 0 (ok). Note the original score report comes as HTML file using color codes (green for ok, red for nok scores).

In Figure 4 we show an example for a product file in which two outliers have been found (mean_width and flux), resulting in a non-zero total score. In the original score report, they display in red and immediately catch attention.

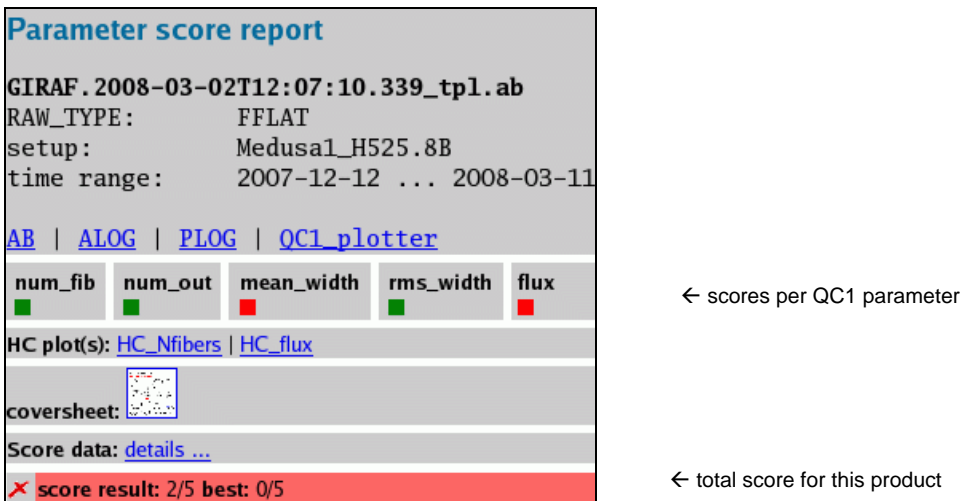
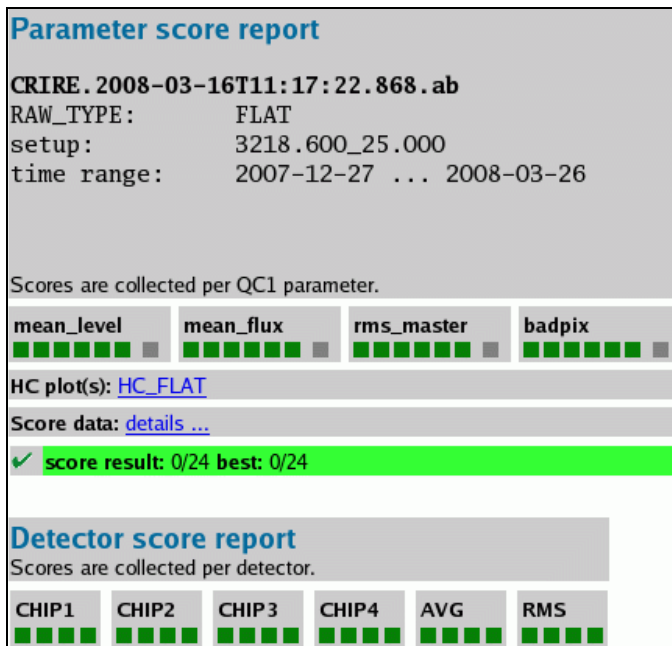


Figure 4. For comparison, this is a score report with two outliers found (mean_width and flux, with alert boxes in darker grey), resulting in an alarm flag.

At a higher level (in a monitor displaying the score results for all product files of that night, see Figure 5) the non-zero score becomes immediately obvious, and the score report then reveals the details of which parameter has failed.

