A&A 616, A3 (2018)
https://doi.org/10.1051/0004-6361/201832712

Astronomy & Astrophysics

Special issue

Gaia Data Release 2

Gaia Data Release 2

Processing of the photometric data

M. Riello[1], F. De Angeli[1], D. W. Evans[1], G. Busso[1], N. C. Hambly[4], M. Davidson[4], P. W. Burgess[1], P. Montegriffo[2], P. J. Osborne[1], A. Kewley[1], J. M. Carrasco[3], C. Fabricius[3], C. Jordi[3], C. Cacciari[2], F. van Leeuwen[1], and G. Holland[1]

[1] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
e-mail: mriello@ast.cam.ac.uk
[2] INAF – Osservatorio Astronomico di Bologna, via Gobetti 93/3, 40129 Bologna, Italy
[3] Institut del Ciències del Cosmos (ICC), Universitat de Barcelona (IEEC-UB), c/ Martí i Franquès, 1, 08028 Barcelona, Spain
[4] Institute for Astronomy, School of Physics and Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK

Received 26 January 2018 / Accepted 14 February 2018

ABSTRACT

Context. The second *Gaia* data release is based on 22 months of mission data with an average of 0.9 billion individual CCD observations per day. A data volume of this size and granularity requires a robust and reliable but still flexible system to achieve the demanding accuracy and precision constraints that *Gaia* is capable of delivering.
Aims. We aim to describe the input data, the treatment of blue photometer/red photometer (BP/RP) low-resolution spectra required to produce the integrated $G_{\rm BP}$ and $G_{\rm RP}$ fluxes, the process used to establish the internal *Gaia* photometric system, and finally, the generation of the mean source photometry from the calibrated epoch data for *Gaia* DR2.
Methods. The internal *Gaia* photometric system was initialised using an iterative process that is solely based on *Gaia* data. A set of calibrations was derived for the entire *Gaia* DR2 baseline and then used to produce the final mean source photometry. The photometric catalogue contains 2.5 billion sources comprised of three different grades depending on the availability of colour information and the procedure used to calibrate them: 1.5 billion gold, 144 million silver, and 0.9 billion bronze. These figures reflect the results of the photometric processing; the content of the data release will be different due to the validation and data quality filters applied during the catalogue preparation. The photometric processing pipeline, PhotPipe, implements all the processing and calibration workflows in terms of Map/Reduce jobs based on the Hadoop platform. This is the first example of a processing system for a large astrophysical survey project to make use of these technologies.
Results. The improvements in the generation of the integrated *G*–band fluxes, in the attitude modelling, in the cross-matching, and and in the identification of spurious detections led to a much cleaner input stream for the photometric processing. This, combined with the improvements in the definition of the internal photometric system and calibration flow, produced high-quality photometry. Hadoop proved to be an excellent platform choice for the implementation of PhotPipe in terms of overall performance, scalability, downtime, and manpower required for operations and maintenance.

Key words. instrumentation: photometers – space vehicles: instruments – techniques: photometric –methods: data analysis – catalogs

1. Introduction

The European Space Agency *Gaia* mission (Gaia Collaboration 2016b) was launched in December 2013. After an extended commissioning period, science operations began on 25 July 2014. In September 2016, the first *Gaia* data release (DR1 Gaia Collaboration 2016a) was made available to the scientific community, and it included an astrometric solution based on a combination of *Gaia*, HIPPARCOS, and *Tycho*-2 data (Lindegren et al. 2016) and *G*–band photometry from the first 14 months of operations.

The second *Gaia* data release (DR2) in April 2018 is based on 22 months of mission data and includes an improved astrometric solution based solely on *Gaia* data (Lindegren et al. 2018) and photometry in *G*–band, $G_{\rm BP}$, and $G_{\rm RP}$ for approximately 1.5 billion sources. This paper focusses on the process of calibrating the raw *G*–band photometry and the processing of the low-resolution spectra to produce and calibrate the $G_{\rm BP}$ and $G_{\rm RP}$ photometry. The validation and scientific quality assessment of the calibrated *Gaia* photometry are discussed in the companion paper, Evans et al. (2018). We recommend that the Carrasco et al. (2016) paper on the principles of the photometric calibration of the *G*–band for *Gaia* DR1 be read in conjunction with this paper.

The data processing effort for the *Gaia* mission happens in the context of the Data Processing and Analysis Consortium (DPAC), which is comprised of more than 400 astronomers and software and IT specialists from over 15 European countries (Gaia Collaboration 2016b). Within DPAC, different groups are set up to handle specific aspects of the data treatment

required to deliver science-ready processed data products to the scientific community. The photometric and low-resolution spectra processing system, PhotPipe, consumes a variety of intermediate data products from other DPAC systems, which, when combined with the low-resolution spectra (see Sect. 2), allows us to derive a consistent set of calibrations that removes most instrumental effects and establishes a sound internal photometric system that is finally tied to the Vega system by means of an external calibration process (Carrasco et al. 2016). A fundamental aspect of the calibration process is that the only stage during which external (non-*Gaia*) data are used is in the determination of the external calibration, which uses a set of well-observed spectro-photometric standard stars (SPSS), see Pancino et al. (2012).

For the *G*–band photometry, the processing done by PhotPipe does not start from the raw data (i.e. the reconstructed satellite telemetry), but from the results of the image parameter determination (IPD), produced by the intermediate data update (IDU) system, comprising the integrated flux resulting from a point-spread function (PSF, for 2D observations) or line-spread function (LSF, for 1D observations) fit of the data, the centroid positions, and relevant statistics and quality metrics. For the $G_{\rm BP}$ and $G_{\rm RP}$, PhotPipe starts from the raw data and performs all the pre-processing steps required to produce the uncalibrated integrated flux. Another critical piece of information used by PhotPipe is the cross-match generated by IDU (Castañeda et al. 2018): this process identifies transits belonging to the same astrophysical source after removing spurious detections that are due mostly to artefacts caused by bright sources. The pre-processing of the blue photometer/red photometer (BP/RP) spectra involves the bias and proximity electronic module non-uniformity mitigation (Hambly et al. 2018), the straylight (Gaia Collaboration 2016a) mitigation, and the determination of the geometric calibration mapping the optical distortions and charge-coupled device (CCD) geometry on the focal plane assembly (FPA). More details on the various pre-processing and the subsequent photometric calibration process are provided in Sects. 4 and 5, respectively.

The overall processing performed by DPAC is iterative: each data release does not simply include more data, but it also involves a complete reprocessing from the beginning of the mission, with improved calibrations and algorithms. In particular, there are a number of significant improvements included in *Gaia* DR2. First, the *G*–band pre-processing and IPD have been performed uniformly on the entire data set. In the first data release, the processing was instead performed on a daily basis by the initial data treatment (IDT; see Fabricius et al. 2016): the strict time constraints on IDT and the complexity of the downlink telemetry scheme meant that it was not always possible to derive and use optimal calibrations and that data set completeness was not always ensured. A more detailed discussion of the differences in the IPD process with respect to the *Gaia* DR1 can be found in Lindegren et al. (2018), Sect. 2. This problem is completely removed in *Gaia* DR2 since IDU processes the entire data set in bulk and therefore can use all the available data to derive the best calibrations (e.g. bias, background, etc.) to then perform the pre-processing and IPD. Another major improvement with respect to *Gaia* DR1 is in the blacklisting (identification of spurious transits) and cross-match process, which has led to fewer spurious sources and a cleaner set of transits to work with (see Castañeda et al. 2018, for more details). Finally, another important improvement is the handling of micro-meteorites and clanks in the reconstructed attitude (Lindegren et al. 2018), which leads to better intra-window source positions. All of these factors have contributed to a cleaner set of input data with higher quality for the photometric processing.

PhotPipe features a number of improvements in terms of both algorithms and processing flow, as we explain in more detail in Sect. 5. Considerable effort has been dedicated over several years to the development of a software system that is robust and deterministic, but still flexible enough to be able to adapt to the needs of a complex space mission such as *Gaia*. With over 51 billion individual transits, contributing 510 billion individual *G*–band CCD transits and 102 billion low-resolution spectra, achieving high-quality photometry is not only a matter of devising a sound calibration model (Carrasco et al. 2016), but also of implementing it in a scalable and resilient fashion.

The processing of *Gaia* data poses several challenges: (1) the intrinsic complexity of the payload and its operation modes (see Sect. 2) leads to a complex data stream in terms of both raw data and intermediate DPAC data products; (2) the large raw and intermediate data volume (tens of terabytes) and the fine granularity (0.9 billion individual CCD observations per day) pose demanding constraints on data storage, I/O, and processing performance; (3) the iterative nature of the DPAC cyclic processing and of some of the processing algorithms, combined with the requirement of keeping the overall processing time fixed at each iteration, poses a demanding requirement on the scalability of the processing systems. Several years before launch, it became clear that a distributed processing architecture is required to meet these challenges successfully. We selected the Hadoop distributed batch-processing system (e.g. White 2012) for the PhotPipe processing architecture. Hadoop provides a reliable distributed file system and a simple parallelisation abstraction based on the Map/Reduce model (e.g. Dean & Ghemawat 2008) to develop distributed data-processing applications. The adoption of Hadoop as the platform for PhotPipe has proven to be very successful in terms of overall performance and robustness, and it is cost effective in terms of manpower required for operation and maintenance. PhotPipe is the first processing system for a large astrophysical survey project, such as *Gaia*, to make use of these technologies. Additional information on the PhotPipe processing software and the Map/Reduce algorithm implementation is provided in Sect. 6.

Section 2 presents a brief overview of the instrument. Section 3 provides a description of the input data used to produce the *Gaia* DR2 calibrated photometry. Section 4 describes the pre-processing treatment, and Sect. 5 describes the calibration processing flow. Section 6 presents the architecture developed for the PhotPipe processing pipeline and some aspects of the distributed implementation of the photometric calibration workflow, and discusses the performance of *Gaia* DR2 processing. Finally, some concluding remarks and planned developments for the near future are given in Sect. 7. For convenience, a list of the acronyms used in this paper can be found in Appendix D.

2. Instrument overview

Although a comprehensive description of the *Gaia* mission and payload can be found in Gaia Collaboration (2016b), in an effort to make this paper more self-contained, this section provides a summary of the mission and payload aspects that are most relevant to this paper. The *Gaia* astrometric measurement concept is based on HIPPARCOS and involves two viewing directions (telescopes) separated by a large fixed angle (the basic angle). The two fields of view (FoV) are then projected onto a single focal plane. The satellite scans the sky continuously with a fixed revolution period of six hours. The scanning law designed for *Gaia*

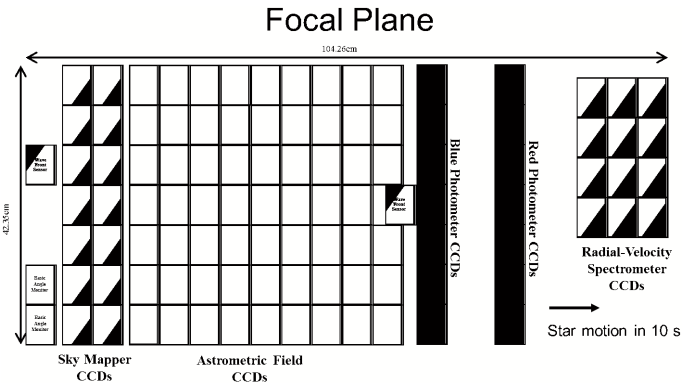

Fig. 1. *Gaia* focal plane, which contains 106 CCDs organised in seven rows. Stellar images travel in the along - scan direction from *left to right*. The 12 CCDs in green are part of the Radial Velocity Spectrometer, which will not be described any further because it is not relevant for this paper.

provides a full sky coverage every six months. The sky coverage is not uniform because some areas (nodes) have a very large number of scans (up to 250 transits per source in the mission nominal length).

Gaia's focal plane uses optical CCDs operated in time-delay integration (TDI) mode. Figure 1 shows a schematic view of the focal plane layout. The CCDs are organised in seven rows. In each row, two CCDs are dedicated to the sky mappers (SM, one per FoV), nine are dedicated to the astrometric field (AF, with the exception of row 4, which uses only eight). There are then two CCDs dedicated to the BP and RP. The satellite scanning direction is aligned with the rows on the focal plane so that a source image will enter the focal plane on the SM CCD appropriate for the FoV in which the source is observed and will then move along the AF CCDs, finally reaching the BP and RP CCDs. A *Gaia* observation of an astrophysical source is called a FoV transit. The crossing time of a source over an individual CCD is approximately 4.4 s, which therefore provides an upper limit for the exposure time of a single CCD observation.

All the CCDs in a given row are controlled by a single video processing unit (VPU) that is responsible for the source detection and confirmation, the definition of the observing mode (see below), and the recording of all the relevant payload and satellite information that is required for the ground-based reconstruction process (Fabricius et al. 2016). The source detection takes place in the SM CCD for each FoV: if the detection is confirmed by AF1 (i.e. the first of the AF CCDs), the VPU assigns a window to the source and determines the observing mode for each of the CCDs in the row. Sources that at detection have an estimated G magnitude of 11.5 or brighter in the SM are automatically confirmed (see de Bruijne et al. 2015). Each CCD observation is acquired by reading a window approximately centred on the source position. The across-scan (AC) position of the window is computed by the VPU for each CCD, taking into account the estimated AC motion over the focal plane (i.e. source images do not travel along a straight line on the CCDs).

