Gaia Early Data Release 3: the celestial reference frame (Gaia-CRF3)

Brown, A.G.A.; Castro Ginard, A.; Kostrzewa, Z.P. et al.

A&A 667, A148 (2022)
https://doi.org/10.1051/0004-6361/202243483

Astronomy & Astrophysics

Gaia Early Data Release 3

Special issue

Gaia Early Data Release 3

The celestial reference frame (*Gaia*-CRF3)

Gaia Collaboration: S. A. Klioner[1], L. Lindegren[2], F. Mignard[3], J. Hernández[4], M. Ramos-Lerate[5], U. Bastian[6], M. Biermann[6], A. Bombrun[7], A. de Torres[7], E. Gerlach[1], R. Geyer[1], T. Hilger[1], D. Hobbs[2], U. L. Lammers[4], P. J. McMillan[2], H. Steidelmüller[1], D. Teyssier[5], C. M. Raiteri[8], S. Bartolomé[9], M. Bernet[9], J. Castañeda[10], M. Clotet[9], M. Davidson[11], C. Fabricius[9], N. Garralda Torres[12], J. J. González-Vidal[9], J. Portell[9], N. Rowell[11], F. Torra[10], J. Torra†[9], A. G. A. Brown[13], A. Vallenari[14], T. Prusti[15], J. H. J. de Bruijne[15], F. Arenou[16], C. Babusiaux[17,16], O. L. Creevey[3], C. Ducourant[18], D. W. Evans[19], L. Eyer[20], R. Guerra[4], A. Hutton[21], C. Jordi[9], X. Luri[9], C. Panem[22], D. Pourbaix†[23,24], S. Randich[25], P. Sartoretti[16], C. Soubiran[18], P. Tanga[3], N. A. Walton[19], C. A. L. Bailer-Jones[26], R. Drimmel[8], F. Jansen[27], D. Katz[16], M. G. Lattanzi[8,28], F. van Leeuwen[19], J. Bakker[4], C. Cacciari[29], F. De Angeli[19], M. Fouesneau[26], Y. Frémat[30], L. Galluccio[3], A. Guerrier[22], U. Heiter[31], E. Masana[9], R. Messineo[32], N. Mowlavi[20], C. Nicolas[22], K. Nienartowicz[33,34], F. Pailler[22], P. Panuzzo[16], F. Riclet[22], W. Roux[22], G. M. Seabroke[35], R. Sordo[14], F. Thévenin[3], G. Gracia-Abril[36,6], M. Altmann[6,37], R. Andrae[26], M. Audard[20,34], I. Bellas-Velidis[38], K. Benson[35], J. Berthier[39], R. Blomme[30], P. W. Burgess[19], D. Busonero[8], G. Busso[19], H. Cánovas[5], B. Carry[3], A. Cellino[8], N. Cheek[40], G. Clementini[29], Y. Damerdji[41,42], P. de Teodoro[4], M. Nuñez Campos[21], L. Delchambre[41], A. Dell'Oro[25], P. Esquej[43], J. Fernández-Hernández[12], E. Fraile[43], D. Garabato[44], P. García-Lario[4], E. Gosset[41,24], R. Haigron[16], J.-L. Halbwachs[45], N. C. Hambly[11], D. L. Harrison[19,46], D. Hestroffer[39], S. T. Hodgkin[19], B. Holl[20,34], K. Janßen[47], G. Jevardat de Fombelle[20], S. Jordan[6], A. Krone-Martins[48,49], A. C. Lanzafame[50,51], W. Löffler[6], O. Marchal[45], P. M. Marrese[52,53], A. Moitinho[48], K. Muinonen[54,55], P. Osborne[19], E. Pancino[25,53], T. Pauwels[30], A. Recio-Blanco[3], C. Reylé[56], M. Riello[19], L. Rimoldini[34], T. Roegiers[57], J. Rybizki[26], L. M. Sarro[58], C. Siopis[23], M. Smith[35], A. Sozzetti[8], E. Utrilla[21], M. van Leeuwen[19], U. Abbas[8], P. Ábrahám[59,60], A. Abreu Aramburu[12], C. Aerts[61,62,26], J. J. Aguado[58], M. Ajaj[16], F. Aldea-Montero[4], G. Altavilla[52,53], M. A. Álvarez[44], J. Alves[63], R. I. Anderson[64], E. Anglada Varela[12], T. Antoja[9], D. Baines[5], S. G. Baker[35], L. Balaguer-Núñez[9], E. Balbinot[65], Z. Balog[6,26], C. Barache[37], D. Barbato[20,8], M. Barros[48], M. A. Barstow[66], J.-L. Bassilana[67], N. Bauchet[16], U. Becciani[50], M. Bellazzini[29], A. Berihuete[68], S. Bertone[69,70,8], L. Bianchi[71], A. Binnenfeld[72], S. Blanco-Cuaresma[73], T. Boch[45], D. Bossini[74], S. Bouquillon[37,75], A. Bragaglia[29], L. Bramante[32], E. Breedt[19], A. Bressan[76], N. Brouillet[18], E. Brugaletta[50], B. Bucciarelli[8,28], A. Burlacu[77], A. G. Butkevich[8], R. Buzzi[8], E. Caffau[16], R. Cancelliere[78], T. Cantat-Gaudin[9,26], R. Carballo[79], T. Carlucci[37], M. I. Carnerero[8], J. M. Carrasco[9], L. Casamiquela[18,16], M. Castellani[52], A. Castro-Ginard[13], L. Chaoul[22], P. Charlot[18], L. Chemin[80], V. Chiaramida[32], A. Chiavassa[3], N. Chornay[19], G. Comoretto[5,81], G. Contursi[3], W. J. Cooper[82,8], T. Cornez[67], S. Cowell[19], F. Crifo[16], M. Cropper[35], M. Crosta[8,83], C. Crowley[7], C. Dafonte[44], A. Dapergolas[38], P. David[39], P. de Laverny[3], F. De Luise[84], R. De March[32], J. De Ridder[61], R. de Souza[85], E. F. del Peloso[6], E. del Pozo[21], M. Delbo[3], A. Delgado[43], J.-B. Delisle[20], C. Demouchy[86], T. E. Dharmawardena[26], S. Diakite[87], C. Diener[19], E. Distefano[50], C. Dolding[35], H. Enke[47], C. Fabre[88], M. Fabrizio[52,53], S. Faigler[89], G. Fedorets[54,90], P. Fernique[45,91], A. Fienga[92,39], F. Figueras[9], Y. Fournier[47], C. Fouron[77], F. Fragkoudi[93,94,95], M. Gai[8], A. Garcia-Gutierrez[9], M. Garcia-Reinaldos[4], M. García-Torres[96], A. Garofalo[29], A. Gavel[31], P. Gavras[43], P. Giacobbe[8], G. Gilmore[19], S. Girona[97], G. Giuffrida[52], R. Gomel[89], A. Gomez[44], J. González-Núñez[40,98], I. González-Santamaría[44], M. Granvik[54,99], P. Guillout[45], J. Guiraud[22], R. Gutiérrez-Sánchez[5], L. P. Guy[34,100], D. Hatzidimitriou[101,38], M. Hauser[26,102], M. Haywood[16], A. Helmer[67], A. Helmi[65], M. H. Sarmiento[21], S. L. Hidalgo[103,104], N. Hładczuk[4,105], G. Holland[19], H. E. Huckle[35], K. Jardine[106], G. Jasniewicz[107], A. Jean-Antoine Piccolo[22], Ó. Jiménez-Arranz[9], J. Juaristi Campillo[6], F. Julbe[9], L. Karbevska[34,108], P. Kervella[109], S. Khanna[65,8], G. Kordopatis[3], A. J. Korn[31], Á. Kóspál[59,26,60], Z. Kostrzewa-Rutkowska[13,110], K. Kruszyńska[111], M. Kun[59], P. Laizeau[112], S. Lambert[37], A. F. Lanza[50], Y. Lasne[67], J.-F. Le Campion[18], Y. Lebreton[109,113], T. Lebzelter[63], S. Leccia[114], N. Leclerc[16], I. Lecoeur-Taibi[34], S. Liao[115,8,116], E. L. Licata[8], H. E. P. Lindstrøm[8,117,118], T. A. Lister[119], E. Livanou[101], A. Lobel[30], A. Lorca[21], C. Loup[45], P. Madrero Pardo[9], A. Magdaleno Romeo[77], S. Managau[67], R. G. Mann[11],

† Deceased.