GIRAF.2008-03-02T05:36:16.720.ab	compl.	OK	giscience	SCL_MED	Medusa2_H525.8B	OK	P_LOG	0.9	DONE			OK
GIRAF.2008-03-02T12:07:10.339_tpl.ab	compl.	OK	gimasterflat	FFLAT	Medusa1_H525.8B	OK	P_LOG	2.7	DONE	✗	(2/5)	REJ
GIRAF.2008-03-02T12:18:52.880.ab	compl.	OK	giwavecalibration	WAVE	Medusa1_H525.8B	OK	P_LOG!	0.7	DONE	✗	(4/4)	REJ
GIRAF.2008-03-02T12:51:32.483_tpl.ab	compl.	OK	gimasterflat	FFLAT	Medusa2_H525.8B	OK	P_LOG	3.0	DONE	✓	(0/5)	OK
GIRAF.2008-03-02T13:02:40.932.ab	compl.	OK	giwavecalibration	WAVE	Medusa2_H525.8B	OK	P_LOG!	0.9	DONE	✓	(0/4)	OK
GIRAF.2008-03-02T13:15:29.979_tpl.ab	compl.	OK	gimasterbias	BIAS	1	OK	P_LOG	0.5	DONE	✓	(0/6)	OK
GIRAF.2008-03-02T13:26:16.207.ab	compl.	OK	giwavecalibration	WAVE	Argus_H525.8B	OK	P_LOG!	1.5	DONE	✓	(0/4)	OK

Figure 5: The product file from Figure 4, in an overview monitor displaying its total score and being flagged as nok and “rejected”. On the original monitor, these flags are displayed in red.



← scores from four QC1 parameters: The first 4 squares denote the QC1 scoring for each detector, while squares 5 and 6 score the average and standard deviation of the QC1 parameter for all 4 detectors. The last square links to an overview plot.

← total score for this product

← scores per detector (1-4 chips, plus average and scatter)

Figure 6. Score report for a CRIRES flat-field. CRIRES is a near-IR spectrograph with 4 detectors, each of which has four QC1 parameters scored. The total score is 0/24. This is immediately visible, despite its complexity.

Figure 6 demonstrates how even a (moderately) complex situation (scoring a CRIRES flat frame from 4 detectors with 4 QC1 parameters) can be assessed and easily analyzed. For multi-detector instruments, we provide, in addition to the QC1 parameter scores, a set of detector scores which usually consists of scores for all detectors, plus their average and scatter values. This is useful to distinguish between *single-detector outliers* and *correlated outliers* (occurring in all detectors at the same time; see Wolff & Hanuschik [2006] and in particular Wolff et al. [2008] with detailed examples).

4.3 Reducing complexity

The last example nicely demonstrates the ability of scoring to reduce complexity, see Figure 7. In addition to the score report, each product has its own set of customized QC reports with detailed graphical information about data quality. In the case of CRIRES, there are four such reports per product file. While in the first implementation of the QC process the QC scientist had in principle to review them all before certification, now there is only one score report to be inspected, and this only in case of non-zero scores. Then, the reviewer can click deeper into detailed reports – the principle of “information on demand” (see Wolff et al. [2008]).