A complex gating and windowing scheme is implemented on board to control the effective exposure time and limit the telemetry data volume. Twelve possible gate configurations are available for each *Gaia* CCD. The gate activation strategy is defined in terms of the onboard detected magnitude and can be configured independently for each CCD and AC position. In the current configuration, the AF CCD observations can be acquired without gate or with one of seven different gates. When activated, a gate will affect all windows that are observed on that CCD during the activation time. This can create unexpected gated observations for faint sources as well as complex gate situations, where a gate affects only part of the window. The window samples can be acquired with or without hardware binning and can be further binned to reduce the number of bytes required for the downlink. Detections brighter than $G = 13$ and $G = 11.5$ are assigned a full-resolution 2D window in AF and BP/RP, respectively. Detections fainter than these limits are assigned a 1D window (obtained by binning the 2D windows in the AC direction at read-out). These different configurations are referred to as window-classes.

The BP/RP low-resolution spectra can be acquired either without gate or with one of five active gate configurations. BP/RP windows are 60 samples long in the along-scan (AL) direction. The spectral dispersion of the photometric instrument is a function of wavelength and varies in BP from 3 to 27 nm pixel^{-1}, covering the wavelength range 330–680 nm. In RP, the wavelength range is 630–1050 nm with a spectral dispersion of 7 to 15 nm pixel^{-1}. Because of their larger size, BP/RP windows are more affected by complex gate cases. Furthermore, contamination and blending issues in dense regions will particularly affect BP/RP spectra. Due to their larger size, it is generally not possible for the VPU to allocate a BP/RP window for every detection. Finally, in both AF and BP/RP CCDs, windows can be truncated in case of overlap. A priority scheme is defined to rule this process. These non-nominal observations and those affected by a complex gate activation have not been included in the *Gaia* DR2 photometric processing and therefore have not contributed to the released photometry.

3. Input data

Gaia DR2 is based on 22 months of observations starting on 25 July 2014 (10:30 UTC) and ending on 23 May 2016 (11:35 UTC), corresponding to 668 days. When discussing mission events, it is more convenient to use the onboard mission timeline (OBMT), expressed in units of nominal satellite revolutions (21 600 s) from an arbitrary origin. An approximate relation to convert beween OBMT revolutions and barycentric coordinate time (TCB) is provided by Eq. (3) in Gaia Collaboration (2016b). Hereafter we use *rev* to mean OBMT revolutions. The period covered by *Gaia* DR2 extends from 1078.38 to 3750.56 rev.

There are a number of events that have to be taken into account in the photometric calibration for *Gaia* DR2: two decontamination and two re-focussing events (see Table 1). Decontaminations are required to mitigate the throughput loss caused by water-based contaminant present in the payload (Gaia Collaboration 2016b). The size of the systematic effect due to the contamination is orders of magnitude larger than the expected level of any systematic effect. Decontamination campaigns are required to recover the optimal performance: they involve actively heating the focal plane and/or some of the mirrors to sublimate the contaminant. Refocussing events have been carried out after each decontamination event.

It should be noted that after a decontamination event is completed (i.e. the active heating is turned off), a much longer time is required for the focal plane to return to its nominal operating temperature. Although science operations resume at the end of a decontamination campaign, the data quality will not be nominal until thermal equilibrium has been reached. Data obtained during the time ranges listed in Table 1 have not been included in the photometric processing. Furthermore, these events create

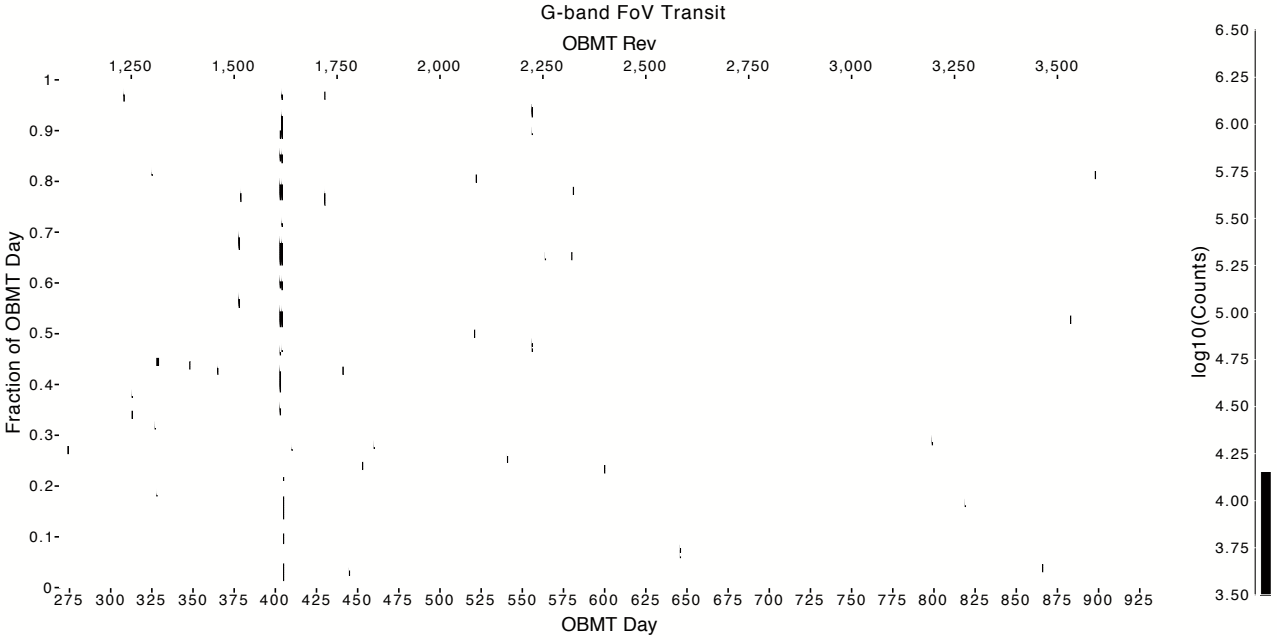

Fig. 2. Temporal density distribution of the ≈51 billion *G*–band observations contributing to *Gaia* DR2. Each column in the heatmap shows the density of observations within a given OBMT day for each OBMT day. The OBMT revolution is shown on the top abscissa axis to facilitate interpretation. The high-density features are the Galactic plane crossing the two FoVs either in the Galaxy inner or outer direction (see the text for more details). The gaps related to the events listed in Table 1 are also visible. Other small gaps are due to telemetry data that could not be included in *Gaia* DR2 because they were affected by processing problems.

Table 1. *Gaia* DR2 mission events relevant for the photometric calibration process.

Event	OBMT range [rev]		Duration
	start	stop	[rev]
Decontamination	1316.492	1324.101	7.609
Refocussing	1443.950	1443.975	0.025
Decontamination	2330.616	2338.962	8.346
Refocussing	2574.644	2574.728	0.084

discontinuities in the instrumental behaviour that can be used as natural breaking points for the definition of the photometric calibrations (see Sect. 5 for more information).

Figure 2 shows the density of *G*–band FoV transits observed by *Gaia* in the time range covered by *Gaia* DR2 (abscissa) with intra-day resolution (ordinate). For a given abscissa position (i.e. one OBMT day), the ordinate shows the density variation within the four OBMT revolutions of that day, thus allowing a much higher level of detail to be visible compared to a standard histogram. Several features are visible, in particular:

- Sixteen daily Galactic plane (GP) crossings: eight in the inner and eight in the outer direction of the galaxy, four for each FoV. The GP features become progressively steeper in the plot because the spacecraft spin axis becomes perpendicular to the GP itself thus leading to a GP scan (GPS) when the two *Gaia* FoVs effectively scan the GP continuously for several days (e.g. at ≈1945 rev and then again at ≈2120 rev, etc.)
- The decontamination events (see Table 1), which manifest as gaps in the data. Refocussing events are harder to spot because their duration is much shorter.
- Outages in the daily processing pipelines, which manifest as minor gaps. These outages meant that some satellite telemetry was not actually available for *Gaia* DR2 processing, but will disappear in future date release. Other gaps are instead caused by genuine spacecraft events and will never disappear.
- The eight thin streaks visible before 1200 rev are due to the LMC crossing the two FoVs at each revolution during the ecliptic poles scanning mode (see below). After this, the LMC is still visible as increased density spots at periodic intervals.

From the start of scientific operations up to 1185.325 rev, *Gaia* observed following the ecliptic poles scanning law (EPSL), which meant that both FoVs were scanning through the north and south ecliptic pole at each revolution (with the scanning direction changing at the same rate as the Sun) and the spin axis moving along the ecliptic at a rate of ≈1° per day. The main aim of this scanning mode was to provide end-of-mission coverage for a limited portion of the sky in a very short amount of time for the purpose of bootstrapping the photometric calibrations and to assess the scientific performance of the mission (see Sect. 5.2 in Gaia Collaboration 2016b, for more information). Unfortunately, the period leading up to the first decontamination proved to be very unsuitable for the purpose of establishing the photometric system as originally planned because of the high level of contamination, and more importantly, because of its strong temporal variation. In Sect. 5 we discuss the implications of this for the flux calibration process. After the EPSL phase, the scanning law was transitioned to the nominal one.

The main inputs to PhotPipe are 1) the IDU pre-processed *G*–band transits, providing centroid, IPD information, and basic acquisition and quality information; 2) the IDU cross-match associating each transit to a source; 3) the source astrometry and reconstructed spacecraft attitude produced by the astrometric global iterative solution (AGIS; Lindegren et al. 2012) system; and 4) the raw BP/RP low-resolution spectra. The IDU PSF (for 2D observations) and LSF (for 1D observations) used

Table 2. Summary of the main input records used by PhotPipe for *Gaia* DR2.

Type	No. records
G–band FoV transit	51,712,381,972
BP/RP raw FoV transit	51,715,475,265
Spurious detections	10,737,486,581
IDU cross-match sources	2,582,614,429
AGIS sources	2,499,375,298

Notes. The spurious detections are a subset of input *G*–band transits and have been excluded from the cross-match because they may be associated with artefacts from bright sources. AGIS sources refer to the number of source records with sky positions determined by AGIS.

for *Gaia* DR2 are similar to those used for *Gaia* DR1 that are described in Fabricius et al. (2016). It is worth mentioning that the PSF/LSF models have been derived from mission data in the range 3350–3365 rev and do not model the colour and time dependencies. This can create systematic effects on the derived fluxes that are time and colour dependent due to the time-varying contamination: these systematics can become more noticeable when handling epoch data, but are less critical for the source photometry, where they will result in increased errors of individual sources.

The *Gaia* onboard detection algorithm (de Bruijne et al. 2015) operates without a source catalogue, which means that the spacecraft telemetry provides only transit-based information: no further knowledge about the association of each transit to a given astrophysical source is available. Associating transits to individual sources is the main goal of the cross-match task performed by IDU (Castañeda et al. 2018). A pre-processing stage identifies spurious detections that are due to artefacts caused by bright sources or extended objects. The cross-match process then associates each individual transit with a source. Although the process reuses the source identifiers that have been created in previous iterations (e.g. *Gaia* DR1 in this case), it should be noted that the source identifier is simply a label: what actually provides an identity to a source are the transits that are associated with it since that will eventually determine the source astrometric parameters and photometry. For this reason, it is not possible to directly compare individual sources between different *Gaia* releases, and *Gaia* DR2 should be treated as a new and independent catalogue. Table 2 summarises the number of input records processed by the PhotPipe system: these represent a superset of the *Gaia* DR2 content as low-quality, incomplete, and/or non-nominal data have been excluded from the release (see Arenou et al. 2018).

4. Pre-processing

As mentioned in the previous section, the raw SM and AF CCD transits are first processed in IDU, which takes care of bias correction, background determination, and removal, and estimates centroid positions and the *G*–band (uncalibrated) flux based on PSF/LSF fitting (see Fabricius et al. 2016). The pre-processing for the BP/RP CCD transits is carried out by PhotPipe instead and is described in this section.

The pre-processing stage is required to prepare the raw integrated epoch fluxes in all bands (at the CCD level) for the calibration step. For all CCD transits, we compute the predicted positions of the image centroid on the CCD from the reconstructed satellite attitude, the geometric calibration (expressed as a correction to the nominal field angles, as described in Lindegren et al. 2012, Sect. 3.4), and the source astrometric parameters as derived by AGIS (Lindegren et al. 2018): this is essentially the inverse of the operation described in Fabricius et al. (2016, Sect. 6.4). Since the AC centroid position of the image is only available for the 2D *G*–band transits, which are ≈1% of the total, the flux calibration models (see Sect. 5) use the predicted AC position as the best available information on the AC location of the source image on the CCD. In *Gaia* DR1, the predicted AC position could not be computed consistently for all transits, and the calibration models therefore used the AC window centre, which is equivalent to assuming that each source image is perfectly centred in the AC direction. The BP/RP integrated fluxes and spectrum shape coefficients (SSCs, see Carrasco et al. 2016) are produced by PhotPipe from the BP/RP spectra after they have undergone several pre-processing steps: correction and mitigation of electronic offset effects (Sect. 4.1), background correction (Sect. 4.2), and AL geometric calibration (Sect. 4.3).