M. Manteiga[120], J. M. Marchant[121], M. Marconi[114], J. Marcos[5], M. M. S. Marcos Santos[40], D. Marín Pina[9], S. Marinoni[52,53], F. Marocco[122], D. J. Marshall[123], L. Martin Polo[40], J. M. Martín-Fleitas[21], G. Marton[59], N. Mary[67], A. Masip[9], D. Massari[29], A. Mastrobuono-Battisti[16], T. Mazeh[89], S. Messina[50], D. Michalik[15], N. R. Millar[19], A. Mints[47], D. Molina[9], R. Molinaro[114], L. Molnár[59,124,60], G. Monari[45], M. Monguió[9], P. Montegriffo[29], A. Montero[21], R. Mor[9], A. Mora[21], R. Morbidelli[8], T. Morel[41], D. Morris[11], T. Muraveva[29], C. P. Murphy[4], I. Musella[114], Z. Nagy[59], L. Noval[67], F. Ocaña[5,125], A. Ogden[19], C. Ordenovic[3], J. O. Osinde[43], C. Pagani[66], I. Pagano[50], L. Palaversa[126,19], P. A. Palicio[3], L. Pallas-Quintela[44], A. Panahi[89], S. Payne-Wardenaar[6], X. Peñalosa Esteller[9], A. Penttilä[54], B. Pichon[3], A. M. Piersimoni[84], F.-X. Pineau[45], E. Plachy[59,124,60], G. Plum[16], E. Poggio[3,8], A. Prša[127], L. Pulone[52], E. Racero[40,125], S. Ragaini[29], M. Rainer[25,128], N. Rambaux[39], P. Ramos[9,45], P. Re Fiorentin[8], S. Regibo[61], P. J. Richards[129], C. Rios Diaz[43], V. Ripepi[114], A. Riva[8], H.-W. Rix[26], G. Rixon[19], N. Robichon[16], A. C. Robin[56], C. Robin[67], M. Roelens[20], H. R. O. Rogues[86], L. Rohrbasser[34], M. Romero-Gómez[9], F. Royer[16], D. Ruz Mieres[19], K. A. Rybicki[111], G. Sadowski[23], A. Sáez Núñez[9], A. Sagristà Sellés[6], J. Sahlmann[43], E. Salguero[12], N. Samaras[30,130], V. Sanchez Gimenez[9], N. Sanna[25], R. Santoveña[44], M. Sarasso[8], M. Schultheis[3], E. Sciacca[50], M. Segol[86], J. C. Segovia[40], D. Ségransan[20], D. Semeux[88], S. Shahaf[131], H. I. Siddiqui[132], A. Siebert[45,91], L. Siltala[54], A. Silvelo[44], E. Slezak[3], I. Slezak[3], R. L. Smart[8], O. N. Snaith[16], E. Solano[133], F. Solitro[32], D. Souami[109,134], J. Souchay[37], A. Spagna[8], L. Spina[14], F. Spoto[73], I. A. Steele[121], C. A. Stephenson[5,135], M. Süveges[136], J. Surdej[41,137], L. Szabados[59], E. Szegedi-Elek[59], F. Taris[37], M. B. Taylor[138], R. Teixeira[85], L. Tolomei[32], N. Tonello[97], G. Torralba Elipe[44], M. Trabucchi[139,20], A. T. Tsounis[140], C. Turon[16], A. Ulla[141], N. Unger[20], M. V. Vaillant[67], E. van Dillen[86], W. van Reeven[142], O. Vanel[16], A. Vecchiato[8], Y. Viala[16], D. Vicente[97], S. Voutsinas[11], M. Weiler[9], T. Wevers[19,143], Ł. Wyrzykowski[111], A. Yoldas[19], P. Yvard[86], H. Zhao[3], J. Zorec[144], S. Zucker[72], and T. Zwitter[145]

(Affiliations can be found after the references)

Received 4 March 2022 / Accepted 15 April 2022

ABSTRACT

Context. *Gaia*-CRF3 is the celestial reference frame for positions and proper motions in the third release of data from the *Gaia* mission, *Gaia* DR3 (and for the early third release, *Gaia* EDR3, which contains identical astrometric results). The reference frame is defined by the positions and proper motions at epoch 2016.0 for a specific set of extragalactic sources in the (E)DR3 catalogue.

Aims. We describe the construction of *Gaia*-CRF3 and its properties in terms of the distributions in magnitude, colour, and astrometric quality.

Methods. Compact extragalactic sources in *Gaia* DR3 were identified by positional cross-matching with 17 external catalogues of quasi-stellar objects (QSO) and active galactic nuclei (AGN), followed by astrometric filtering designed to remove stellar contaminants. Selecting a clean sample was favoured over including a higher number of extragalactic sources. For the final sample, the random and systematic errors in the proper motions are analysed, as well as the radio–optical offsets in position for sources in the third realisation of the International Celestial Reference Frame (ICRF3).

Results. *Gaia*-CRF3 comprises about 1.6 million QSO-like sources, of which 1.2 million have five-parameter astrometric solutions in *Gaia* DR3 and 0.4 million have six-parameter solutions. The sources span the magnitude range G = 13–21 with a peak density at 20.6 mag, at which the typical positional uncertainty is about 1 mas. The proper motions show systematic errors on the level of 12 μas yr^{-1} on angular scales greater than 15 deg. For the 3142 optical counterparts of ICRF3 sources in the S/X frequency bands, the median offset from the radio positions is about 0.5 mas, but it exceeds 4 mas in either coordinate for 127 sources. We outline the future of *Gaia*-CRF in the next *Gaia* data releases. Appendices give further details on the external catalogues used, how to extract information about the *Gaia*-CRF3 sources, potential (Galactic) confusion sources, and the estimation of the spin and orientation of an astrometric solution.

Key words. astrometry – reference systems – proper motions – catalogs – quasars: general

1. Introduction

Gaia is the second highly successful astrometric space mission of the European Space Agency (ESA[1]; Gaia Collaboration 2016), vastly extending and improving the achievements of its predecessor HIPPARCOS (ESA 1997; Perryman et al. 1997; van Leeuwen 2007). While the latter delivered astrometry at the level of a milliarcsecond (mas) for nearly 120 000 preselected stars at the mean epoch 1991.25, the latest data release *Gaia* EDR3 contains astrometry for about 1.8 billion sources with accuracies from a few tens of microarcseconds (μas) to ∼1 mas, mainly depending on the brightness (Gaia Collaboration 2021a; Lindegren et al. 2021b).

It is well known that the observational principles of *Gaia* (and HIPPARCOS) do not allow one to fully fix the reference frame, that is the coordinate system of positions and proper motions on the sky. *Gaia* effectively measures angular separations of sources on the sky, and these separations remain

[1] Some further acronyms used in this paper: EDR3 – *Gaia* Early third Data Release; DR3 – *Gaia* third Data Release; QSO – quasi stellar object (here understood as all compact extragalactic sources); AGN – active galactic nuclei; ICRS – International Celestial Reference System; ICRF3 – third realisation of the International Celestial Reference Frame; *Gaia*-CRF*n* – *n*th realisation of the *Gaia* Celestial Reference Frame; VLBI – very long baseline interferometry; S/X, K, X/Ka (after ICRF3) denote the radio frequency bands used in VLBI.

unchanged if one rotates the whole solution by an arbitrary angle at each moment of time. Although the degeneracy with rotation is not mathematically exact, owing to the effects of aberration and gravitational light bending which are coupled to the fixed solar system and *Gaia* ephemerides, those effects are too small to be of any practical help in fixing the reference frame. In the standard astrometric solution, the apparent motion of a source on the sky is modelled by five freely adjusted parameters: two angular components of its position at some reference epoch, the parallax, and two components of the mean proper motion over the observed time interval. This model gives a six-fold near-degeneracy in the astrometric solution: three constant angles of rotation for the positions at the reference epoch (the 'orientation' of the solution), and three components of the constant angular velocity of rotation that affect the proper motions (the 'spin' of the solution). In short, although there are many more observations than unknowns to fit, the resulting system always has a rank deficiency of six. To fix the reference frame of a *Gaia* solution, external information should be used to constrain precisely these six degrees of freedom, and nothing more.

Gaia routinely delivers astrometry for sources as faint as $G = 21$ mag and, unlike HIPPARCOS, it can observe a large number of extragalactic sources. This capability plays an important role for the whole project. By directly linking the astrometric solution to a set of remote extragalactic sources, *Gaia* implements the principles set forth in the International Celestial Reference System (ICRS; Arias et al. 1995), a concept originally devised for the use in radio astrometry observing radio-loud quasars at cosmological distances. In practice, the term 'extragalactic' here refers to objects more distant than about 50 Mpc, that is well beyond the Local Group. The reasoning for choosing remote extragalactic sources is simple: because they are not affected by the complex motions in our Galaxy and have negligible apparent motions on the sky (thanks to their large distances), they provide reliable and stable beacons for astrometry. The suggestion to rely on very distant and nearly motionless sources can be traced back to W. Herschel and P.S. Laplace two centuries ago, well before quasars were spotted (see Sect. 6 of Fricke 1985 for a brief historical account).