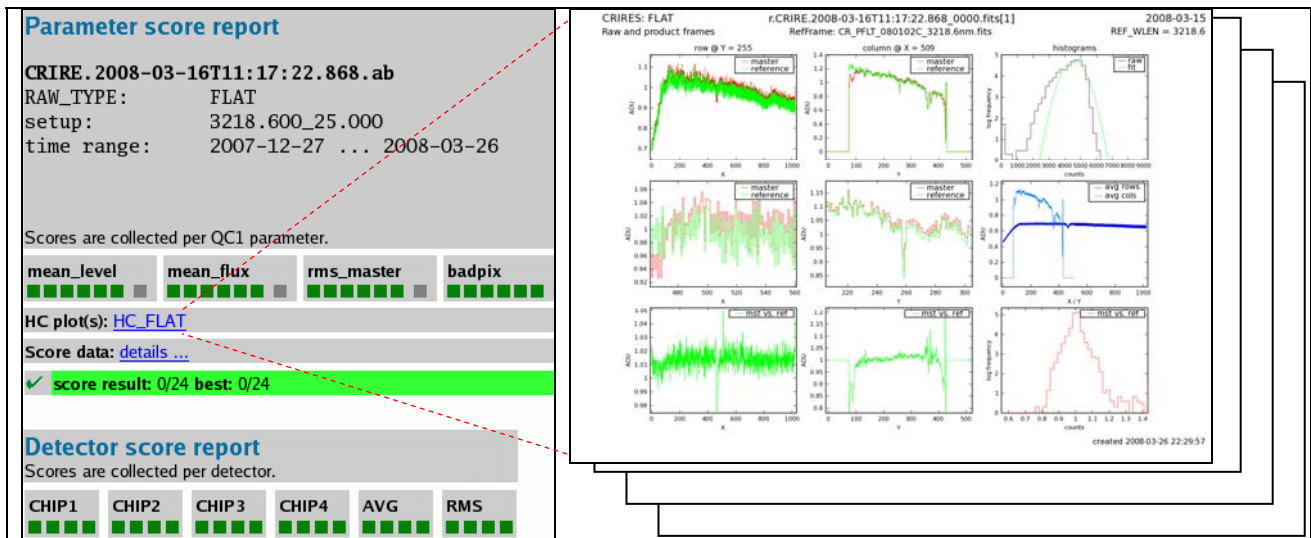


Figure 7. The score report from Figure 6 and the full set of QC reports made visible on demand (four in total, right panel).

4.4 Health Check reports

The Health Check (HC) reports represent the trending of QC1 parameters in a schema which is usually not file-based but more related to instrument properties. All values for a QC1 parameter in a certain time interval are now collected, trended and scored (Figure 7).

Now we can again apply the power of scoring and replace the exact values of the QC1 parameters by their scores, to obtain the quick-look version of the HC report. It focuses on the scores only. This is implemented as a quick-look version and displays only the scores of the last few days. The quick-look version of the HC report is displayed in Figure 8.

Usually the reviewer (VLT astronomers on Paranal and/or QC astronomers in Garching) only needs to inspect the scores to adequately assess the health of the instrument and the quality of the data. Detailed information, if demanded, is readily accessible.