4.1. Correction and mitigation of electronic offset effects

The electronic zero-point offset on the CCD amplification stage (commonly referred to as the bias level) is in principle separable from nearly all other calibrations. However, the complexity of the *Gaia* CCD design and operation leads to quasi-stable behaviour that in turn considerably complicates the determination of the additive correction to be applied to the data at the beginning of the processing chain (Hambly et al. 2018). In addition to the normal zero point of the digitised sample data (which in the case of *Gaia* is measured via periodic prescan samples), offset excursions are present on any given data with amplitudes of up to ≈16 ADU (≈64e^-) in BP/RP depending on the timing of that sample in the serial scan and on the number (if any) of fast-flushed pixels preceding the sample. Furthermore, the onset of the excursions and recovery as normal samples are read is a non-trivial function of the flushing, reading, and occasional pausing in the serial scan (exhaustive detail is given in Hambly et al. 2018). Hence the full mitigation of these electronic effects involves effectively reconstructing the readout history of the CCD in a window of ≈30 s centred on each detection. For *Gaia* DR2, all the required calibrations are determined in the IDT and in the First Look CCD one-day calibration subsystems, but the process of determining the correct bias level for each sample still requires readout reconstruction from observation log records that are telemetered as part of the auxiliary data streams into the on-ground processing pipelines.

Figure 3 shows an example of the effectiveness of the offset instability correction procedure for the BP CCD in row 3 of the *Gaia* focal plane. This device shows the largest excursions from the gross electronic zero point amongst the astro-photometric devices. For this illustration we have chosen samples that have been affected by a gate 5 activation. We note that these are not samples from windows containing objects that have triggered a gate 5 activation; we have chosen instead samples from empty windows (also known as "virtual objects", VO) that are observed at the same time as such a window containing a very bright star, but at different AC positions within the same CCD. The integration time of these samples is limited to 32 ms, resulting in very small photoelectric background correction and hence no possibility of significant residual systematic errors from that correction, accurate calibration of which also depends on bias correction, of course. These selected samples are the closest approximation we can achieve to "dark" observations in *Gaia*,

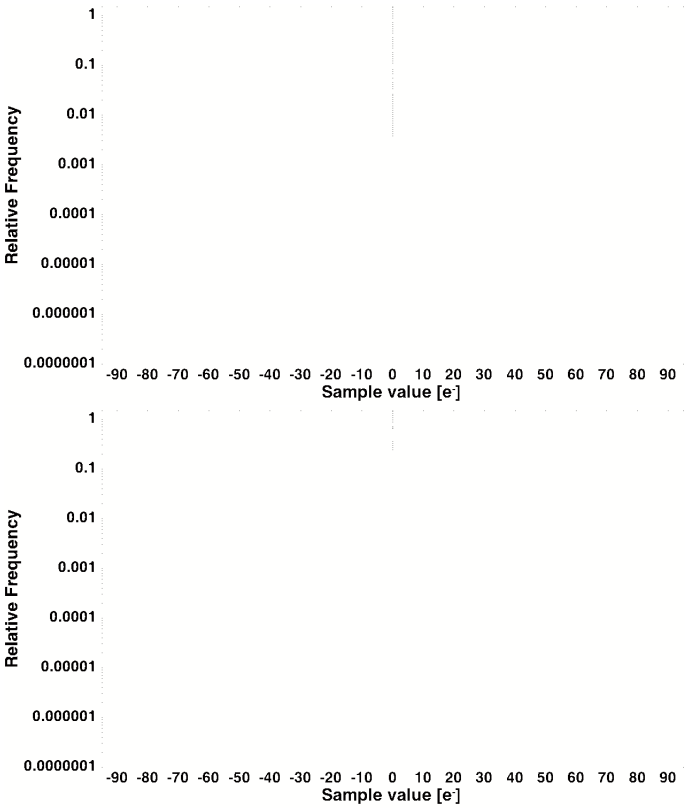

Fig. 3. Sample distribution in empty windows affected by gate 5 activation (and hence limited in integration time to 32 ms) when they are bias corrected by scalar prescan level only (*top*) and when they are bias corrected using the full bias instability model (*bottom*). The sample distribution in the latter has a near-normal distribution (disregarding the negligible few outliers on the positive side of the distribution resulting from stray photoelectric flux and prompt-particle events, etc.) with Gaussian equivalent $\sigma = 5.2e^{-}$ and is dominated by the video chain read-noise-limited performance for that device (see Hambly et al. 2018 for further details). In both panels the magenta line shows the cumulative distribution.

which scans continuously with no shutter. The distribution of sample values corrected for prescan level only shows a high-amplitude systematic residual pattern that is introduced by the offset instability excursions resulting from the multifarious sample serial timings as the observed windows transit the CCD. The core distribution of samples corrected with the full bias model is, however, limited to a near-normal distribution equivalent to the distribution expected given the video chain detection noise performance. We note that a relatively small number of samples remain uncorrected in this example, which features data from 1973–2297 rev. These arise in *Gaia* DR2 in situations where the data stream is incomplete and on-ground readout reconstruction is consequently inaccurate. This problem affected *Gaia* DR1, is significantly reduced in DR2, and will be further reduced in DR3.

4.2. BP/RP straylight mitigation

As has been reported for *Gaia* DR1, the large-scale background has a major contribution in all instruments from light scattered by loose fibres on the solar shield edges that enter the FPA via illegal optical paths (Fabricius et al. 2016, Sect. 5.1.3). However, cosmic sources also contribute significantly. The background mitigation for the G-band is performed in IDU and involves

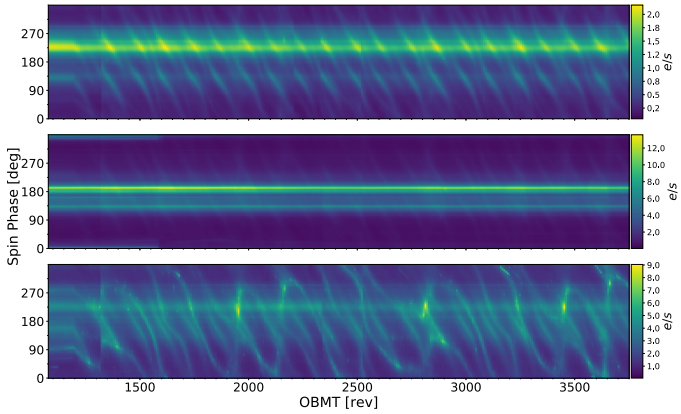

Fig. 4. Straylight background level evolution in BP. Each panel show the straylight level as a function of time and satellite spin phase. *Top panel*: BP row 1, which has the lowest overall straylight level. *Central panel*: BP row 7, which has the highest straylight level, but shows a very stable pattern. *Bottom panel*: BP row 5, which has a higher and extremely variable straylight level. See the text for further discussion.

fitting a 2D spline as a function of time and AC position for each CCD. The variation of the straylight with time in SM/AF is therefore captured reasonably well.

In both *Gaia* DR1 and DR2, the BP/RP background mitigation performed by PhotPipe primarily involves the determination of the straylight component. The straylight pattern depends on the spin phase of the satellite and is stable over several tens of revolutions. Instead of explicitly modelling the time dependence of the straylight pattern for each CCD, an independent solution is determined on every set of consecutive ≈8 rev time intervals in the *Gaia* DR2 dataset (excluding the events in Table 1). The calibration uses the VO empty windows, which are allocated by each VPU according to a predefined spatial and temporal pattern. For each VO, PhotPipe determines the median level and constructs the AC versus spin-phase straylight map by taking the median level from all contributions in each AC/phase bin. For *Gaia* DR2, the resolution of the maps was ≈100 pixels in the AC directions (20 bins AC) and 1° in the phase direction (360 bins in phase).

The resulting maps can occasionally contain gaps (i.e. empty bins) caused by missing data or gaps in the reconstructed attitude. To reduce the impact of gaps, the maps are processed on the fly to fill the gaps via interpolation. The process first attempts to fill the gaps by interpolating along the phase and then fills any remaining empty bins by interpolating along the AC dimension. In both cases, we used linear interpolation by searching the nearest non-empty bin within a configurable range (four bins for phase and three bins for AC). Even after this interpolation process, it is possible for empty bins to be present in the case of very large gaps. These bins are assigned a default value equal to their nearest phase bin and are flagged to ensure that they will not be used by the background level estimation process[1]. The straylight level to be removed from each transit is then determined from the appropriate pre-processed map via bicubic interpolation.

One effective way to visually evaluate the stability of the straylight background over time is to create an animation from the individual straylight maps. An alternative approach is shown in Fig. 4, where we generate an average straylight profile (level versus spin phase) from each map and then display all

[1] This stage is required because empty bins will otherwise confuse the bicubic interpolator used for the straylight level estimation process.

the profiles as a function of OBMT revolution. Three cases are shown to illustrate the challenges faced in mitigating the straylight background. The top panel shows BP row 1: the CCD least affected by straylight, peaking at nearly 3 e^- s^{-1}, the main peak is located at a phase of ≈225°, and its location appears to be quite stable over time. A secondary, fainter feature is visible at a phase of ≈130°. Both features are very stable during the EPSL period, while they appear to progressively drift in phase when the satellite follows the nominal scanning law. The central panel of Fig. 4 shows the straylight evolution for BP row 7: the CCD with the highest level of straylight. Although the level is much higher than in row 1 (see the different range covered by the colour scale in the corresponding plots), the pattern is very stable for all of the three main features: the brightest at phase ≈190°, and two fainter features at phase ≈135° and ≈170°. Finally, the bottom panel shows BP row 5, where the overall straighlight level is not as strong as in row 7, but shows a very complex temporal evolution. There are four peaks at phases ≈90°, ≈150°, ≈225°, and ≈280° that are stable during the EPSL period, but then appear to drift along the entire phase range in a cyclic fashion, creating a complex pattern. The peak at ≈225° also appears to maintain a stable component over the entire time range.

4.3. BP/RP AL geometric calibration

Although low-resolution spectral data are not part of *Gaia* DR2, some aspects of the BP/RP spectral processing are very important in the generation of the photometric catalogue and should therefore be described in this paper. As in the case of *G*–band observations, spectra are also collected in small windows centred around the detected sources. The incoming light is dispersed in the AL direction by a prism. However, the flux at each wavelength is additionally spread over a range of samples according to the LSF appropriate for that wavelength. This means that the flux collected in each sample will have contributions from a range of wavelengths whose width depends on the FWHM of the LSF at the various wavelengths. The size of these contributions depends on the source spectral energy distribution.

Several calibrations are required to convert the acquired flux per sample into the flux in a band covering a specific wavelength range. While for the generation of integrated $G_{\rm BP}$ and $G_{\rm RP}$ a simple sum of the flux in all the samples of a window is sufficient, an accurate calibration of the AL coordinate within the window, in terms of absolute wavelength, is required to extract more detailed colour information in the form of SSCs (see Carrasco et al. 2016, and Appendix A). The dispersion calibration provides a relation between the AL coordinate within the window and the absolute wavelength scale. The nominal pre-launch dispersion calibration has been adopted for *Gaia* DR2. This was derived from chief-ray analysis from fitting a polynomial function to the unperturbed EADS-Astrium *Gaia* optical design with a maximum fit uncertainty of 0.01 AL pixel (Boyadjian 2008). However, the polynomial function is defined with respect to the location within the window of a specific reference wavelength, chosen to be well centred in the instrument wavelength range, and therefore its application is complicated by the fact that sources are often not well centred (due to inaccuracies in onboard detection window assignment). The location of the source centroid within the window can be predicted using our knowledge of the source astrophysical coordinates, the satellite attitude, and the layout of the CCDs in the focal plane, i.e. their geometry. Astrophysical coordinates and satellite attitude are best calibrated using the *G*–band data in the AGIS system (Lindegren et al. 2012), while the geometry of the BP/RP CCDs is calibrated as part of the PhotPipe pre-processing.

The AL geometric calibration is computed differentially with respect to the expected nominal geometry based on pre-launch knowledge of the CCD layout. An initial guess of the source location within the window is obtained by adopting the nominal geometry. The calibration process aims at modelling the corrections to be applied to the nominal predicted positions. For more details on the calibration procedure and on the model definition, see Carrasco et al. (2016).

Figure 5 shows the epoch spectra available for one of the SPSS (Pancino et al. 2012) used in the external calibration. This source was chosen because it is quite bright and has a large and well-distributed number of epochs. The top row shows the epochs calibrated using only our nominal knowledge of the CCD geometry (BP spectra are shown on the left and RP spectra are shown on the right). The bottom row shows the same epochs after the application of the calibration produced by PhotPipe. In all panels of Fig. 5, the location of each sample and the corresponding flux have been shifted and scaled respectively according to the differential dispersion across the focal plane. This creates an internal reference system that is referred to as a pseudo-wavelength scale. Figure 6 shows the actual calibration evaluated at different locations (CCD edge cases in the AC direction are shown with dashed and dotted lines, while the value in the centre of the CCD is shown with a solid line) on the various CCDs (in different colours, red corresponding to row 1 and purple to row 7) for BP in the top panel and RP in the bottom panel at various times in the period covered by *Gaia* DR2.

When comparing Figs. 5 and 6, it is easy to see that the calibrations can reproduce the significant offsets observed for approximately simultaneous spectra from different FoVs. For instance, epochs in the period 3000–5000 rev (colour-coded in blue in Fig. 5) show an offset between the two FoVs of several AL pixels and a wide separation in the calibrations for the two FoVs. The difference between the two FoVs is much smaller in other periods, hardly noticeable, for instance, in the period 1750–2000 rev (colour-coded in yellow in Fig. 5), which is confirmed by the calibrations evaluated in the same period. Discontinuities in the calibrations shown in Fig. 6 are clearly related to decontamination and refocus activities (see Table 1). In general, the RP calibration is noisier because the features in the RP spectrum are smoother than those in the BP spectrum. The standard deviation of the solution evaluated in the period 2700–3700 rev is 0.05 for BP and 0.12 for RP in AL pixels. These are equivalent to 0.4 nm and 1.3 nm, respectively, in the central part of the spectrum. Uncertainties of this size are negligible when computing the spectrum shape coefficients used for the photometric calibrations (for more details, see Appendix A).