The ICRS embraces two ideas of a very different nature. The first one concerns the orientation of a celestial reference frame at a given epoch while the second one deals with the rotational state (spin) of the reference frame. The orientation of a celestial reference frame at a given epoch is a pure convention, with no physical significance to the particular choice of axes. In this respect the choice of origin for the right ascension in ICRS is as arbitrary as, for example, the choice of Greenwich as the Prime Meridian for the geographic longitude in 1884. *Gaia* uses its own astrometric observations of the optical counterparts to sources from the well-established radio astrometric catalogue ICRF3 (Charlot et al. 2020) to fix the orientation of the *Gaia* catalogue. This choice is of a practical nature, namely to ensure consistency between the optical and radio frames and continuity of the orientation used in astronomy over past decades.

The second aspect, much more significant, is the use of remote extragalactic sources to fix the rotational state of the reference frame. Individual sources may show some (even non-linear) positional variations in time, owing to their peculiar motions or intrinsic variations of their structure (e.g. Taris et al. 2018), but it is assumed that objects at cosmological distances, on the average, show 'no rotation'.

In the framework of Newtonian physics, the assumption of no rotation is interpreted as no rotation with respect to Newtonian absolute space. In the modern framework of the General Theory of Relativity, the procedure results in a reference frame in which the Hubble flow has no rotational component. One can argue that the no rotation assumption is related to Mach's principle (e.g. Schiff 1964). However, we stress that *Gaia*'s observations of distant extragalactic objects can neither prove nor disprove that principle. To test Mach's principle is a separate task beyond the scope of the current discussion.

Besides the task of fixing the *Gaia* reference frame, having a set of astrometrically well-behaved extragalactic sources is also important for assessing systematic errors in the *Gaia* astrometric solution. Again, this is possible because we can assume that the true values of the proper motions and parallaxes of these sources are all very close to zero. Such applications of the *Gaia*-CRF3 sample are found, for example, in Lindegren et al. (2021a,b).

Based on the way *Gaia* automatically detects and observes all point-like objects of sufficient brightness (de Bruijne et al. 2015), *Gaia* EDR3 is expected to contain a few million extragalactic objects suitable for defining a reference frame. However, it is not easy to reliably identify those sources in the *Gaia* catalogue. As was the case for *Gaia*-CRF1 (Mignard et al. 2016) and *Gaia*-CRF2 (Gaia Collaboration 2018), the identification of well-behaved extragalactic sources for *Gaia* EDR3 relies on a dedicated cross-matching with external catalogues of quasi-stellar objects (QSOs; also often dubbed 'quasars' irrespective of their radio loudness) and active galactic nuclei (AGNs). For the purpose of this work the actual nature of the compact extragalactic sources plays no role, and we refer to them as 'QSO-like' sources, whether they are genuine QSOs of various kinds, galaxies with astrometrically compact nuclei, or other types of extragalactic objects. The resulting set of QSO-like objects selected on this basis constitutes *Gaia*-CRF3 – the third version of the *Gaia*-CRF. This paper describes the procedure used to select the sources of *Gaia*-CRF3 and presents its most important properties and its relation to the radio ICRF.

According to the principles of the ICRS, *Gaia*-CRF3 is entirely defined by the set of ∼1.6 million QSO-like sources described in this paper. It is implicitly assumed that the positions and proper motions of all other (mostly stellar) sources in *Gaia* DR3 are expressed in the same reference frame, and thus constitute a secondary, much denser reference frame that extends to the brightest magnitudes. For stars of similar magnitude and colour as the QSO-like sources, the consistency of the reference frames should be guaranteed by the methods of observation and data analysis in *Gaia*, which are exactly the same for both kinds of objects. For much brighter sources, and perhaps also for objects of extreme colours, this may not be the case owing to various instrumental effects and deficiencies in their modelling. Indeed, Cantat-Gaudin & Brandt (2021) have shown that the proper motions of bright ($G \lesssim 13$ mag) sources in *Gaia* EDR3 have a residual spin with respect to the fainter stars by up to 80 μas yr^{-1}, and provide a recipe for their correction. Similarly, the alignment errors for the bright sources in *Gaia* EDR3 were investigated by Lunz et al. (2021), who found differences of about 0.5 mas. The properties of the secondary (stellar) reference frame in *Gaia* EDR3 are not further discussed in this paper.

The structure of the paper is as follows. Section 2 describes the multi-step procedure used to select the *Gaia*-CRF3 sources. The properties of those sources are discussed in Sect. 3. The *Gaia*-CRF3 sources matched to the radio sources of ICRF3 give an important link to radio astrometry and are discussed in Sect. 4. Section 5 reviews the role of the *Gaia*-CRF3 sources in the 'frame rotator' of the primary astrometric solution for *Gaia* EDR3, which is where the reference frame was actually

Table 1. Statistics of the cross-matches of *Gaia* EDR3 to external QSO catalogues used in the construction of *Gaia*-CRF3.

Catalogue	Sources	Unique matches	Filtered sources	Retained sources	Code	Reference
ICRF3 S/X	4536	3477	3142	3142	icrf3sx	Charlot et al. (2020)
ICRF3 K	824	715	660	660	icrf3k	Charlot et al. (2020)
ICRF3 X/Ka	678	611	576	576	icrf3xka	Charlot et al. (2020)
OCARS	7607	5337	4723	4723	ocars	Malkin (2018)
AllWISE	1 354 775	733 462	580 403	580 403	aw15	Secrest et al. (2015)
Milliquas v6.5	1 980 903	1 347 414	1 065 936*	1 039 610	m65	Flesch (2019)
R90	4 543 530	1 331 547	1 022 081	1 022 081	r90	Assef et al. (2018)
C75	20 907 127	2 068 813	1 265 419*	1 169 431	c75	Assef et al. (2018)
SDSS DR14Q	526 356	368 013	308 608	308 608	dr14q	Pâris et al. (2018)
LQAC-5	592 809	421 289	348 085	348 085	lqac5	Souchay et al. (2019)
LAMOST phase 1, DR1-5	42 578	42 255	39 886	39 886	lamost5	Yao et al. (2019)
LQRF	100 165	98 902	94 839	94 839	lqrf	Andrei et al. (2009)
2QZ	28 495	24 192	21 569	21 569	cat2qz	Croom et al. (2004)
Roma-BZCAT, release 5	3561	3343	2986	2986	bzcat5	Massaro et al. (2015)
2WHSP	1691	982	413	413	cat2whsp	Chang et al. (2017)
ALMA calibrators	3361	2746	2209	2209	alma19	Bonato et al. (2019)
Gaia-unWISE C75	2 734 464	2 726 201	2 061 253*	1 569 359	guw	Shu et al. (2019)
Total retained matches				1 614 173		

Notes. The external QSO catalogues used to construct *Gaia*-CRF3 are identified in Cols. 1 and 7. Column 2 is the total number of sources in the catalogue. Column 3 gives the number of unique matches in *Gaia* EDR3 to the catalogue, using a matching radius of $\Delta_{\max} = 100$ mas for the four VLBI-based catalogues (ICRF3 and OCARS) and $\Delta_{\max} = 2$ arcsec for the other catalogues. Column 4 contains the number of matched sources satisfying Eqs. (1)–(5) (where the last two criteria were not used for the VLBI-based catalogues). An asterisk means that the filtered sources show evidence of stellar contamination. Column 5 gives the number of sources retained in *Gaia*-CRF3 after the second stage of the filtering described in Sect. 2.3. Column 6 gives the code for the catalogue in the cross-match table `gaiadr3.gaia_crf3_xm` (see Appendix C).

fixed. Some conclusions and prospects for future versions of the *Gaia*-CRF are given in Sect. 6.

The appendices provide more detailed information on a few more technical issues. Appendix B provides some additional details of the external catalogues used for *Gaia*-CRF3. Appendix C demonstrates how to select the *Gaia*-CRF3 sources in the *Gaia* Archive and trace which external catalogues have matches of those sources. In Appendix D we discuss the 'confusion sources' in *Gaia* EDR3, that is sources, mainly in the Galaxy or Local Group, whose astrometric parameters are roughly consistent with them being at cosmological distances, and that could therefore potentially be confused with QSO-like sources. Finally, Appendix E describes the algorithm used by the frame rotator to estimate the spin and orientation corrections for astrometric solutions.

2. Selection of *Gaia* sources

The goal of the source selection process for the *Gaia*-CRF is to find a set of QSO-like sources that combines three conflicting requirements: it should be as large as possible, as clean as possible (that is, free of stellar contaminants), and homogeneously distributed on the sky. We expect that this selection can eventually be made exclusively from *Gaia*'s own observations, using the astrophysical characterisation of sources based on a combination of astrometric and photometric information (e.g. Bailer-Jones et al. 2013, 2019; Delchambre 2018; Heintz et al. 2018). This was not possible for *Gaia* EDR3, and similarly to what was done for the first two *Gaia* data releases, *Gaia* EDR3 uses external catalogues of QSOs and AGNs to identify QSO-like sources among the nearly two billion *Gaia* sources. As detailed below, the *Gaia*-CRF3 content is the result of cross-matches with 17 external catalogues, followed by a two-stage astrometric filtering. These external catalogues were all publicly available at the time of the final construction of *Gaia*-CRF3 in mid-June 2020.