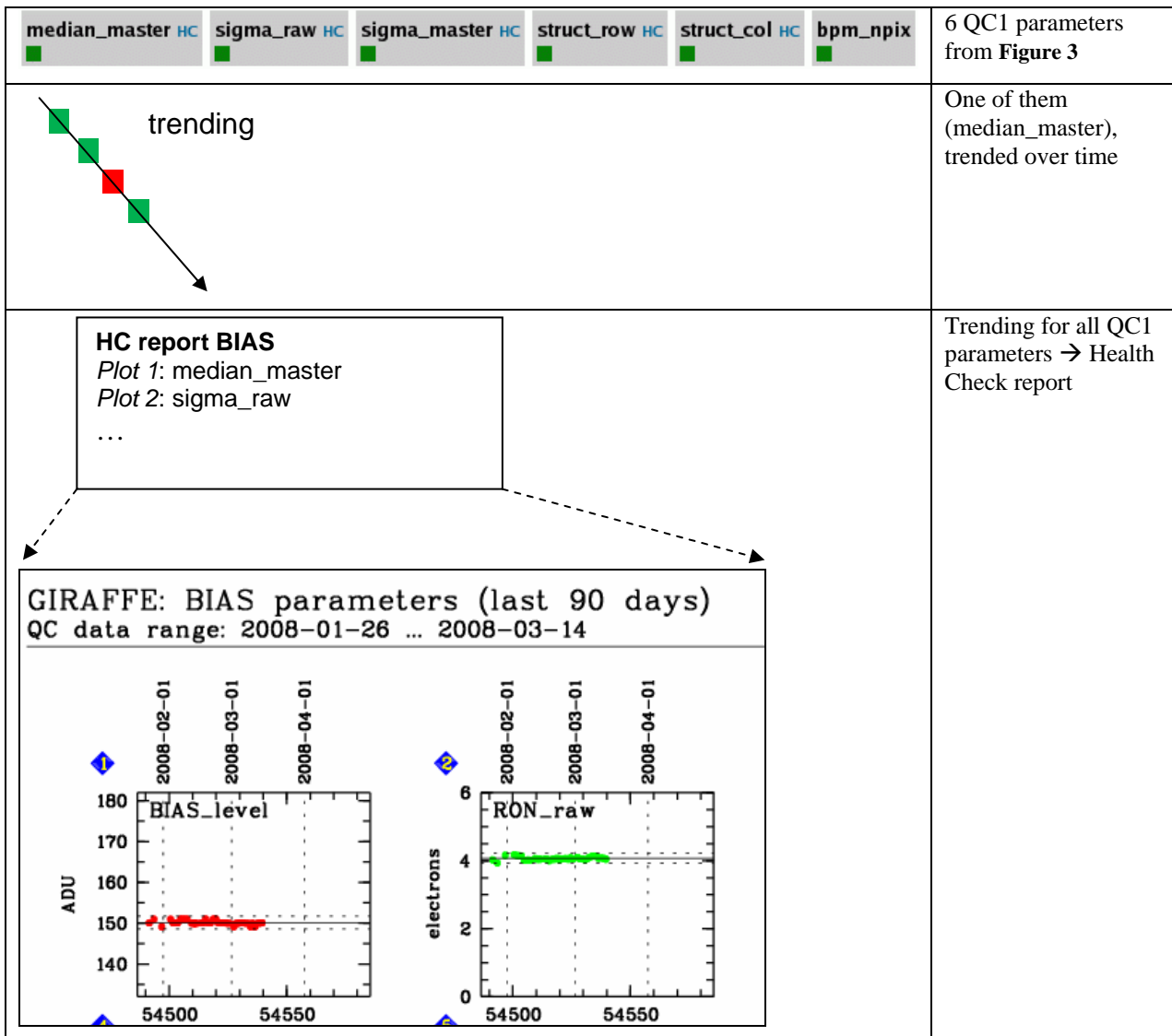


Figure 8. From score report to trending and HC report. A time record of QC1 parameters, sorted by instrument properties, is displayed for a selected time range. The current version of this plot can be found under the URL: http://www.eso.org/qc/GIRAFFE/reports/HEALTH/trend_report_BIAS_HC.html

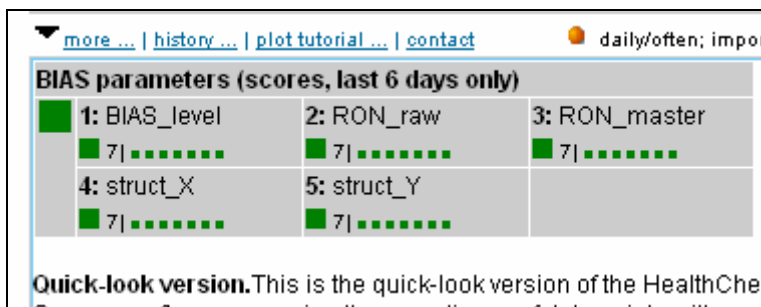


Figure 9. The quick-look score version of Figure 7 (the current version can be found at: http://www.eso.org/qc/GIRAFFE/reports/HEALTH/trend_report_BIAS_QUICK.html)

4.5 Hierarchy of scores

With the introduction of scoring, a powerful hierarchical scheme can be easily implemented. Moving from scoring levels 5 to 0, the scoring detail branches out (increases) and becomes more specific:

Level	Meaning	Example
Level 0	Scores per QC1 parameters per product file and detector	Figure 6
Level 1	Scores per QC1 parameter and product file	Figure 3, Figure 4
Level 2	Scores per QC1 parameter: all level 1 scores from last days	Figure 8, larger squares per parameter
Level 3	Scores per report: all level 2 aggregated	Figure 8, single big square at left
Level 4	Scores per report group: all level 3 aggregated, score per instrument component	navigation bar of URL http://www.eso.org/qc/GIRAFFE/reports/HEALTH/trend_report_BIAS_QUICK.html
Level 5	Scores per instrument: all level 4 scores aggregated	as for level 4

4.6 A closed loop between HC reports and score reports

As sketched in Figure 2, all file scores go into a database. The new values of all QC1 parameters, and their scores, are now evaluated in the HC report. As part of the creation of the HC report, updated statistics of the data set are calculated (among them average and scatter). If the scoring is configured as “dynamic” (thresholds taken from HC results), these updated HC thresholds will now become the reference thresholds for the next file scoring:

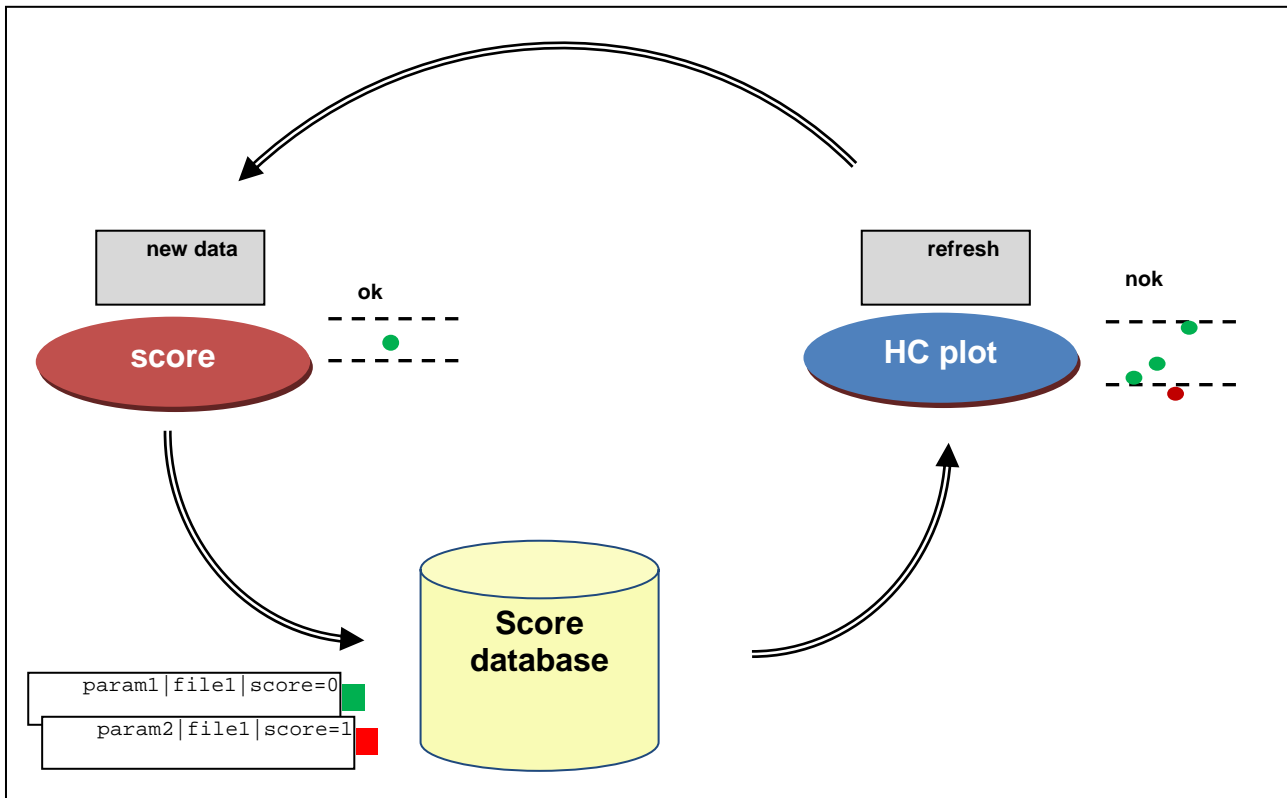


Figure 10: Closed feedback loop with scores and thresholds

5. CONCLUSIONS

We have presented our novel concept of scoring which is a powerful approach for automatic certification of complex and large sets of product data. It checks measured QC parameters against thresholds and returns either a 0 for compliance, or 1 for non-compliance. Thresholds can either be configured as static, specified values, or as dynamic, statistical values. In the latter case, the statistics are derived in trending plots and fed back in a closed loop to the scoring configuration. The concept can be applied to multi-detector instruments and can be implemented in hierarchical form, from a score per detector in a certain product file to a total score for the whole instrument.

This concept has been developed and implemented by the QC group for the certification process. It is currently extended to the Health Check plots and will soon become a core feature. Scoring can effectively help reduce complexity, insulate the quality assessor from a veritable flood of redundant data, and provide detailed information only when demanded.

Recently, scoring has been very effectively used on a large data set in the UVES reprocessing scheme. It was used to automatically assess and certify pipeline products from more than 50,000 UVES exposures (Hanuschik 2007).

REFERENCES

- [1] Ballester, P., Banse, K., Castro, S., Hanuschik, R., et al., "Data reduction pipelines for the Very Large telescope", Proc. SPIE 6270, 62700T-1...10 (2006).
- [2] Wolff, B., Hanuschik R., "Quality Control for multi-detector instruments: a test bed with VIMOS", Proc. SPIE 6270, 62701S-1...10 (2006).
- [3] Wolff, B., Hanuschik, R., Hummel, W., Neeser, M., "Solutions for Quality Control of multi-detector instruments and their application to CRIRES and VIMOS", Proc. SPIE 7016, these proceedings (2008).
- [4] Hanuschik R., <http://www.eso.org/qc/reproUVES/processing.html>