Systematic errors in the geometric calibration parameters would not affect the photometric calibrations as they will simply result in a slightly different set of SSC bands being used for the definition of the colour information. After dispersion and geometry calibrations have been carried out, it becomes possible to estimate the flux in given passbands, such as those used to define the SSCs. We call these "raw" or "uncalibrated" SSCs, even though the calibration process described above has been applied in order to generate them. The fluxes obtained at this stage will still be affected by differences in the CCD response and in the LSF across the BP/RP strips of CCDs. None of the spectra shown in this paper are calibrated for response and LSF. These effects are calibrated out in the internal calibration step. For *Gaia* DR2, the uncalibrated SSCs and integrated $G_{\rm BP}$ and

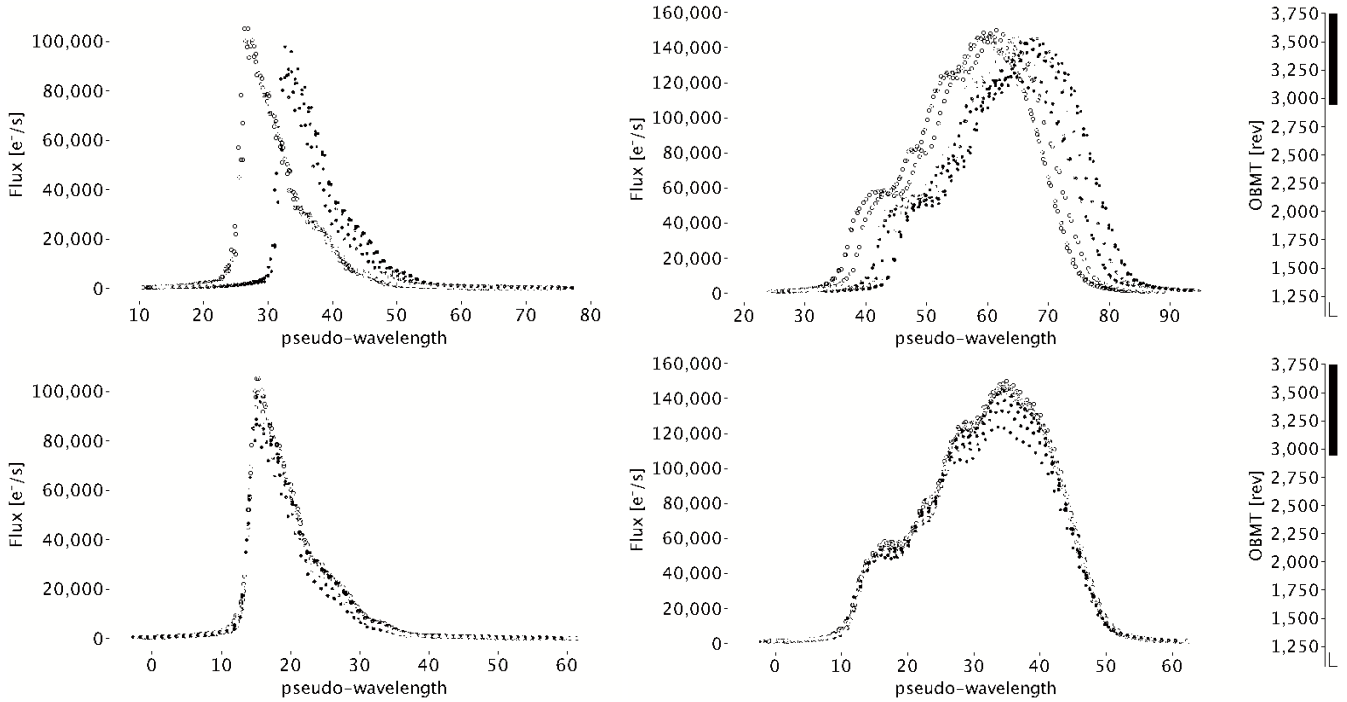

Fig. 5. Epoch spectra for one of the SPSS sources. BP and RP are shown in the *left and right panels*, respectively. The *top row* shows the epoch spectra aligned using the nominal geometric calibration only. The *bottom row* shows the same epoch spectra after application of the differential geometric calibration computed by PhotPipe. Filled symbols are used for the preceding FoV, while open symbols show the following FoV. Symbols are colour-coded by time in OBMT-rev as indicated by the colour bar.

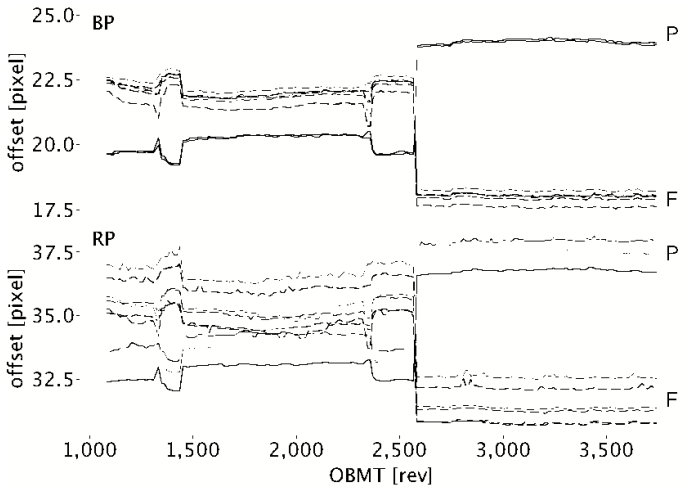

Fig. 6. Evolution in time of the geometric calibration, relative to the nominal geometry, evaluated at the centre of the CCD in the across-scan direction and for different CCDs (BP in the *top panel*, RP in the *bottom panel*; rows from 1 to 7 are shown in red, orange, yellow, green, blue, purple, and pink). The preceding FoV is shown with brighter colours and solid lines, while darker shades and dashed lines are used for the following FoV, as indicated by the labels P and F in the plot area.

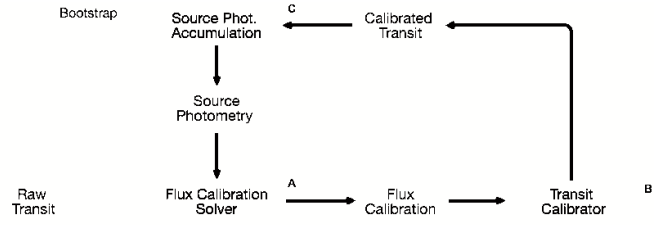

Fig. 7. Iterative internal calibration flowchart. The process is started by bootstrapping the reference source photometry from the raw transits, and it then proceeds by iteratively deriving new flux calibrations, which are then used to produce an updated set of reference fluxes. The calibration loop is represented by the three processes labelled A, B, and C; the dataflow is represented by thicker arrows. See the text for additional information.

G_{RP} are calibrated following the same procedure as applied to the G–band and is described in Sect. 5.

5. Calibration of *Gaia* integrated photometry

5.1. Overview

The calibration of the G–band and BP/RP integrated photometry is based on the principle of first performing a self-calibration on an internal system using only *Gaia* data, followed by an external calibration to link the internal to the external system (Carrasco et al. 2016). The internal calibration workflow is illustrated in Fig. 7 and involves establishing the internal *Gaia* photometric system as defined by a set of standard sources with defined reference G–band and BP/RP integrated fluxes. These standards are then used to derive the set of photometric calibrations required to calibrate all individual epochs. These calibrated epochs are then combined to derive the source photometry in the internal photometric system. Since the reference fluxes for the standard sources are not known a priori, they are derived via a simple iterative bootstrap procedure from the uncalibrated source photometry. Each step is described in greater detail below:

1. Compute the raw source photometry from uncalibrated epochs to use as starting values for the reference fluxes. This step is represented in Fig. 7 by the process labelled

C, operating on raw transit input as represented by the thin dashed line labelled "bootstrap".
2. Derive a set of calibrations based on the current set of reference fluxes. This step is represented in Fig. 7 by the process labelled A, operating on the raw transits and the reference fluxes for the corresponding sources. The calibration model is described in more detail in Carrasco et al. (2016).
3. Apply the calibration to the individual epochs. This step is represented in Fig. 7 by the process labelled B, operating on the raw transits and using the calibration derived at the previous step to produce the calibrated epoch photometry.
4. Recompute the source photometry to provide an updated set of reference fluxes. This step is labelled C; it is the same as the first step, but operating this time on calibrated transits instead of raw ones.
5. Iterate by repeating steps from two to four until convergence is reached. This is the calibration loop that is shown in Fig. 7 as the three processes labelled A, B, and C; the dataflow is represented by thicker lines.

The calibration model is composed of a large-scale (LS) and a small-scale (SS) component. The LS component tracks the fast changes in the instrument over timescales of a few revolutions (≈one day), whereas the SS component tracks more stable sensitivity variations and effectively provides a 1D flat-field equivalent. Instead of explicitly modelling the time dependence in the LS calibrations, we simply computed a number of independent solutions on ≈4 rev time ranges spanning the *Gaia* DR2 dataset, but excluding the decontamination and refocussing events listed in Table 1. The LS model used for *Gaia* DR2 features a quadratic dependency on the AC position of the transit and a linear dependency from the source colour. The colour information is expressed in terms of the spectral shape coefficients (SSC), which are derived by integrating the BP/RP spectra on a predefined set of top-hat bands providing four integrated fluxes from the BP spectrum and four from the RP spectrum (see Sect. 4 in Carrasco et al. 2016). The main advantage of using SSC-based colours is that it allows the use of lower order dependencies in the calibration models than when using a plain (e.g. $G_{\rm BP} - G_{\rm RP}$) colour by providing more detailed colour information. However, for the SS calibration, the *Gaia* DR2 model simply involves a zero point.

5.2. Robustness

When computing the least-squares (LSQ) solutions for the LS and SS models, we exclude from the solutions non-nominal observations, that is, observations that are either truncated or that have been acquired with a complex gate configuration. In addition, we filter observations that have been flagged as problematic in the acquisition or IPD processes: these observations are tagged with the corresponding set of problems and are then used to generate a report, attached to each individual solution, describing how many observation have been excluded and for which reasons. One additional filter is used to exclude outliers that might originate from cross-match problems by excluding any observation exhibiting a difference between the transit flux and the source reference flux larger than 1 magnitude. These filters are only applied when solving for the calibrations. A different robustness process handles the rejection of unsuitable epochs when generating the calibrated source photometry, as described in Sect. 5.6.

Each LSQ solution is computed iteratively: at a given iteration, we use the solution computed at the previous iteration to reject observations that are discrepant by more than $N\sigma$. At each iteration, the rejection process will evaluate the residuals of all measurements (including those that were rejected in a previous iteration). In *Gaia* DR2, the rejection process has been configured with a 5σ rejection threshold and a maximum number of ten iterations. The rejection process is attempted only if there are at least 20 observations contributing to the solution. This approach requires all available observations to be kept in memory during the iteration process: since the calibration models have a low number of parameters, this is never a problem, even when there are millions of observations contributing to a given calibration solution.

5.3. Time-link calibration

In the early PhotPipe test runs after the start of nominal operations (November 2014), it was discovered that a time-dependent level of contamination was causing linear trends of ≈0.0023 mag/day in the EPSL epoch photometry produced by PhotPipe. This linear trend was caused by the varying level of contamination that affected the data. Contamination introduces a systematic offset in the bootstrap reference flux of a given source, which is a function of the time distribution of the individual transits and the source colour (since the size of the systematic effect caused by contamination on a given transit depends on the colour of the source). This systematic offset is imprinted on the reference fluxes and is not efficiently removed by the iterative calibration loop described above. When solving for the various LS calibrations, this effect causes over-/under-corrections resulting in the linear trend reported in the test campaign. With additional iterations, the linear trend was reduced to ≈0.0021, ≈0.0018, and ≈0.0016 mag/day, thus showing a rather slow decrease in the size of the effect. In order to mitigate for this systematic effect without requiring a large number of iterations, we introduced a new calibration that tracked the differential contamination level as a function of time. The model fits the magnitude differences of epochs of a given source as a function of time and source colour using a cubic time dependence and a linear colour-time cross term, as explained in more detail below.

To measure the throughput loss due to contamination, we could compare the observed raw flux of sources to their known true flux from other catalogues. However, to avoid introducing uncertainties and systematic effects due to passband differences and colour transformations, we preferred not to use any external catalogue. We instead devised a method to recover the variation in throughput with an arbitrary constant defining the real throughput at a given time. This method allowed us to recover the throughput evolution using only *Gaia* data.