2.1. Cross-matching

For the construction of *Gaia*-CRF3 we cross-matched the *Gaia* EDR3 astrometric catalogue with the external catalogues listed in Table 1 (see also Appendix B). Only astrometric information was used for the cross-matches. The first four entries in the table – three ICRF3 catalogues and OCARS – are based on VLBI astrometry in the radio domain and have an astrometric quality comparable to that of *Gaia* EDR3. For those catalogues the maximal cross-match radius was set to $\Delta_{\max} = 100$ mas. The other catalogues have various levels of astrometric quality and for them the cross-match radius was set to $\Delta_{\max} = 2$ arcsec. For virtually all catalogues the cross-matching results in a (relatively small) number of ambiguous matches, where more than one *Gaia* source was matched to a source in the external catalogue, or more than one source in the external catalogue was matched to the same *Gaia* source. In all such cases only the closest positional match was retained. For each external catalogue this gives a list of 'unique' matches in the form of pairs '*Gaia* source identifier' – 'external source identifier', in which all identifiers appear only once. Table 1 gives the total number of sources in the catalogue and, in column 3, the number of unique matches in *Gaia* EDR3, to which the two-stage filtering described below was applied.

This cross-matching procedure adopted for *Gaia*-CRF3 is not perfect and could be refined in future releases. For example, the maximum cross-matching radius could be better tuned to the precisions of the individual external catalogues and the expected densities of stars and quasars in the different situations. A particular case concerns multiple lensed quasar images, where

ideally all the images should be retained instead of just the closest match, as is presently done. In some ambiguous cases the cross-matching could benefit from using also the photometric information available in the external catalogues. This is however far from straightforward owing to the expected variability of the QSOs, differences in wavelength bands and resolution between the instruments, etc., which may even require a manual check of individual cases. For the current cross-matching we have chosen not to use such information.

2.2. Astrometric filtering: Individual sources

The first stage of the astrometric filtering ensures that the retained *Gaia* sources have five- and six-parameter astrometric solutions in *Gaia* EDR3 compatible with the assumed extragalactic nature of the source. For all the catalogues in Table 1 we applied the three criteria in Eqs. (1)–(3) below. This was found to be sufficient for the VLBI catalogues (the first four entries in the table), but for the remaining catalogues the two criteria in Eqs. (4) and (5) were also applied. Some of the criteria are alternatively formulated as parts of a query in the *Gaia* Archive[2] using the Astronomical Data Query Language (ADQL).

The first criterion is that the source has a valid parallax and proper motion. This is ensured by selecting sources with five- or six-parameter astrometric solutions:

```
astrometric_params_solved = 31 OR
astrometric_params_solved = 95.
```
(1)

We note that `astrometric_params_solved` is a binary flag indicating which astrometric parameters are provided for the source, that is 11_2, 11111_2, and 1011111_2 for two-, five-, and six-parameter solutions, with the seventh bit representing pseudo-colour.

Secondly, the measured parallax ϖ of the source, corrected for the median parallax bias, should be zero within five times the formal uncertainty σ_ϖ:

$$\left|\frac{\varpi + 0.017\,\mathrm{mas}}{\sigma_\varpi}\right| < 5\,, \tag{2}$$

which translates to the ADQL query

```
abs((parallax + 0.017)/parallax_error) < 5.
```

This takes into account the median parallax of −0.017 mas obtained from a provisional selection of QSO-like sources in *Gaia* EDR3, but ignores the complicated dependence of the parallax bias on the magnitude and other parameters discussed by Lindegren et al. (2021a). As shown by their Fig. 5, the variation of the parallax bias for the QSO-like sources is typically less than ±0.010 mas. Because this is many times smaller than the individual uncertainties (for example, 99% of the *Gaia*-CRF3 sources have $\sigma_\varpi > 0.1$ mas), these variations were deemed uncritical for the astrometric filtering. Assuming Gaussian errors, $(\varpi + 0.017\,\mathrm{mas})/\sigma_\varpi$ is then expected to follow the standard normal distribution, and the probability that Eq. (2) is violated by random errors is, nominally, $\mathrm{erfc}\left(5/\sqrt{2}\right) \simeq 5.7\times10^{-7}$.

The third criterion is that the proper motion components $\mu_{\alpha*}$ and μ_δ should be small compared to their uncertainties $\sigma_{\mu\alpha*}$ and $\sigma_{\mu\delta}$. Here we take into account the correlation $\rho(\mu_{\alpha*}, \mu_\delta)$ between the two parameters. We require

$$X^2_\mu \equiv \begin{bmatrix}\mu_{\alpha*} & \mu_\delta\end{bmatrix} \mathrm{Cov}(\boldsymbol{\mu})^{-1} \begin{bmatrix}\mu_{\alpha*}\\ \mu_\delta\end{bmatrix} < 25\,, \tag{3}$$

[2] https://gea.esac.esa.int/archive/

where $\mathrm{Cov}(\boldsymbol{\mu})$ is the covariance matrix given by Eq. (E.4). This corresponds to the query

```
( power(pmra/pmra_error,2)
  +power(pmdec/pmdec_error,2)
  -2*pmra_pmdec_corr
  *pmra/pmra_error
  *pmdec/pmdec_error )
/(1-power(pmra_pmdec_corr,2)) < 25.
```

For Gaussian errors of the proper motion components, X^2_μ is expected to follow the chi-squared distribution with two degrees of freedom (see Appendix B of Mignard et al. 2016). In this case the nominal probability that a source violates Eq. (3) is $\exp(-25/2) \simeq 3.7\times10^{-6}$. We neglect here the systematic proper motions of extragalactic objects caused by the acceleration of the solar system barycentre (e.g. Gaia Collaboration 2021b). The amplitude of this effect, about 0.005 mas yr^{-1}, is much smaller than the typical proper-motion uncertainty of QSO-like sources in *Gaia* EDR3 (see below) and is therefore ignored in the filtering. We note that X_μ is independent of the reference frame, so that, in particular, it plays no role if the equatorial or, say, galactic components of the proper motions are used in Eq. (3).

For the VLBI catalogues based on high-accuracy radio astrometry (the first four entries in Table 1), the application of Eqs. (1)–(3) results in filtered samples that show no traces of stellar contamination. This is to be expected thanks to the high positional accuracy of the VLBI data and the small cross-match radius used (100 mas). Consequently, no additional filtering was considered for these catalogues.

The situation is very different for the remaining (non-VLBI) catalogues in Table 1. An analysis of their matches, filtered by Eqs. (1)–(3), revealed that a significant number of chance matches to (stellar) sources in *Gaia* EDR3 happen through a combination of two factors: (i) the high density of potentially matching stellar sources in certain areas of the sky; and (ii) the lower positional accuracy of these catalogues which motivated the use of a much larger cross-matching radius (2 arcsec) than for the VLBI catalogues.

To investigate the likely effects of chance matches we make use of the 'confusion sources' discussed in Appendix D. These are all the sources in *Gaia* EDR3 that satisfy all three conditions in Eqs. (1)–(3), irrespective of their true nature. The total number of confusion sources in *Gaia* EDR3 is approximately 213.5 million, of which 30.7 million have five-parameter solutions and 182.8 million have six-parameter solutions.

The galactic plane is especially problematic for the cross-matching: 23% of the confusion sources with five-parameter astrometric solutions and 50% of the confusion sources with six-parameter solutions are located in the 10% of the celestial sphere with the galactic latitude b such that $|\sin b| < 0.1$. The analysis of the matches from non-VLBI catalogues showed that the resulting samples of QSO candidates in that area are too much contaminated by stars and cannot reasonably be further cleaned by means of astrometric criteria alone. We therefore decided to completely avoid this area of the sky for *Gaia*-CRF3 for the matches to the non-VLBI catalogues.

Considering the distribution of the *Gaia*-CRF3 sources and the confusion sources in galactic latitude shown on Figs. 5 and D.2, respectively, our results for *Gaia* EDR3 generally agree with the discussion of the stellar contamination at various galactic latitudes provided by Heintz et al. (2020; see their Fig. 9).

The second factor important for the chance matches – the lower quality of the astrometric data in the non-VLBI catalogues – is also more problematic in the areas of higher density of

confusion sources. Figure D.2 shows that the density of confusion sources is a strong function of the Galactic latitude b. To avoid too many chance matches at low $|b|$ we require a smaller matching distance Δ towards the Galactic plane. For the matches to the non-VLBI catalogues we therefore add the two criteria

$$|\sin b| > 0.1 \tag{4}$$

and

$$\Delta < (2\ \mathrm{arcsec}) \times |\sin b|\,. \tag{5}$$

We note that, after the initial cross-matching, this last criterion is the only one that uses astrometric data (positions) from the external catalogue. The number of unique matches that survived the filtering by Eqs. (1)–(5), or Eqs. (1)–(3) for the VLBI catalogues, is given in the fourth column of Table 1.