We assume the throughput loss, in magnitiudes, to be a function of time t and source colour C and express it using Chebyshev polynomials T_n as basis functions, allowing for cross-terms between time and colour:

$$\tau(t, C) = \sum a_n T_n(t) + \sum b_m T_m(C) + \sum c_m T_m(t) T_m(C). \quad (1)$$

We can thus express the observed variation in throughput between two observations of the same source k as the difference between the throughput function τ evaluated at the two times t_i and t_j,

$$\Delta\tau(t_i, t_j, C_k) = \tau(t_i, C_k) - \tau(t_j, C_k). \quad (2)$$

From the definition above, it is clear that the $\Delta\tau(t_i, t_j, C)$ polynomial does not have a zero point or a linear colour term as the

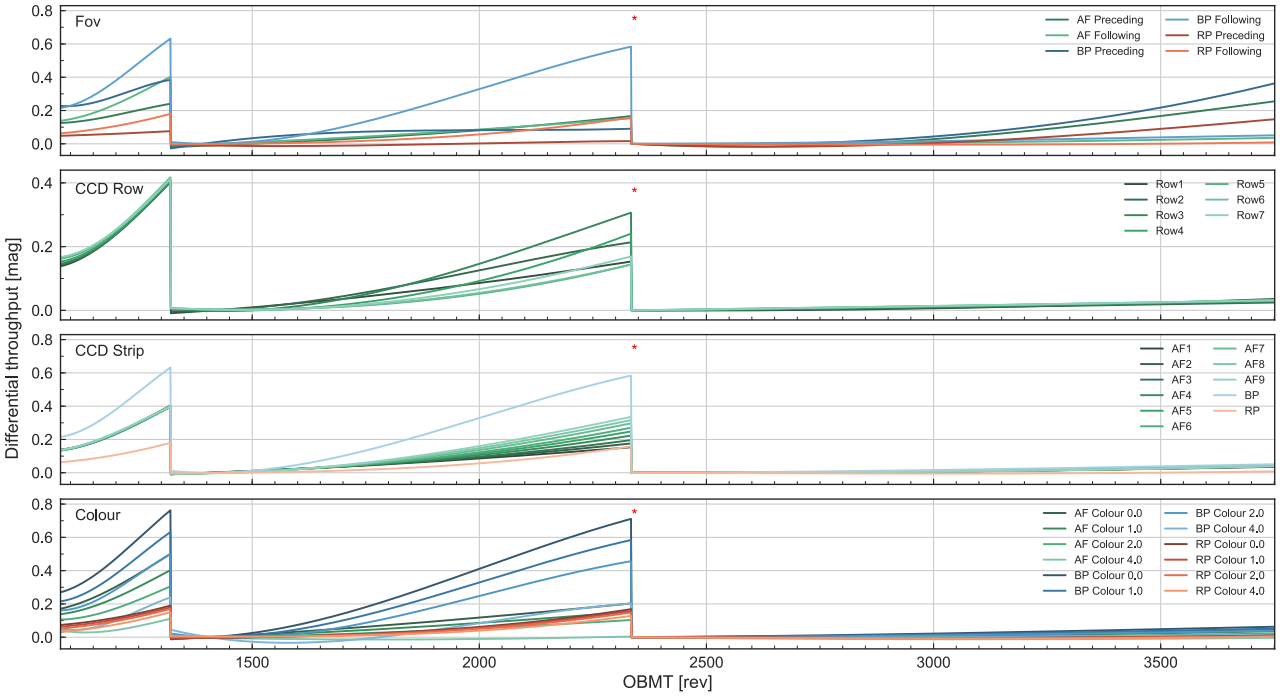

Fig. 8. Differential throughput $\Delta\tau$ (see Eq. (3)) w.r.t. 2335 rev as a function of time, the reference epoch is indicated by the red asterisk. *First panel*: variation in throughput for the two FoV in CCD row 1 for AF1 (green), BP (blue), and RP (red) showing larger contamination in the following FoV at least up to the second decontamination. *Second panel*: variation in throughput for AF1 and following FoV for different CCD rows showing stronger contamination in the bottom part of the focal plane (low CCD row number). The range of the ordinate axis in this panel is more compressed since it does not show BP and RP. *Third panel*: variation in throughput in the following FoV for all CCDs in row 1 showing an increase in contamination from AF1 to AF9; the maximum effect is in BP and the lowest effect in RP, as expected. *Fourth panel*: variation in throughput in the following FoV, CCD row 1 for AF1, BP, and RP and different source colours showing that bluer sources are more heavily affected by contamination, hence the overall much larger systematic in the BP band.

corresponding terms cancel out. For the throughput function τ, we consider a cubic time dependence ($n = 0, \ldots, 3$), a linear colour, and time-colour dependence ($m = 1$). Since during the period of interest there are two decontaminations, the full time-line can be modelled piece-wise by having a $\Delta\tau$ function for each of the three time ranges plus a discontinuity in time and colour at each decontamination event: $K_n(t_i, t_j, C) = d_o + d_1 T_1(C)$. This leads us to the following formulation:

$$\begin{aligned}\Delta\tau(t_i, t_j, C_k) = \Delta\tau_0(t_i, t_j, C_k) + K_1(t_i, t_j, C_k)\\ +\Delta\tau_1(t_i, t_j, C_k) + K_2(t_i, t_j, C_k)\\ +\Delta\tau_2(t_i, t_j, C_k).\end{aligned} \quad (3)$$

When producing the least-squares solution for Eq. (3), a pair of observations of a given source k only contributes to the $\Delta\tau$ polynomials containing one of the two observations, and it will only activate the discontinuity function K_i if the two observations are separated by the ith decontamination. Given a set of observations of a source, there are many different ways to form pairs of observations: we found that a good coverage of the time range span by the observations is created by interleaving the N observations: $(z_i, z_{N/2+i})$, where $i = 0, \ldots, N/2$.

To solve for the differential throughput function, we need sources with observations providing a good coverage of the time-line: the simplest way to achieve this is to select in each HEALPix pixel (of level six, see Górski et al. 2005) the N sources with the most observations (we used $N = 20$). To limit the total number of sources to a manageable level, we selected sources in the (uncalibrated) magnitude range $G = [13.0, 13.5]$ and introduced a colour restriction to the (uncalibrated) range $G_{\rm BP} - G_{\rm RP} = [0.0, 4.0]$ to normalise the colour for use with the Chebyshev basis.

Figure 8 shows the variation in contamination with respect to a reference epoch of 2335 rev (i.e. shortly after the end of the second decontamination, as indicated by the red asterisk) since it seems safe to assume that the overall contamination is at its lowest absolute level at this time. For all top three panels, we considered a source colour of $G_{\rm BP} - G_{\rm RP} = 1.0$. The first panel (top) shows that the contamination level is stronger in the following FoV, at least up to the second decontamination (higher throughput loss), and that it is much stronger in BP than in RP, as expected from the wavelength dependence of the contamination. The second panel shows that the contamination is generally stronger at the bottom of the focal plane (lower row number, see Fig. 1) using the AF1 CCD and the following FoV calibrations. The third panel shows that the contamination increases along CCD row 1, and the effect is stronger in AF9 for the G–band. The fourth (bottom) panel shows the colour-dependence of the throughput in the following FoV for CCD row 1 in AF1, BP, and RP using a source colour $G_{\rm BP} - G_{\rm RP} = 0.0, 1.0, 2.0$, and 4.0.

In *Gaia* DR1, the introduction of this new link calibration reduced the linear trend in the EPSL to 0.00008 mag/day, but it did not completely remove it. The reason probably was that although the model provides a reasonable approximation, it is not sophisticated enough to reproduce all the systematic effects caused by contamination. This is especially true for the EPSL period, when the contamination level was both most intense and showed the strongest time-variation (contamination was higher during the commissioning phase, but this paper is

only concerned with the observations obtained during science operations). Section 5.5 discusses our improved mitigation of contamination for *Gaia* DR2.

5.4. Gate window-class link calibration

The calibration process is complicated by the multitude of instrumental configurations with which observations can be acquired. Each configuration is effectively a different instrument, and for a given time range, the PhotPipe system therefore produces a set of calibrations, one for each instrumental configuration. For simplicity, we call these configurations calibration units (CU). For the LS, a CU is identified by the FoV, the CCD row and strip (i.e. a given CCD), the active gate, and the window class. This leads to a total of 2108 CUs for each time interval. As mentioned in Sect. 2, it is possible that some transits are acquired with a non-nominal configuration when a higher priority simultaneous transit triggers a gate activation. A total of 1848 possible non-nominal configurations exist, which means that in a given time interval, there will be at most 3956 individual LS calibration solutions.

For the SS, a CU is identified by the FoV, the CCD row and strip, the active gate, and a 4-pixel-wide AC bin. Instead of explicitly modelling the high-frequency spatial variations of the AC CCD response (e.g. due to bad/hot columns), we map the AC 1D flat field by computing an independent solution in equally sized bins of 4 pixels. This leads to a total of 1 120 416 nominal CUs per time interval, which increases to 2 332 704 CUs when all the possible non-nominal configurations are considered.

The LS and SS solutions are derived independently for each CU. The fact that sources are observed multiple times in different configurations ensures that CUs are linked together (since all solutions are computed using the same set of reference source fluxes), and therefore, the internal photometric system is homogeneous over the entire instrument. Unfortunately, this is not true for all CUs. Owing to the combination of narrow gate-activation magnitude ranges, the small number of bright sources available and small uncertainties in the onboard magnitude detection, there is insufficient mixing between some CUs. The iterative process used to establish the photometric system should be able to take care of this, but the convergence could be very slow. To speed this process up, an additional link calibration has been introduced for G–band and BP/RP integrated photometry. This calibration provides the link between the different window-class and gate configurations and is applied only at stage 1 in the calibration process (see Sect. 5.1) when the initial set of raw reference fluxes that are used to bootstrap the iterative calibration process described above is derived. The links are computed from multiple observations of the same source in different configurations.

In *Gaia* DR1, a single set of calibrations was computed using ≈10 rev, and this was then used to calibrate the entire dataset. The results were not optimal, as revealed by the features in the errors of the final source photometry (see Sect. 7 in Evans et al. 2017). For *Gaia* DR2, we computed a set of calibrations for each week using ≈8 consecutive revs, therefore calibrating possible time variations for this effect.

5.5. DR2 calibration strategy

In *Gaia* DR1, the calibration process described in Sect. 5.1 was applied to the entire dataset: this was necessary to ensure a sufficient number of good-quality sources to establish the internal photometric system. Since *Gaia* DR2 spans nearly two years, it provides a much better sky coverage than DR1. This allowed us to be more selective in which data to use for the initialisation of the photometric system. In particular, it is clear from Fig. 8 that in the period after the second decontamination, the throughput loss is much lower and more stable. This period spans ≈354 days and therefore provides nearly two complete sky coverages. We therefore decided to use this subset of data, which we call the INIT dataset/period, to initialise the photometric system following the procedure outlined in Sect. 5.1. We refer to the subset from the beginning of operations up to the second decontamination as the CALONLY dataset/period; the motivation for the name is described below.

Using all observations in the INIT dataset, we generated the uncalibrated source photometry from which we selected the sources to be used to compute the time-link calibration. Although the effect of contamination is much reduced in the INIT dataset, the variations in throughput are non-negligible, and therefore it is still appropriate to perform the time-link calibration. The main difference with respect to the model described in Sect. 5.3 is that only a single period is required as there are no discontinuities to take into account. The time-link calibrations were then applied to the entire INIT dataset to generate a new set of reference source fluxes that were used to solve for the gate window-class link calibration as described in Sect. 5.4. Finally, we computed a new set of reference source fluxes by applying both the time-link and gate window-class link calibrations to all applicable observations in the INIT dataset. This provides an improved set of reference fluxes to bootstrap the internal photometric calibration, as described in Sect. 5.1 and shown in Fig. 7.

The next stage is the calibration loop, in which we iteratively solve for the LS calibration and then produce an updated set of reference source fluxes to be used in the subsequent iteration. We performed a total of five iterations and refer to the process as the "LS iterations". The final stage of the initialisation of the internal photometric system involves the introduction of the SS calibration (see Sect. 5.1 and also Sect. 4 in Carrasco et al. 2016). In this stage, we iterate between the LS and SS calibrations without updating the reference source fluxes. An iteration is composed of two steps: in the first, we solve for the SS calibration using LS-calibrated observations; in the second, we solve for the LS calibration using SS-calibrated observations based on the SS calibrations obtained in the first step. We performed two of these iterations. The final internal photometric system is then established by generating a new set of reference source fluxes by applying the last set of SS and LS calibrations. A single set of SS calibrations (composed of 1 749 013 independent solutions) was computed using the entire INIT dataset: we had indeed already confirmed in *Gaia* DR1 that the SS calibration is very stable and has no significant variations over a timescale of one year (Evans et al. 2017, 2018).

Figure 9 shows the mode of the error distribution on the weighted mean G–band magnitude as a function of G magnitude derived using the INIT dataset for the source photometry generated from uncalibrated observations, observations calibrated using the first LS solution from the initialisation loop (see above and Sect. 5.1), and the observations calibrated using the final set of SS and LS solutions. As expected, the introduction of the LS calibration is a great improvement, but further improvement due to the subsequent iterations in the initialisation loop and the introduction of the SS calibration is also quite noticeable, especially at $G < 16$. The large scatter at $G < 6$ is mainly due to saturation, whereas the various bumps are caused by a

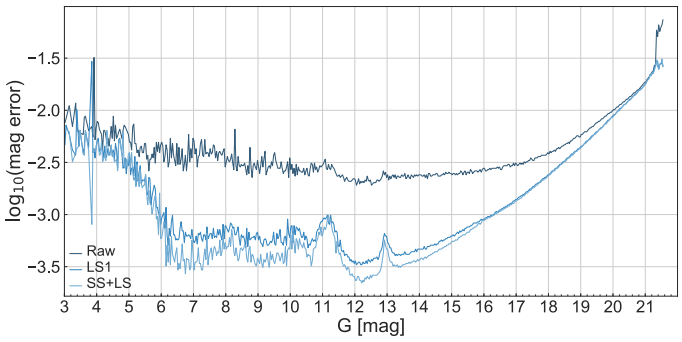

Fig. 9. Mode of the error distribution on the weighted mean G–band magnitude as a function of G magnitude for (1) the uncalibrated source photometry (Raw), (2) the source photometry obtained after the first LS solution in the calibration intialisation loop described in Sect. 5.1 (LS1), and (3) the source photometry obtained after the last SS+LS iteration. The increased scatter at $G \leq 13$ is related to changes in the observation configuration (e.g. window class/gates), the limitations in the IPD algorithms, and the handling of saturation and flux loss. A more in-depth discussion can be found in Evans et al. (2017, 2018).

combination of changes in the instrumental configuration (window class, gate) and the limitations in the PSF/LSF models used in *Gaia* DR2. We refer to Evans et al. (2017, 2018) for a detailed analysis of the error properties of the *Gaia* photometry and a discussion of the various features in the error distributions.