2.3. Astrometric filtering: Sample statistics

The second stage of the filtering is based on an assessment of the level of stellar contamination among the filtered unique matches in the different external catalogues. To this end we analysed histograms of the normalised parallaxes ϖ/σ_ϖ and normalised proper motion components $\mu_{\alpha*}/\sigma_{\mu_{\alpha*}}$ and $\mu_\delta/\sigma_{\mu_\delta}$. Ignoring the small systematic and random offsets discussed in Sect. 2.2, we expect that, for a clean sample of QSOs, the normalised quantities should have Gaussian distributions with mean values close to zero and standard deviations only slightly greater than one (reflecting the slight underestimation of the formal uncertainties; see Sect. 3). Conversely, if their distributions deviate significantly from the expected Gaussian, it indicates that the sample is contaminated by stellar objects with non-zero parallaxes and/or proper motions.

Although the criteria in Eqs. (1)–(3) are based on the Gaussian hypothesis, they absolutely do not guarantee that the normalised parallaxes and proper motions are Gaussian after the filtering. Figures D.3 and D.4 show examples of the distributions of the normalised parameters for general *Gaia* EDR3 sources satisfying the criteria. While a fraction of the sources in these samples are QSOs (including *Gaia*-CRF3 sources), most of them are stars in our Galaxy or in nearby satellite galaxies such as the Large Magellanic Cloud (LMC) and Small Magellanic Cloud (SMC), and the histograms of the normalised quantities are very far from Gaussian.

A high level of stellar contamination in a sample can be directly seen in the histograms of the normalised parameters as a distortion of the expected Gaussian distribution. Any sub-sample large enough to create representative histograms can be checked separately. This gives a flexible tool for screening samples of QSO candidates against stellar contamination.

For the proper motions the same effect is illustrated in Fig. 1. This shows the relative frequencies of X^2_μ from Eq. (3) for the filtered matches to the catalogues in Table 1. If X^2_μ comes purely from the Gaussian errors of the *Gaia* proper motions, it should follow the χ^2 distribution with two degrees of freedom, with normalised frequency $\exp(-\frac{1}{2}x/s^2)$. Here s is a scaling factor ($\simeq$1) that takes into account that the uncertainties of the astrometric parameters in *Gaia* EDR3 are slightly underestimated. In Fig. 1 this expected frequency is a straight line through unity at $X^2_\mu = 0$.

Applying this tool to the full sets of filtered matches to each external catalogue in Table 1 reveals that only the sources matched to three of the catalogues (C75, Milliquas v6.5, and *Gaia*-unWISE) show clear evidence of stellar contamination. These catalogues are marked with an asterisk in the table. The

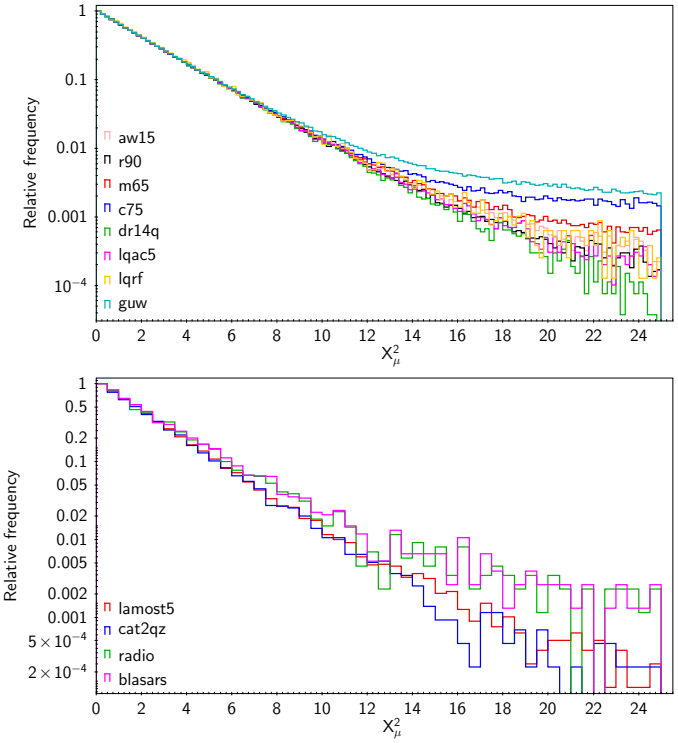

Fig. 1. Relative frequencies of X^2_μ for the filtered matches to the catalogues in Table 1. The catalogues are identified by the codes given in column 6 of the table, with the eight larger catalogues in the *upper panel* and the smaller catalogues in the *lower panel*. The four radio catalogues (ICRF3 S/X, ICRF3 K, ICRF X/Ka, and OCARS) are combined under 'radio', and the three blasar catalogues (Roma-BZCAT, 2WHSP, and ALMA calibrators) are combined under 'blasars'.

filtered matches of the remaining 14 catalogues show no obvious traces of stellar contamination and therefore qualify as being part of *Gaia*-CRF3.

The same *Gaia* source usually appears in the filtered cross-matches to several external catalogues. One reason for this, as discussed in Appendix B, is that many of the catalogues are at least partially based on the same observational material (the IR photometry of WISE, certain data releases of SDSS, etc.), although their criteria for the selection of QSOs may be different. But even when the observational material is different, it is no wonder that different approaches and algorithms result in strongly overlapping samples, namely if they tend to find the same objects. It is therefore reasonable to expect that the circumstance that a *Gaia* source has matches in several external catalogues increases the reliability of its classification as a QSO. If we split the set of *Gaia* sources matching a given external catalogue (after the astrometric filtering) into subsets that are present also in other catalogues and a subset that appears only in the cross-match with this particular catalogue, we can therefore use an analysis of these subsets to improve the purity of the final QSO selection.

Using this idea, we selected three sets of *Gaia* sources that appear only among the filtered matches of C75, Milliquas v6.5, and *Gaia*-unWISE, but do not appear in any of the other catalogues: we call these the c75-only, m65-only, and guw-only sets. All three sets show significant stellar contamination. We found no obvious way to clean the c75-only sources and this whole set of 95 988 sources was therefore dropped from further consideration. The Milliquas v6.5 and *Gaia*-unWISE catalogues include

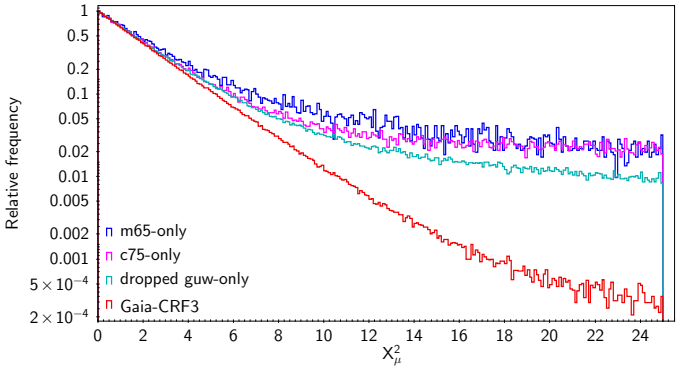

Fig. 2. Relative frequencies of X^2_μ for the final selection of *Gaia*-CRF3 sources, and for the filtered matches to C75, Milliquas v6.5, and *Gaia*-unWISE that were dropped in the final selection.

estimates of the probability that a source is a QSO. Selecting from the m65-only set only the sources with the highest probability of being a QSO did not remove the contamination and that set of 26 326 sources was therefore also dropped. For the guw-only set we found, on the contrary, that the subset of sources with the highest QSO probability (PROB_RF = 1) appears to be free of stellar contaminants. This gave an additional set of 229 914 sources for *Gaia*-CRF3. The other 491 894 guw-only sources were dropped (any relaxation of PROB_RF = 1 immediately gave samples with clear signs of stellar contamination). The number of sources from the cross-matches of each catalogue that passed also this second stage of astrometric filtering is given in the fifth column of Table 1, summing up to a total of 1 614 173 sources. The relative frequencies of X^2_μ in *Gaia*-CRF3 and in the three sets of sources dropped from C75, Milliquas v6.5, and *Gaia*-unWISE are shown in Fig. 2. Deviations from the exponential distribution expected for Gaussian errors are very obvious for the dropped sets; for the final selection the distribution is briefly discussed below (Sect. 2.4).

A separate investigation was done for the QSO candidates in the areas of the LMC and SMC, which have a very high density of confusion sources (Fig. D.1) and therefore a higher risk of chance matches. The statistics of confusion sources in these areas are discussed in Appendix D. The *Gaia*-CRF3 contains 5411 and 2779 sources, respectively, in the LMC and SMC areas (as defined in the appendix). The distributions of the normalised quantities indicate a low level of stellar contamination in these sets.