The set of source reference fluxes generated from the INIT dataset can now be used to produce the LS and SS calibrations for the CALONLY dataset. This involves two SS-LS iterations in the same fashion as for the INIT dataset: a single set of SS calibrations was also derived for this period. We then perform one final SS and LS calibration on the INIT dataset to have a consistent set of SS and LS calibrations for both the INIT and CALONLY datasets based on the same photometric system. The linear trend in the EPSL (see Sect. 5.3) period caused by varying contamination has now been further reduced from 0.00008 mag/day of DR1 to 0.000015 mag/day in *Gaia* DR2: this amounts to 0.4 milli-magnitudes over the 28 days of EPSL.

Figure 10 provides an example of the time evolution of the standard deviation and zero point of the final LS calibrations. The most striking features are the decrease in the overall standard deviation "floor" after each decontamination event and the remarkable agreement between the calibration zero-point time evolution in the CCD rows and FoV and what was independently measured via the time-link calibration (see Sect. 5.3 and Fig. 8). We also note that in the period leading to the first decontamination, the standard deviation of the solutions progressivily deteriorated as the contamination built up. This trend is only marginally visible in the preceding FoV (top panel) in the central period between first and second decontamination. This is explained by the fact that the PSF/LSF models were generated using 15 rev in the period after the second decontamination and therefore become progressively worse at representing the data when the contamination level increases and varies. In the period after the second decontamination, the following FoV calibrations are extremely stable, whereas in the preceding FoV, it is possible to see a significant correction at the time of the second refocus (≈2750 rev) and the more pronounced throughput loss in the lower rows caused by the increase in contamination level in the period. Occasional δ-functions such as spikes in the standard deviation are due to individual calibration solutions that are affected by an anomalously large number of poor observations that are mainly caused by sub-optimal calibrations (e.g. background) used in the IPD process and are not a cause of major concern since they are naturally taken care of by the DPAC iterative processing.

5.6. Source photometry

The set of LS and SS calibrations for the CALONLY and INIT periods and the reference source SSCs can now be used to produce the final source photometry. We note that the error distribution of the individual transits of a given source is, in general, heteroscedastic since the observations are taken under a variety of different instrumental configurations. We therefore generate the source photomety as the weighted mean of the individual calibrated observations using the inverse variance as the weight (see Sect. 6 in Carrasco et al. 2016, for more details). The source photometry is produced by applying the SS and LS calibrations to all individual transits of a source followed by the computation of the weighted mean from all calibrated transits. For the G band we included only the AF CCDs since the SM is always observed with the Gate12 configuration. This means that saturation and low photon counts will always be a problem at the bright and faint end, respectively. Moreover, the SM observations are obtained in 2D windows with a sampling such that the effective pixels are twice the size of a standard AF CCD.

When validating the source photometry published in *Gaia* DR1, we discovered that in some cases, it was highly affected by outliers. In particular, because the cross-match and spurious detection black-listing process was still sub-optimal, transits from different sources could occasionally be assigned by the cross-match to the same source. If the magnitude difference between the sources was significant, the epochs from the fainter sources would bias the weighted mean towards the faint end since their associated weights would be much lower than for the brighter epochs. Occasional poor IPD results or the use of poor photometric calibrations could also lead to similar results. For *Gaia* DR2, the robustness of the source photometry determination is improved by introducing a rejection process based on median statistics. We first determine the median and median absolute deviation (MAD) of all valid calibrated observations (in a given band) and then reject all observations that are more than 5σ from the median (the standard deviation was obtained as $\sigma = 1.4826$ MAD). An observation is considered valid if it has been both SS and LS calibrated and if the calibrated flux is higher than 1 $e^- s^{-1}$ ($G \approx 26$); this is a very generous lower limit for a physically meaningful flux.

In order to calibrate a transit of a given source, it is necessary to have the reference SSCs for the source and the reference integrated BP and RP fluxes, which are required in the normalisation of the SSC fluxes (see Appendix A). When deriving the link-calibrated source photometry for the bootstrapping of the photometric system initialisation loop, the time-link calibration could only be applied to sources within the colour range used by the model ($G_{BP} - G_{RP} = [0.0, 4.0]$). All epochs of bluer and redder sources could therefore not be calibrated in the standard procedure and were excluded from the calibration process altogether.

In *Gaia* DR2, we used three different approaches to generate the source photometry, which depends on the availability of colour information for the source. We call the three procedures and the corresponding samples of sources gold, silver, and bronze.

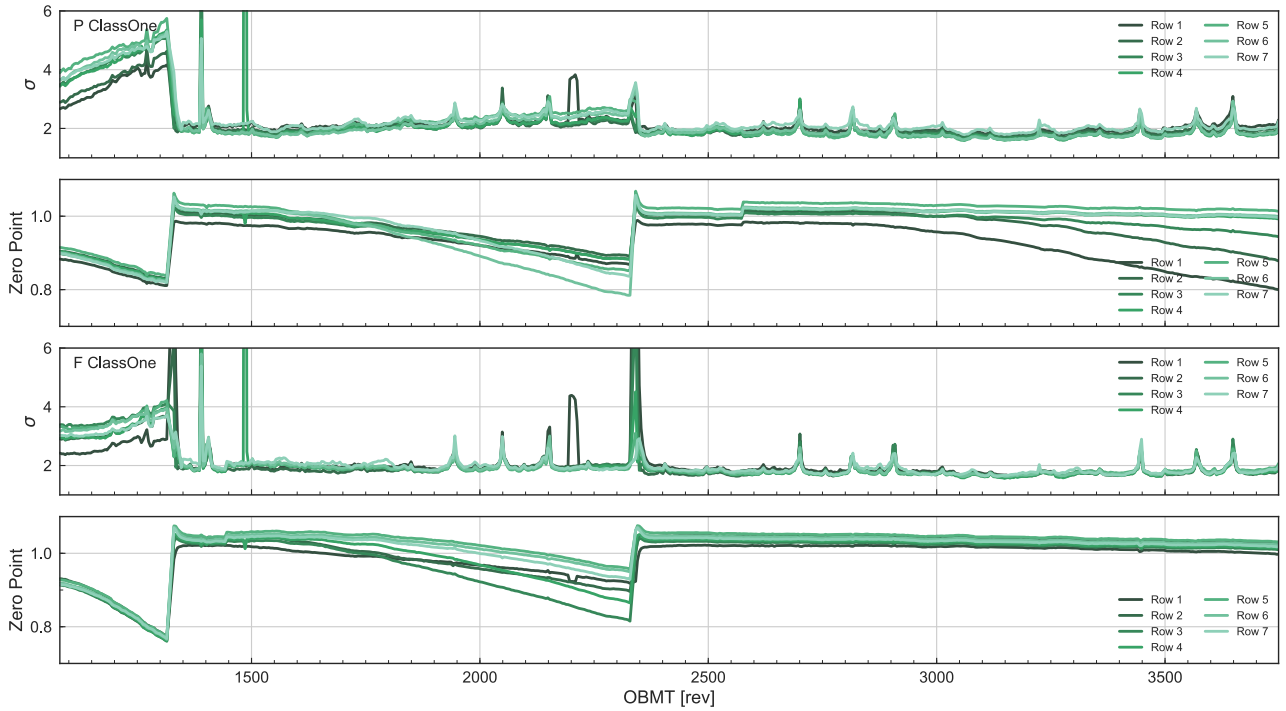

Fig. 10. Time evolution of the LS calibration standard deviation (sigma) and zero point for the preceding FoV (*panels 1 and 2*, respectively) and following FoV (*panels 3 and 4*, respectively) for each CCD row. All solutions only consider the ungated AF1 observations acquired with window class 1 (corresponding approximately to $13 < G < 16$). The two decontaminations are clearly visible in both FoVs as a major discontinuity in the calibration zero point. The first refocus can be seen as a very small step in the zero point for the following FoV, whereas the second refocus is more visible as a slightly larger step in the zero point for the preceding FoV. Additional features visible in both the standard deviation and zero point are discussed in the text.

5.6.1. Gold sources

We define as gold any source for which the photometry was produced by the full calibration process described in Sect. 5.5. PhotPipe produced a total of 1 527 436 167 gold sources. The actual number in the *Gaia* DR2 archive will probably be lower because various data-quality filters are applied during the catalogue preparation (see Arenou et al. 2018, for more detail). Sources were excluded from the gold sample mainly by the colour selection introduced by the time-link calibration. However a number of sources that were originally within the time-link calibration colour range dropped out of the gold sample during the iterative calibration process described in Sect. 5.5. These dropouts are only a small fraction of the initial sample, and the main cause is probably related to a small number of BP/RP transits that fail to be calibrated at some stage during the iterations (see also Appendix A).

5.6.2. Silver sources

To recover sources that were excluded from the gold sample, we implemented an iterative calibration process that uses the SS and LS calibration produced from the CALONLY and INIT datasets to update the mean source photometry starting from the uncalibrated mean source photometry. The effect of the iterations is to produce progressively better source photometry by making use of improved mean source colour information (SSCs). In *Gaia* DR2, we define as silver any source that went through this iterative calibration process: PhotPipe produced a total of 144 944 018 silver sources. The actual number in the *Gaia* DR2 archive will probably be lower because various data-quality filters are applied during the catalogue preparation (see Arenou et al. 2018, for more detail). Sources with incomplete reference source colour information (see Appendix A) could not be calibrated using this iterative process and therefore are not part of the silver sample. As we noted for the gold sample, a small fraction of sources that were originally part of the silver sample (i.e. at the first iteration) dropped out of the sample during the iterative calibration process: the same conclusions as drawn for the gold sources apply.

5.6.3. Bronze sources

Transits for the remaining set of sources in principle are not calibratable since they miss the colour information required to apply the LS calibrations. As a compromise between quality of the photometry and completeness of the *Gaia* DR2 catalogue, we calibrated the remaining sample of sources using a set of default SSC colours. These default colours were obtained from a subset of sources by converting the individual source SSC fluxes into colours, then taking the median value of each SSC colour, and finally renormalising these median SSCs to ensure that their sum is equal to one (see Appendix A). PhotPipe produced a total of 901 338 610 bronze sources, of which 861 630 440 have available G–band photometry, 194 652 181 have integrated BP photometry, and 226 114 046 have integrated RP photometry. The actual number in the *Gaia* DR2 archive will probably be lower because various data-quality filters are applied during the catalogue preparation (see Arenou et al. 2018, for more detail). For all bronze sources available in the *Gaia* DR2 archive, only the G–band photometry is published. We note that several of the bronze sources are likely to have extreme colours (which would explain why so many are missing either BP or RP). Since ≈44%

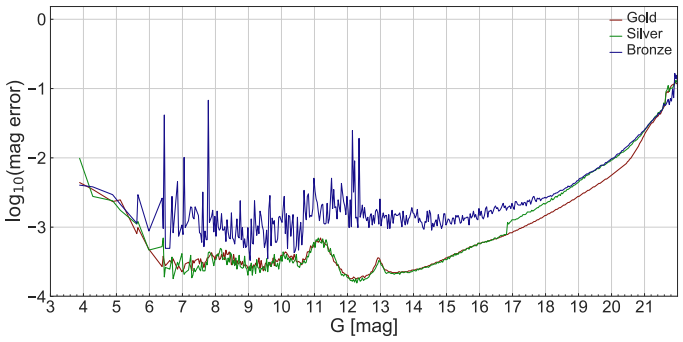

Fig. 11. Mode of the error distribution on the weighted mean G–band magnitude as a function of G magnitude for the gold (red), silver (green), and bronze sources (blue).

of the bronze sources have $G > 21$, it is likely that a significant fraction of these sources are not real and are caused instead by spurious detections.

Figure 11 shows a comparison of the mode of the magnitude error distribution versus magnitude for the G band for gold, silver, and bronze sources. There is very good agreement between the gold and silver source photometry errors: the two samples are indistinguishable up to $G \approx 16.8$, where the error distribution mode is clearly discontinuous because towards the faint end, the error distribution of the silver sources appears to be bimodal. The reasons for this are not yet clear. It is possible that the bimodality is caused by selection effects in the population of sources that end up in the silver calibration mode combined (or not) with processing problems (e.g. background underestimation or crowding effects in BP/RP). These effects are not visible in the bronze sample because the BP/RP information is not used at all. The bronze photometry has considerably larger errors and scatter, as shown by the noise on the mode line. We refer to Evans et al. (2018) for a more detailed discussions of the scientific quality of the source photometry.