Further checks of the selection were done by splitting the *Gaia*-CRF3 sources into subsets according to magnitude, colour, and Galactic or ecliptic latitude, and examining the distributions of normalised parallaxes and proper motions. No clear contamination was detected in any of the subsets.

Gaia-CRF3 contains only 195 sources with $|\sin b| < 0.1$. They all come from the VLBI-based catalogues, for which Eqs. (4) and (5) were not applied. We made an attempt to find a clean set of QSO-like sources also in the Galactic plane area, that is without applying the criterion in Eq. (4). Using the cross-matches and the remaining four criteria, we found nearly 100 000 additional QSO candidates with $|\sin b| < 0.1$. However, the distributions of their normalised parameters indicate a high level of stellar contamination. This demonstrates that the cross-matches and/or external astrophysical classifications are unreliable in this crowded part of the sky, and none of these sources were further considered for *Gaia*-CRF3. However, we note that this set should contain a significant percentage of genuine QSOs, part of which may be recovered in future versions of the *Gaia*-CRF (see Sect. 2.5).

2.4. Final selection

The union of all *Gaia* sources that passed both stages of the filtering contains 1 614 173 sources and constitutes the final source list of *Gaia*-CRF3. Of these, 1 215 942 (75%) have five-parameter solutions in *Gaia* EDR3 and 398 231 (25%) have six-parameter solution. Appendix C explains how to find the astrometric parameters of these sources and how to trace the matches of these sources in the external catalogues.

Figure 3 shows the distributions of the normalised parallaxes and proper motion components for both kinds of solutions. The standard deviations of the three normalised quantities are 1.052, 1.055, and 1.063 for the five-parameter solutions and 1.073, 1.099, and 1.116 for the six-parameter solutions. All distributions are very symmetric and close to the Gaussian shape (i.e. parabolic in the logarithmic plots), suggesting a low level of stellar contamination. The small excess of sources with normalised parameters beyond ±4 is not necessarily a sign of contamination, but could be produced by source structure or instrumental effects as well as by statistical fluctuations in the sample.

The number of (non-QSO) chance matches in *Gaia*-CRF3 can be estimated by counting the confusion sources that satisfy Eqs. (4) and (5) around positions that are slightly offset from the *Gaia*-CRF3 positions. Using offsets of ±0.1 deg in $\Delta\alpha\cos\delta$ or $\Delta\delta$ we conclude that *Gaia*-CRF3 is expected to contain up to 420 chance matches, of which 150 have five-parameter solutions and 270 have six-parameter solutions. This corresponds to a probable contamination of 0.012 and 0.068%, respectively, of the five- and six-parameter sources in *Gaia*-CRF3.

The chance matches are obviously not the only cause of stellar contamination in *Gaia*-CRF3. Further contaminants may come from erroneous classification of some sources in the external catalogues, even among the sources that passed the astrometric filtering in Eqs. (1)–(5). We have found no means to quantify this effect, but expect it to be quite small.

A certain hint on the possible (total) level of contamination could be given by the relative frequency of X^2_μ for *Gaia*-CRF3, as shown by the red curve in Fig. 2. This distribution can be rather well described as the mixture of two Gaussian distributions: one with standard deviation of 1.051 containing 98.0% of the sources, the other with standard deviation of 2.038 containing 2.0% of the sources. We stress that this cannot directly be interpreted in terms of the purity of the *Gaia*-CRF3: on one hand, genuine QSOs may show deviations from Gaussianity for other reasons (as mentioned above); on the other hand, some contaminants will have measured proper motion that are perfectly consistent with the extragalactic hypothesis. Nevertheless, we consider it unlikely that the level of contamination in *Gaia*-CRF3 is higher than 2%. Further investigations are needed to show if this claim is justified.

2.5. Discussion of the selection

Assuming that QSO-like sources have zero true values for the parallaxes and proper motions, and that the astrometric errors are Gaussian with uncertainties as given in *Gaia* EDR3, we expect that very few actual QSOs in a sample of a few million candidates will violate Eqs. (2) or (3). The numerical boundaries in Eqs. (2) and (3) were selected empirically, based on the data themselves, and represent a compromise taking into account possible non-Gaussian systematic errors and effects of complex

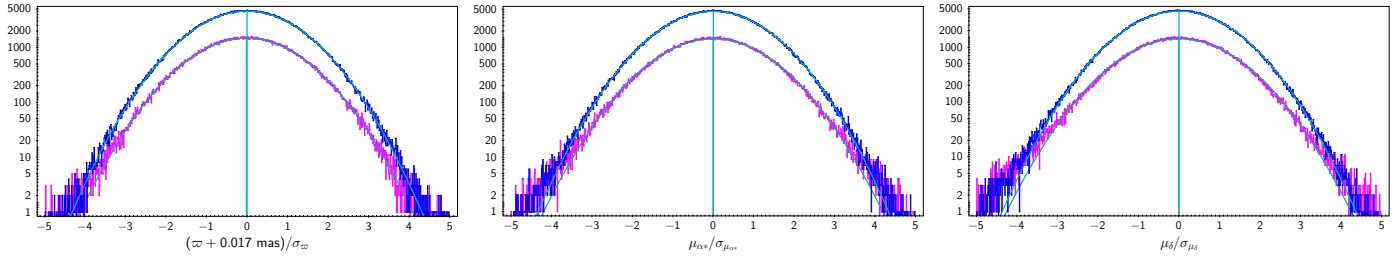

Fig. 3. Distributions of the normalised parallaxes and proper motion components for the *Gaia*-CRF3 sources with five-parameter (blue) and six-parameter (magenta) solutions. The turquoise curves show the corresponding best-fit Gaussian distributions.

and sometimes time-dependent source structure. The latter are to be expected for QSOs and could produce apparent proper motions over several years of observations, while parallaxes are much less affected by such effects. This justifies setting separate criteria for parallaxes and proper motions as described above. In future *Gaia* data releases, when these effects may be better known and statistically quantified, it will be possible to adopt a mathematically more consistent filtering approach based on the 3×3 covariance matrix of the parallax and proper motion components.

As already mentioned, we expect that future versions of *Gaia*-CRF will to a higher degree depend on *Gaia*'s own (astrometric, photometric, and spectroscopic) observations, and that external catalogues of AGNs will consequently become less important for the selection of QSO-like sources. These catalogues will continue to be important for validation purposes, but their uses in the selection process may become more complex if the astrometry in external catalogues is already *Gaia*-based. For example, a criterion such as Eq. (5) may not be meaningful if the position in the external catalogue comes from a previous *Gaia* data release.

The search for the QSOs in the Galactic plane is an important scientific problem that requires further investigation. We note that more advanced methods to combine different datasets in the search for QSOs in the Galactic plane are under development (e.g., Fu et al. 2021). Combined with the expected use of *Gaia*'s own object classification, one can hope that these methods will result in a much improved completeness of the *Gaia*-CRF in the Galactic plane in the future.

The selection of QSO-like objects described here is mainly driven by the reliability of the resulting set of sources. The resulting set of QSO-like sources can be considered 'reliable', but by no means 'complete'. It is clear that the selection algorithm rejects some genuine quasars, for example if they show observable mean proper motions for the time span of the *Gaia* EDR3 data. A prominent example here is 3C273 – the brightest quasar on the sky, which has significant proper motion in declination in *Gaia* EDR3 and $X^2_{\mu} = 29.7$ (Eq. (3)).

For the purposes of aligning the reference frame of *Gaia* EDR3 with the ICRF, and especially for ensuring that it has negligible spin with respect to distant galaxies, one can use a much smaller set of QSO-like sources than the full *Gaia*-CRF3 without significantly increasing the statistical uncertainties of the global system. Indeed, this system was defined in the final stages of the primary astrometric solution, as described in Sect. 5, using only about 0.4 million 'frame rotation' sources. However, the purpose of *Gaia*-CRF3 is not just to fix the system for the *Gaia* astrometric solution, but to provide the community with a direct access to this system with sources as faint as permitted by the sensitivity of *Gaia*, while still having a very low level of stellar contamination.

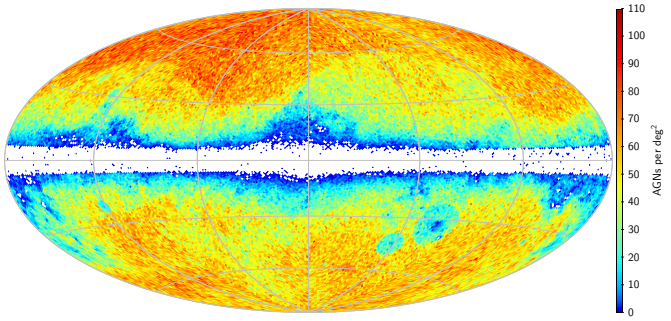

Fig. 4. Sky distribution of the *Gaia*-CRF3 sources. The plot shows the density of sources per square degree computed from the source counts per pixel using HEALPix of level 6 (pixel size $\simeq$0.84 deg^2). The magnitude distribution and further properties of the sources are shown on Fig. 6. This full-sky map uses a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left.