6. PhotPipe implementation details

The data-processing platform adopted for PhotPipe is the open-source Hadoop distributed processing system. Hadoop is a mature system with wide adoption in industry for a variety of data processing executed on large datasets. Hadoop has been designed to operate well with commodity hardware and is composed of a distributed file system (HDFS) that provides good resilience against hardware and network failure and against data loss by means of data replication. The other core component is an application resource management layer that allows the scheduling of distributed applications running on the cluster.

The version of the PhotPipe processing system used for *Gaia* DR2 is entirely based on the Map/Reduce programming model (Dean & Ghemawat 2008). The Map/Reduce paradigm is a very simple parallelisation model that involves a data transformation stage (Map), a sorting and grouping by some user-defined key, and a final transformation of the values associated to a given key (Reduce). The Hadoop implementation distributes the processing tasks optimally by scheduling them on the node that holds a local copy of the data. This approach provides both horizontally scalable I/O and processing capacity. In Sect. 6.1 we briefly recall the key concepts of the distributed Map/Reduce framework (see Dean & Ghemawat 2008, for more details), and then we describe in the following section the implementation of the iterative initialisation of the photometric system described in Sect. 5.1.

6.1. Distributed Map/Reduce overview

Given an input stream of key/value pairs of type $\{a, x\}$, the Map/Reduce model involves applying a Map function transforming the ith input key/value pair into a sequence of n key/value pairs of $\{b, y\}$

$$\text{map} : \{a_i, x_i\} \rightarrow (\{b, y\}_n).$$

Let the output key b have K distinct values, the output of the Map function is grouped into K sets composed of the kth b key value and the sequence of values associated to that key. The Reduce function is then applied to each of these sets transforming the input sequence (y_j) keyed by b_k into a sequence of p key/value pairs of type $\{c, z\}$,

$$\text{reduce} : \{b_k, (y_j)\} \rightarrow (\{c, z\}_p).$$

This simple model is implemented by Hadoop in a distributed fashion. The input dataset is stored in the distributed file system, the data are segmented into blocks of equal size, each block is stored on a node of the cluster, and the system ensures that there are always R copies of any given data block stored on R different nodes (where R is configurable per file, based on requirements of performance and robustness against node failure). Assuming that the input data are composed of B blocks, Hadoop will schedule B parallel Map tasks, each one applying the user-defined map function to all key/value pairs in the assigned block. Hadoop attempts to schedule each of the tasks on a node that holds a copy of the block to maximise I/O performance. If a map task fails (e.g. because of hardware/network glitches or outages), Hadoop will automatically re-schedule the task on a different node; if all nodes holding a copy of the input block are busy, Hadoop will schedule the task to another node and the input data will be transferred over the network. The parallelisation of the map stage is determined by the number of blocks in the input data and the the cluster size (i.e. how many nodes and how many tasks per node can be executed).

The size of the input dataset for the Reduce stage is unknown at scheduling time, so that the parallelisation is defined by specifying the number P of partitions in which the dataset set should be subdivided. Each map output key/value pair is assigned by Hadoop to one of the P partitions using a partitioning function: each partition is assigned to a single reduce task. The next stage is called shuffle and involves collecting all records belonging to a given partition on the node that has been assigned the task of processing that partition. This stage involves fetching the data over the network from multiple nodes. Each reducer process then merge-sorts the input data, groups them by key, and applies the user-defined reduce function to each set $\{b_k, (y)_j\}$. The $\{c, z\}_p$ outputs of each partition are then written back to the distributed filesystem. The merge-sort process is very efficient since the outputs of each individual Map task are also merge-sorted and therefore each individual Reduce node need only do one final merge-sort of the partial Map outputs. The default behaviour is for Hadoop to use hash-based partitioning and lexicographic byte order for the sorting and grouping, but each one of these phases can be customised by supplying a user-defined function: this feature is heavily used in the implementation of the photometric calibration workflow. In the following section we present an example of how this simple model involving the definition of a map function, a reduce function, and, optionally, a sorting

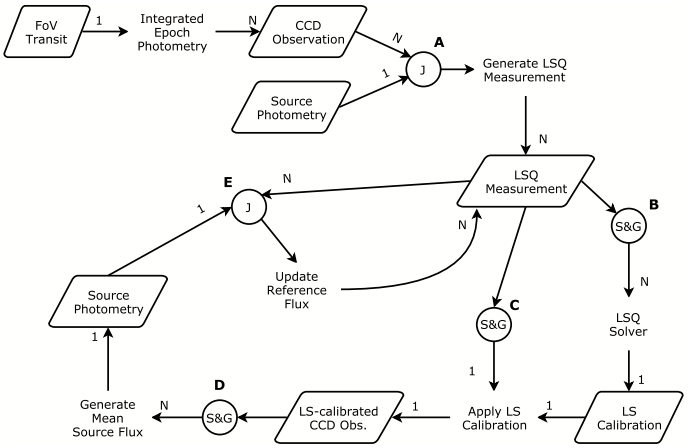

Fig. 12. Map/Reduce workflow for the initialisation of the photometric system (LS iterations, see Sect. 5.1). The workflow is composed of five Map/Reduce jobs labelled **A** to **E**. The red circles represent the global distributed join (**J**) or sorting and grouping by key (S&G) operations. The green boxes represent processes: the input(s) are represented by the incoming arrow, and the output by the outgoing arrow. The data cardinality is shown on the labels as 1 when a single record is handled at a time, or N when multiple records are handled simultaneously. See the text for further information on the distributed workflow.

function and a grouping function can be used to implement the calibration workflow described in Sect. 5 in PhotPipe.

6.2. Distributed LS initialisation

The Map/Reduce implementation of the LS calibration iteration loop (see Fig. 7 and Sect. 5.1) is shown in Fig. 12. The workflow is composed by five jobs: the one-off bootstrap job **A,** and the four jobs that comprise an LS iteration, **B** to **E**.

The first stage of the LS iterations is executed only once and involves generating the individual CCD observations, including integrated BP/RP and SSCs, and attaching to each one the appropriate reference flux information (from the source). This bootstrap job **A** consumes two input streams: (1) the uncalibrated FoV transits that are converted into the integrated epoch photometry (composed of the IPD G–band fluxes, integrated BP/RP fluxes, and SSC) and separated into the individual CCD components keyed by the source identifier; and (2) the reference source photometry records, which are simply read and output keyed by source identifier. At the reduce stage, all records associated with a given source identifier are collected and processed by a single call of the reduce function, which will attach the appropriate reference flux and SSC information to each indivudal CCD transit. We call the output type an LSQ measurement since it represents an individual contribution to one of the LSQ problems producing a given LS calibration solution.

The first job, **B**, of the calibration loop produces the set of LS calibration solutions, one solution is produced per calibration unit and per time-range. This can be easily implemented by assigning each input LSQ measurement to the corresponding time range and calibration unit: a single reduce call will then receive all the LSQ measurements that contribute to a single LS calibration. In this approach, however, the order in which the measurements are processed at the reduce stage is not deterministic since it will depend on the order of completion of the various map tasks. Round-off errors in the LSQ solution could then produce slightly different LS calibrations. In order to make the LS solution deterministic (i.e. fully reproducible), we adopt a compound key containing the time range, CU, and transit identifier. We then use a custom sorting function that orders the LSQ measurements by calibration unit (using the default lexicographic order), and for a given CU, by increasing transit identifier. We then provide a custom grouping function that will perform the grouping based only on the CU and time-range ignoring the transit identifier.

The next stage in the calibration loop involves calibrating the CCD observations, using the set of LS calibrations produced by job **B**, to then generate a new version of the mean source photometry. In principle, we could implement this as a single Map/Reduce job since the LS calibration application process is performed on individual LSQ measurements (and hence could be taken care of in the Map stage), and the source photometry requires all calibrated observations for a given source that can be grouped by outputting the calibrated CCD observations keyed by source identifier at the Map stage. The overall performance (total execution time) of a Map/Reduce job depends heavily on its concurrency: everything else being equal, the more Map (and later Reduce) tasks that can be executed simultaneously on the cluster, the faster the job will complete. The maximum number of concurrent tasks that can run on a single cluster node is limited by the amount of memory available; to maximise concurrency, it is therefore important to minimise the memory footprint of the individual tasks. In our specific case, the input LSQ measurements to be calibrated are not time ordered (because job **A** involves a join by source identifier and hence the output of a given reduce task will be ordered by lexicographic byte of the source identifier hash value) this means that a given Map task would need to keep in memory the entire set of LS calibrations (amounting to several gigabytes). This process can be made more memory efficient by ordering the input LSQ measurements in time: since a Map task only processes a subset of records, it would thus be necessary to keep in memory only a small subset of the LS calibration (several megabytes). For this reason, we perform the generation of the new source photometry in two Map/Reduce jobs. Job **C** reads the LSQ measurements and outputs them keyed by transit identifier: the LSQ measurements reach the reducer sorted by time (since the 42 most significant bits of the transit identifier represent the acquisition time of the AF1 CCD of that transit). At the Reduce stage, we only need to keep a limited number of calibrations in memory since the input data are time ordered. The reducer outputs calibrated CCD observations. Since the source photometry is composed of several passbands (G, $G_{\rm BP}$, $G_{\rm RP}$, and the eight BP/RP SSCs) we define a key containing the source identifier, the transit identifier, and the CCD/SSC information. Job **D** then reads these calibrated CCD observations and outputs them keyed by the compound key defined above. The job uses a custom sorting function that orders the CCD observations for a given source by increasing transit identifier (i.e. increasing time) and increasing CCD/SSC. The job also uses a custom grouping function that will only consider the source identifier component, thus collecting all CCD observations for each source in the order specified above. In this way, the reducer can efficiently compute the source photometry for each band by simply accumulating (see Appendix B) the input calibrated fluxes until a change in CCD/SSC is detected: at that point, the mean flux for this band is finalised and the computation for the following band started. Finally, job **E** closes the calibration loop by updating the original LSQ measurements with the new reference source fluxes produced by job **D**. This is a simple join that supplies each call of the reduce function with all LSQ measurements and the mean photometry (in all bands) for a given source.

6.3. Performance

The processing required for the generation of the *Gaia* DR2 calibrated photometry involved a variety of Map/Reduce jobs with different properties: some were I/O-bound, some were CPU-bound, some involved memory-intensive operations, and others had to perform demanding join operations on hundreds of billions of records. A detailed analysis of the performance properties considerering all these factors is clearly beyond the scope of this paper. However, since this paper presents the first use of the Hadoop and Map/Reduce algorithms in a large-scale astrophysical survey, we believe it is still appropriate to provide some general performance figures that highlight how successful the choice of this processing platform has been for the *Gaia* photometric data.

Since we have covered the Map/Reduce implementation of the LS iteration loop in Sect. 6.2 in more detail, we consider the overall performance of this processing sequence in the context of the initialisation of the photometric system (see Sect. 5.1). The performance metrics for the five LS iterations are listed in Table 3. Each iteration is composed of jobs from **B** to **E**, as described in the previous section, and the jobs were run on the Cambridge Hadoop cluster (see Appendix C for more information). Overall, the 20 jobs completed in 2.8 days, corresponding to nearly 21.5 CPU years. The reduce stage dominates the run time, while the map stage only accounts for ≈38%. This is expected since for all these jobs, the map stage is essentially just reading data from HDFS and applying trivial transformations (e.g. key generation). On the other hand, the reduce stage is responsible both for the LSQ solutions (with iterative rejection) producing the LS calibration and for the computation of the mean source photometry. The job duration is not equal to the sum of the Map and Reduce stage durations since the reducer tasks are normally started before the completion of the map stage. This allows starting the transfer of the outputs of the completed Map tasks to the nodes that will reduce the corresponding partition, hence reducing the delay in starting the actual Reduce stage when the Map stage is completed.

Table 3 shows that the vast majority of the Map tasks have been reading from a local disk: this is indeed crucial in allowing the distributed PhotPipe processing to scale with the input data volume. The read throughput is higher than expected, averaging ≈3 GB s^{-1}, because the operating system can use some of the memory for caching to further optimise the I/O. Finally, we note that the Map stage experiences a very low number, ≈0.04%, of task failures caused by hardware and network glitches, which is normal when processing hundreds of terabytes of data on a system of this scale (see Appendix C).

Table 3. Cumulative performance metrics for the 20 Map/Reduce jobs required to run five LS iterations for the initialisation of the photometric system.

Metric	Map	Reduce	Total
Wall clock time	26.19 h	61.22 h	68.21 h
CPU time	3194.28 h	4650.61 h	7844.89 h
Number of tasks	296,047	46,005	342,052
Data local tasks[†]	96.2%	–	–
Failed tasks	141	1	142
Records I/O[*]	7.74×10^{12}	3.78×10^{12}	–
HDFS I/O[*]	274.77 TB	132.42 TB	–

Notes. A single iteration is composed of job **B** to **E,** as described in Sect. 6.1. See the text for the discussion. ([†]) The concept of a data-local task is only meaningful for the Map stage. ([*]) When reporting I/O figures, the Map stage reports the input figure (since the data are read off the distributed file system), and the Reduce stage reports the output figure (since the job results are written to the distributed file system).

7. Concluding remarks and future developments

Producing science-grade data products from the *Gaia* raw data poses several challenges that are due to the intrinsic complexity of the payload and acquisition system, the huge data volume and granularity, and the necessity of a self-calibration approach. There are simply no full-sky surveys with the same spatial resolution, high accuracy, and precision that *Gaia* could use for the purpose of photometric calibration. In this paper we presented how these challenges were successfully overcome to design and implement a distributed photometric processing system, PhotPipe, which was used to produce the *Gaia* DR2 source photometry in *G* band and BP/RP.