3. Statistical properties of the *Gaia*-CRF3 sources

In this section we present the overall properties of *Gaia*-CRF3 primarily in terms of its distributions in position, magnitude, colour, and astrometric quality.

3.1. Distribution of the sources on the sky

The spatial distribution of the ∼1.6 million *Gaia*-CRF3 sources is shown in Fig. 4. Their distributions in galactic longitude and latitude are shown in Fig. 5. In the avoidance zone $|\sin b| < 0.1$ *Gaia*-CRF3 has only 195 sources matched to the radio catalogues as described in Sect. 2. Outside the avoidance zone the average density is about 42 deg^{-2}, but with significant variations with galactic latitude; the density is typically below average for $|b| \lesssim 30°$ and above at higher latitudes. This comes from the source catalogues used in our compilation combined with the selection function in Eq. (4), designed to avoid stellar contaminants near the Galactic plane. From Fig. B.1 one can see that also most of the external catalogues avoid crowded regions on the sky. We note that the density of *Gaia*-CRF3 reaches about 110 deg^{-2} in some areas on the sky. Given the expected isotropic distribution of QSOs and the effects of Galactic extinction, a complete version of a *Gaia*-CRF could therefore potentially comprise about 4 million sources. This estimate is also suggested by recent studies (e.g. Shu et al. 2019; Bailer-Jones et al. 2019; Fu et al. 2021) of the QSOs as well as by the extragalactic results of *Gaia* DR3.

3.2. Main characteristics of the sources

Table 2 and Fig. 6 show the median values and distributions of several important properties of the *Gaia*-CRF3 sources, split

Table 2. Characteristics of the *Gaia*-CRF3 sources.

Type of solution	Number of sources	G [mag]	$G_{BP} - G_{RP}$ [mag]	ν_{eff} [μm^{-1}]	RUWE	$\sigma_{pos,max}$ [μas]	$\sigma_{\mu\alpha*}$ [$\mu as\,yr^{-1}$]	$\sigma_{\mu\delta}$ [$\mu as\,yr^{-1}$]
Five-parameter	1 215 942	19.92	0.64	1.589	1.013	385	457	423
Six-parameter	398 231	20.46	0.92	1.541	1.044	749	892	832
All	1 614 173	20.06	0.68	1.585	1.019	447	531	493

Notes. Columns 3–9 give median values of the G magnitude, the $G_{BP} - G_{RP}$ colour index, the effective wavenumber ν_{eff} (or the pseudo-colour for six-parameter solutions), the astrometric quality indicator RUWE (see text), the positional uncertainty at epoch J2016.0 (semi-major axis of the error ellipse), and the uncertainties of the proper motion components in α and δ. The last line ('All') is for the whole set of *Gaia*-CRF3 sources.

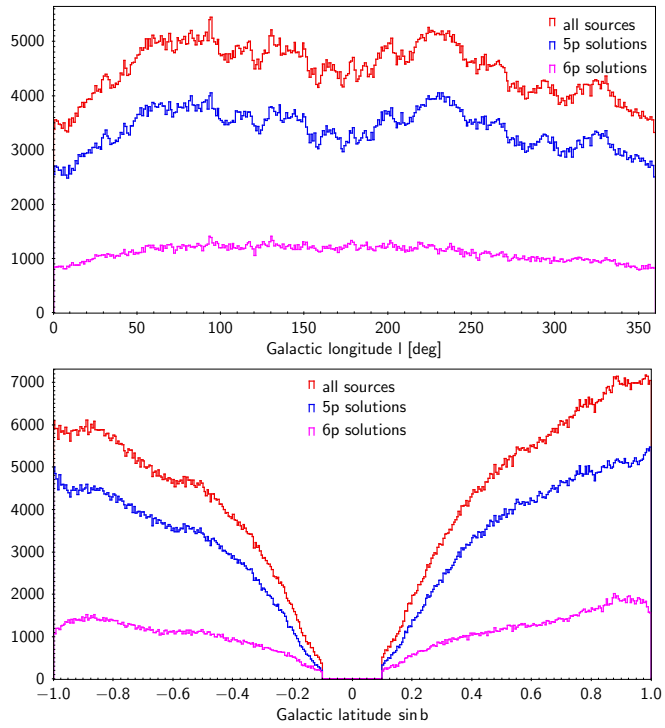

Fig. 5. Distribution of the *Gaia*-CRF3 sources in galactic coordinates. There are only 195 sources in the avoidance zone $|\sin b| < 0.1$. The distribution is shown for the whole *Gaia*-CRF3 and separately for the sources with five- and six-parameter astrometric solutions.

according to the nature of the solution (five- or six-parameter solution): magnitude (G), colour index ($G_{BP} - G_{RP}$), and the renormalised unit weight error (RUWE; Lindegren et al. 2021b). The RUWE is a measure of the goodness-of-fit of the astrometric model to the observations of the source. The expected value for a good fit is 1.0. A higher value could indicate that the source is not point-like at the optical resolution of *Gaia* ($\simeq 0.1''$), or that it has a time-variable structure. As seen in the left panel of Fig. 6, the proportion of six-parameter solutions increases for $G \gtrsim 19$, a feature that also implies that these sources are on average of lower astrometric accuracy, not only because they are fainter but also because the model fit is often worse, as shown in the right panel. The sources with six-parameter solutions are also on average redder (middle panel).

The sky distributions of the median magnitude G and the colour index $G_{BP} - G_{RP}$ (Fig. 7) clearly show the effects of the Galactic extinction and reddening. We also see that the selection favours brighter sources in the crowded areas such as the Galactic plane and the LMC and SMC areas.

3.3. Astrometric parameters of the sources

The left panel of Fig. 8 shows the distribution of the formal positional uncertainty in *Gaia*-CRF3 separately for the five- and six-parameter solution. The positional uncertainty, $\sigma_{pos,max}$, is defined as the semi-major axis of the uncertainty ellipse (see Eq. (1) in Gaia Collaboration 2018) at the reference epoch J2016.0. The middle panel is a scatter plot of $\sigma_{pos,max}$ versus G, with the smoothed median indicated by the dashed curve. In the right panel, the black curve (labeled 'all sources') gives the fraction of the full sky that has a density of *Gaia*-CRF3 sources exceeding the value on the horizontal axis. The other curves give the fractions when only sources with $\sigma_{pos,max}$ below a certain level are counted. In the 10% of the sky nearest to the Galactic plane ($|\sin b| < 0.1$) the source density is negligible (cf. Fig. 4), which explains why none of the curves reaches above 90% sky coverage. To compute the fractions, the celestial sphere was divided into 49 152 pixels of equal area ($\simeq 0.8393$ deg^2, i.e. HEALPix level 6) and the density computed in each pixel. It should be noted that the curves therefore refer to the source densities at a spatial resolution of about 1 deg and would not be the same at a different resolution (pixel size). The plot shows, for example, that for a minimum density of 35 *Gaia*-CRF3 sources per square degree, about 62% of the sky could be covered; this fraction is reduced to 55 or 20% if we require, in addition, that $\sigma_{pos,max}$ is below 1.0 or 0.5 mas. Conversely, at 62% coverage there could be at least 30, 18, and 3 sources per square degree with $\sigma_{pos,max} < 1$, 1.0 or 0.5 mas, if in each case the most populated pixels are chosen. The corresponding magnitude distributions can be gleaned from the scatter plot in the middle panel, where the different levels of positional uncertainty have been marked.

Figure 9 shows the distributions of the median parallax and proper motion components on the sky. Both large-scale systematics and small-scale inhomogeneities are visible in these maps. The large-scale systematics are discussed in Sect. 3.4. Small-scale inhomogeneities, seen as fluctuating median values between neighbouring HEALPix cells in the maps, are mainly caused by the statistical scatter of the astrometric parameters in each cell. They are generally stronger in areas of low source density (Fig. 4) or fainter median magnitude (Fig. 7, top), and also in the ecliptic belt (ecliptic latitude $|\beta| \lesssim 45°$), where the scanning law usually results in less accurate astrometric solutions than at high ecliptic latitudes (see, for example, Sect. 5.4 and Figs. 8 and 9 in Lindegren et al. 2021b). Numerous streaks, edges, and other small-scale features can also be traced to the scanning law. The overall scatter of the median values in the three maps is about 0.11 mas (ϖ), 0.12 mas yr^{-1} ($\mu_{\alpha*}$), and 0.11 mas yr^{-1} (μ_δ), if only the cells with at least five QSOs are counted (covering 85% of the sky). The positional errors are expected to be of a roughly similar size as the parallax scatter. Thus, we conclude that