The software architecture, design, and implementation have proved to be very stable during the entire processing phase. Hadoop has proven to be an excellent choice for the core processing architecture with zero downtime due to hardware/system problems.

A significant portion of the overall processing time was dedicated to the validation of the photometric calibration process. These validation tasks have also been implemented as map/reduce jobs: with nearly two billion sources (and two orders of magnitude more epochs), visual inspection is clearly not an option. We therefore took the approach of generating the distributions of various key metrics to be able to quickly assess the quality and progress of the processing. Several examples have been shown in Evans et al. (2017) for *Gaia* DR1 and more are available in Evans et al. (2018) for *Gaia* DR2.

Although the pre-processing stages that were run for *Gaia* DR2 are reliable and able to mitigate the instrumental effects they deal with, the background mitigation in BP/RP requires further improvement to better handle the cases in which the astrophysical background dominates the straylight contribution. This mostly affects the faint sources (e.g. $G > 18$) where the local background becomes a significant fraction of the overall flux. Another improvement that will be introduced in future data releases for BP/RP is related to the handling of crowding effects, which are not limited to the faint end, but affect the full magnitude range and can mimic variability because epochs are acquired with different scanning directions and different overlapping fields of view, as dictated by the satellite scanning law.

Overall, the iterative initialisation of the photometric system performed very well and produced a noticeably better system than in DR1 (see Evans et al. 2018), not only thanks to the improvements in the IPD and *G*–band pre-processing, but also because of the possibility of using only mission data with an overall lower and more stable level of contamination. Although we expect the approach described in this paper to lead to an even better source photometry when the PSF/LSF models used in the IPD process will include time and colour dependencies, some improvements are planned in the calibration process itself. In particular, the gate and window-class link calibration (see Sect. 5.4) does not yet fully remove the discontinuities between the different instrumental configurations (see e.g. Fig. 9 and also Evans et al. 2018). At the bright end ($G < 13$), saturation and flux-loss effects become important. In principle, both effects are

handled by the IPD process, but only for the full 2D windows (i.e. $G < 11$). For the 1D windows, the IPD will handle saturation effects, but not flux loss (due to the lack of AC resolution). Although flux loss does not appear explicitly in the calibration models used for *Gaia* DR2, we should note that the current model provides already a calibration for the average flux-loss experienced by a source. This is equivalent to the case where the source is perfectly centred within the window and there is no AC motion of the source along the CCD transit. The centring and AC motion are different for each epoch of a given source and essentially depend on the scanning law and the random errors in the VPU detection process.

An analysis of the residuals of the epoch photometry error (defined as the difference between the predicted AC position of the source on the CCD, derived from the source astrometry and satellite reconstructed attitude, and the window centre) shows that the centring error distribution of the epochs acquired with window class 1 has a standard deviation of 0.44 pixels and is centred on zero. The residual distribution of the epochs with a centring error in the 1σ range has a standard deviation of 0.005 mag. When considering this second-order effect on the source photometry, the size of this systematic will be much smaller (depending on the number of epochs). Although this effect could in principle be included by the LS calibration model, we note that at the faint end, the effect will be harder to measure reliably because other systematic effects become more important (e.g. background problems and crowding effects due to unresolved sources). In *Gaia* DR2, we decided not to include terms providing the second-order correction of the flux-loss (due to centring error and AC motion) because we expected the effect to be smaller than the overall improvements introduced by the better IPD and calibration strategy. The choice of whether to include these terms will have to be re-evaluted for the next data release: since the IPD process will be different (including colour and time dependency in the PSF/LSF models), it is hard to establish using the current data whether the centring error and AC motion effects will be the same as is seen in the *Gaia* DR2 data.

Although we are very pleased with the overall performance achieved by PhotPipe on the Cambridge cluster (see Appendix C) for this data release, more work is required to ensure that PhotPipe is able to perform equally well in future data releases. One major challenge is posed by the fact that the Map/Reduce allows only one global distributed operation (i.e. sorting+grouping/joining): this means that when the algorithms involve more than one distributed operation, the implementation requires the chaining of several Map/Reduce jobs, therefore generating a large amount of intermediate data. Writing the intermediate data and then reading it back (in the next Map/Reduce job) will progressively slow down the processing as the data volume increases. An obvious way to keep the system scaling is to avoid the persistence of these intermediate data products as much as possible. This can be achieved by rephrasing a given processing flow in terms of a directed acyclic graph (DAG) instead of a linear concatenation of Map/Reduce stages by using, for example, Apache Spark[2]. Although the approach is different, it can be easily implemented without requiring a complete rewrite of the existing software because Spark adopts a functional model for the definition of the dataflow DAG. This allows one to "re-wire" the existing modules defining the various Map and Reduce stages using the Spark API. We have performed extensive testing of this approach that has confirmed the benefits in terms of performance.

Acknowledgements. This work presents results from the European Space Agency (ESA) space mission *Gaia*. *Gaia* data are being processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the *Gaia* MultiLateral Agreement (MLA). The *Gaia* mission website is `https://www.cosmos.esa.int/gaia`. The *Gaia* Archive website is `http://gea.esac.esa.int/archive/`. This work has been supported by the United Kingdom Rutherford Appleton Laboratory, the United Kingdom Science and Technology Facilities Council (STFC) through grant ST/L006553/1, and the United Kingdom Space Agency (UKSA) through grant ST/N000641/1. This work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and ESP2014-55996-C2-1-R (MINECO/FEDER, UE) and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). We also thank the Agenzia Spaziale Italiana (ASI) through grants I/037/08/0, I/058/10/0, 2014-025-R.0, and 2014-025-R.1.2015 to INAF, and the Italian Istituto Nazionale di Astrofisica (INAF). We thank the referee, Mike Bessell, for suggestions that helped improve this paper.

References

Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17 (*Gaia* 2 SI)
Bird, R. 2010, Pearls of Functional Algorithm Design, 1st edn. (New York: Cambridge University Press)
Boyadjian, J. 2008, EADS Astrium/ESA Gaia Instruments Optical Performances Delivery Explanatory Note, GAIA.ASF.TCN.PLM.00108
Carrasco, J. M., Evans, D. W., Montegriffo, P., et al. 2016, A&A, 595, A7
Castañeda, J., Clotet, M., González-Vidal, J. J., et al. 2018, A&A, submitted (*Gaia* 2 SI)
Dean, J., & Ghemawat, S. 2008, Commun. ACM, 51, 107
de Bruijne, J. H. J., Allen, M., Azaz, S., et al. 2015, A&A, 576, A74
Evans, D. W., Riello, M., De Angeli, F., et al. 2017, A&A, 600, A51
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4 (*Gaia* 2 SI)
Fabricius, C., Bastian, U., Portell, J., et al. 2016, A&A, 595, A3
Gaia Collaboration (Brown, A. G. A., et al.) 2016a, A&A, 595, A2
Gaia Collaboration (Prusti, T., et al.) 2016b, A&A, 595, A1
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Hambly, N. C., Cropper, M., Boudreault, S., et al. 2018, A&A, 616, A15 (*Gaia* 2 SI)
Lindegren, L., Lammers, U., Hobbs, D., et al. 2012, A&A, 538, A78
Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2 (*Gaia* 2 SI)
Pancino, E., Altavilla, G., Marinoni, S., et al. 2012, MNRAS, 426, 1767
van Leeuwen, F. 1997, Space Sci. Rev., 81, 201
White, T. 2012, Hadoop: The Definitive Guide, ed. O'Reilly Media, 688

[2] `https://spark.apache.org/`

Appendix A: Spectral shape coefficients

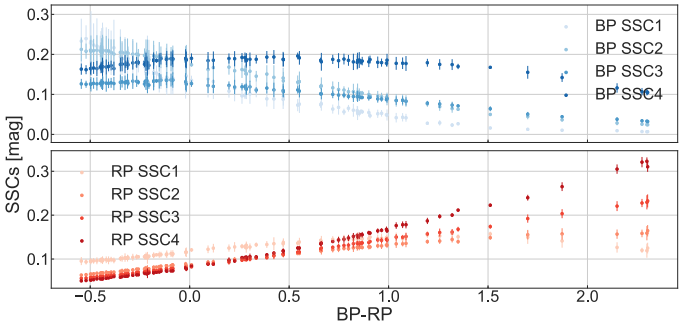

Fig. A.1. Distribution of the 8 SSCs derived from the BP (*top*) and RP (*bottom*) spectra for the set of sources used in the external calibration process.

The SSCs and their use in the photometric calibrations to provide colour information are described in Sects. 4 and 5 of Carrasco et al. (2016). In this section we briefly recall the key concepts to clarify the discussion of the source photometry grade (gold, silver, bronze) in Sect. 5.6.

For both BP and RP, we defined four rectangular bands and produced the integrated flux in each band for each transit: the results of this synthetic photometry are four BP SSCs and four RP SSCs per transit. All eight SSCs are independently calibrated in the same fashion as the G band and integrated BP/RP. The calibrated epoch SSCs for each source are then used to compute the weighted-average source SSCs fluxes. These calibrated source SSC fluxes are then used in the LS calibration model to provide colour information. The four BP SSCs are normalised so that their sum is equal to one, and the same is done for RP. These colour SSCs are then used in the BP and RP LS calibration models. For the calibration of the G band instead both the BP and RP SSCs are used. In this case we apply an additional normalisation to the colour SSCs such that their total sum is equal to one and the ratio of the sum of the BP and RP colour SSCs is equal to the ratio of the integrated BP and RP fluxes. Figure A.1 shows the distribution of the source SSCs for the SPSS used in the external calibration. In this case the same normalisation as was used for the G–band calibration was applied.

In Sect. 4.3 we described that the level of uncertainty in the geometric calibration does not significantly affect the computation of the source SSCs. To confirm this, we have compared the set of source SSCs shown in Fig. A.1 with an alternative set of SSCs computed by applying a geometric calibration with an additional random noise of the same level as the scatter of the geometric calibration solution in a period with no discontinuities. The difference between the two sets of source SSCs is shown in Fig. A.2, where the range covered by the plots is equivalent to a few milli-magnitudes. There is no evidence for systematic differences with colour, and the overall scatter is well within the uncertainties of the SSCs themselves.

One important requirement of using source mean SSCs in the calibration model is that for any given source

– all four BP SSC average source fluxes must be available in order to apply the LS calibration solution to an epoch BP flux to produce the internally calibrated flux (and analogously for RP),
– all eight SSC average source fluxes and the integrated BP/RP average source fluxes must be available in order to apply the LS calibration solution to a G–band epoch flux to produce the internally calibrated flux.

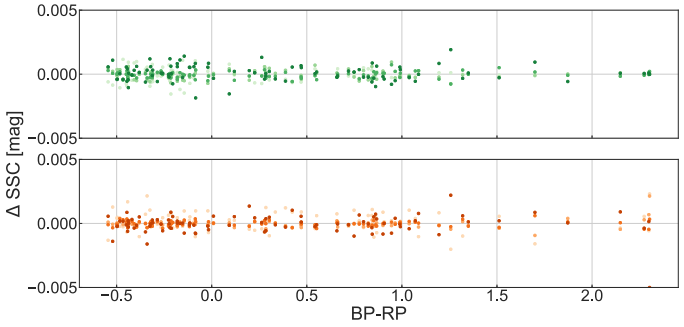

Fig. A.2. Comparison between two sets of SSCs: the first is computed using the final geometric calibrations generated for *Gaia* DR2, and the second is computed adding random noise of the same level as the scatter in the geometric calibration. The *top panel* shows the result for the BP SSCs, and the *bottom panel* shows the comparison for the RP SSCs.

These requirements can become problematic, especially at the faint end and for sources with more extreme colours because the synthetic photometry of the epoch spectra might fail to produce a valid flux for one or more of the SSC bands. If this happens systematically for all transits of a given source, then it will not be possible to calibrate these transits since the source colour information (as represented by the eight SSC fluxes) might be incomplete or missing altogether.

Appendix B: Weighted-mean source photometry by accumulation

For efficiency, PhotPipe implements the computation of the weighted-mean flux in a given band for a given source as a left-fold operation (e.g. Bird 2010) on the sequence of calibrated observations. This is implemented by adding the contribution of an individual calibrated observation to three accumulators that can then be used to generate the weighted-mean flux and error, χ^2, variance and scatter measures, and an estimate of the additional scatter caused by variability (see Sect. 6 in Carrasco et al. 2016 and Eq. (87) of van Leeuwen 1997). For a given passband, the fold operation is based on three accumulators and the total number of contributing observations, N:

$$A_1 = \sum w_i \tag{B.1}$$

$$A_2 = \sum f_i w_i \tag{B.2}$$

$$A_3 = \sum f_i^2 w_i, \tag{B.3}$$

where f_i represents the flux of the ith observation, and $w_i = 1/\sigma_i^2$ is the associated weight defined as the inverse variance.

Appendix C: Hadoop cluster

The Hadoop cluster used for the *Gaia* DR2 photometric data processing is hosted by the High Performance Computing Service of the Univeristy of Cambridge, UK. The cluster is composed of 218 identical nodes that serve both as storage nodes (i.e. contributing to the Hadoop distributed file system) and as compute nodes (i.e. to run the distributed PhotPipe processing jobs). Each node features dual 12 core Intel E5-2650v4 2.2 GHz processors, 256 GB RAM, a 64GB system SSD, and 6 2TB 7.2k SAS hard drives. Overall, the cluster provides a raw capacity of 2.32 PB of storage, 54.5 TB RAM, and 5232 physical cores