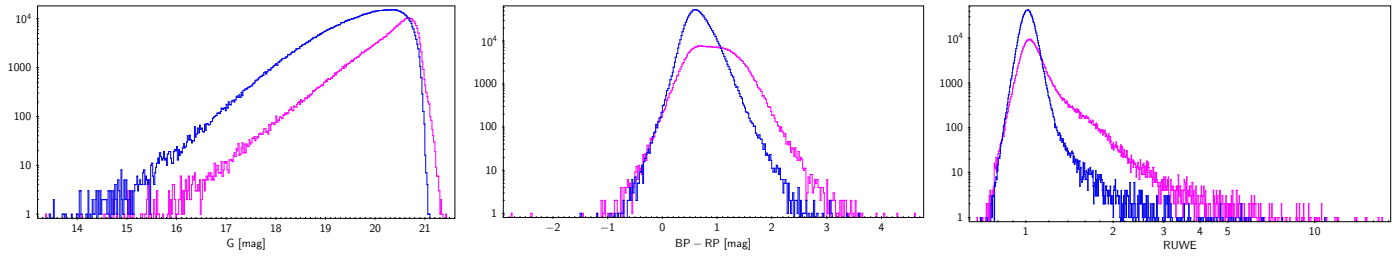

Fig. 6. Histograms of some important characteristics of the *Gaia*-CRF3 sources with five-parameter solutions (blue) and six-parameter solutions (magenta). *Left to right*: G magnitudes, colour index $G_{\rm BP} - G_{\rm RP}$, and the astrometric quality indicator RUWE.

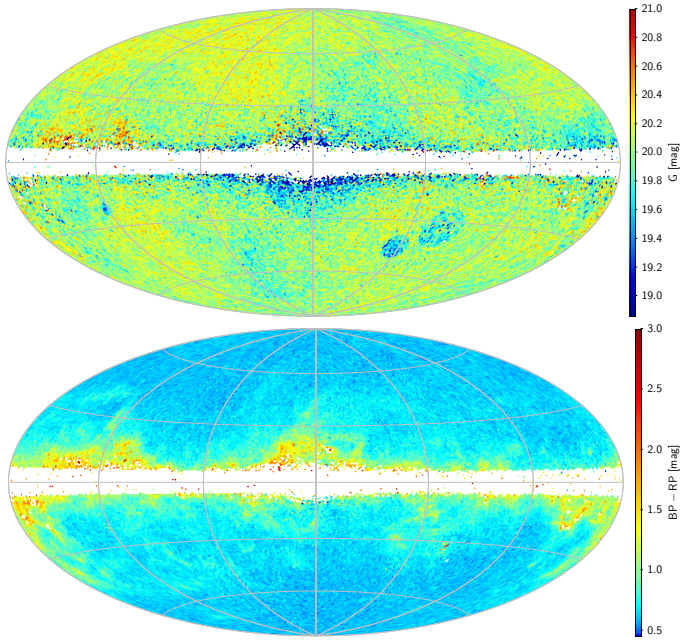

Fig. 7. Median G magnitude (*top*) and median colour index $G_{\rm BP} - G_{\rm RP}$ (*bottom*) of the *Gaia*-CRF3 sources. These maps use a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left.

Gaia-CRF3, on an angular scale of about 1° and over most of the sky, defines a local reference frame accurate to better than 0.2 mas in position and 0.2 mas yr^{-1} in proper motion.

3.4. Systematic errors

Systematic errors in *Gaia*-CRF3 were discussed already in Lindegren et al. (2021b, Sect. 5.6). In particular, the left column of their Fig. 13 contains smoothed maps of the parallaxes and proper motions of the *Gaia*-CRF3 sources that have five-parameter astrometric solutions. These maps were obtained by a special smoothing procedure applied to the parameters of the individual sources. Including also the less numerous and fainter sources with six-parameter solutions does not noticeably change these maps.

Here we present the results of a different way to analyse systematics in the *Gaia*-CRF3 proper motions, namely by fitting an expansion of vector spherical harmonics (VSH; Mignard & Klioner 2012) up to a certain degree ℓ. Although the VSH provide a smooth model of the systematic errors also in the regions where only a few sources are present, such as in the Galactic plane, the higher-degree harmonics ($\ell \gtrsim 15$) are only weakly constrained in these areas, resulting in unreliable estimates of the VSH coefficients also in other parts of the sky because of the global scope of the fitted VSH functions.

Analogous to the discussion of parallax systematics in Sect. 5.7 of Lindegren et al. (2021b), using (scalar) spherical harmonics, the VSH allow us to estimate the angular power spectrum of the systematics in the *Gaia*-CRF3 proper motions. To this end we used the set of 1 215 942 *Gaia*-CRF3 sources with five-parameter solutions. Given the vector field of proper motions of those sources we computed the VSH expansion as given by Eq. (30) of Mignard & Klioner (2012), including all the terms of degree $1 \le \ell \le L$. It was found that the VSH fit is reasonably stable for $L \le 12$. When higher degrees are included, the fits become questionable, mainly because of the inhomogeneous distribution of the sources on the sky (see Sect. 3.1 above and Figs. 3 and 4 of Gaia Collaboration 2021b). All subsequent computations used $L = 12$. The fitted VSH coefficients confirmed the maps of systematic errors in proper motion shown on Fig. 13 of Lindegren et al. (2021b). From the same coefficients we obtained the powers P^t_ℓ, P^s_ℓ of the toroidal and spheroidal signals of degree ℓ, as defined by Eq. (76) of Mignard & Klioner (2012). The total power of degree ℓ is $P_\ell = P^t_\ell + P^s_\ell$. The mean powers of the toroidal and spheroidal harmonics of degree ℓ are then $C^t_\ell = P^t_\ell/(2\ell + 1)$ and $C^s_\ell = P^s_\ell/(2\ell + 1)$, where $2\ell + 1$ is the number of harmonics at the degree ℓ. This is fully analogous to the definitions for the scalar spherical harmonics given by Eq. (31) of Lindegren et al. (2021b).

We note that the VSH components of degree $\ell = 1$ in the proper motions should be analysed separately from other systematic effects. The toroidal harmonic of degree $\ell = 1$ represents a residual spin of the reference frame, and should in principle be negligible thanks to the way the reference frame is constructed (see Sect. 5 below). The spheroidal harmonic of degree $\ell = 1$ represents a 'glide' component of the proper motions that is expected as a consequence of the acceleration of the Solar system in the rest frame of remote sources (Gaia Collaboration 2021b). Although these terms need to be included in the VSH fit, the subsequent discussion of systematics in the *Gaia*-CRF3 proper motions only considers the VSH components of degree $\ell = 2, \ldots, 12$.

Formally, the powers and the mean powers are weighted sums of the squared VSH coefficients. Appendix A of Gaia Collaboration (2021b) demonstrates that such quantities are biased (overestimated) when the VSH coefficient are estimated in a least squares fit. Equation (A.1) of that appendix gives the exact formula to correct the bias. In particular, the corrected (unbiased) values of the mean powers are

$$\hat{C}^t_\ell = C^t_\ell - N^t_\ell\,, \tag{6}$$

$$\hat{C}^s_\ell = C^s_\ell - N^s_\ell\,, \tag{7}$$

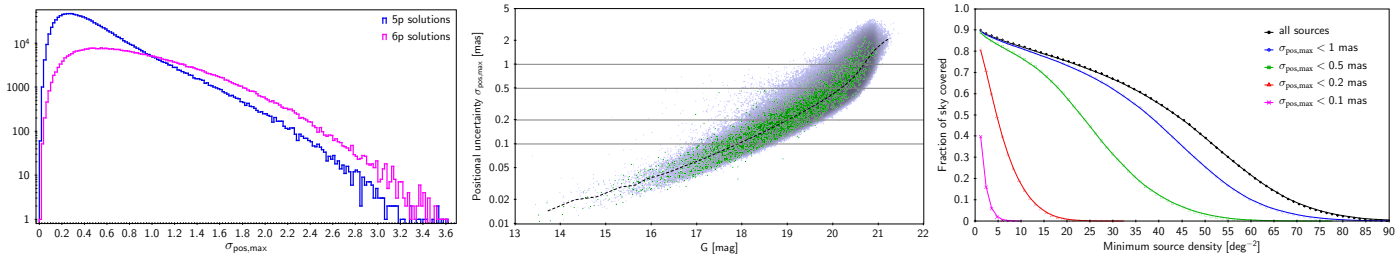

Fig. 8. Positional uncertainty and source density distribution in *Gaia*-CRF3. *Left*: histogram of the positional uncertainty $\sigma_{\rm pos,max}$ for *Gaia*-CRF3 sources with five-parameter solutions (blue) and six-parameter solutions (magenta). *Middle*: scatter plot of $\sigma_{\rm pos,max}$ versus the G magnitude. The dashed curve is the smoothed median uncertainty versus magnitude. The green points are the sources matched to ICRF3 S/X. *Right*: fraction of sky covered by a given minimum density of *Gaia*-CRF3 sources at different levels of positional uncertainty (see text for further explanation).

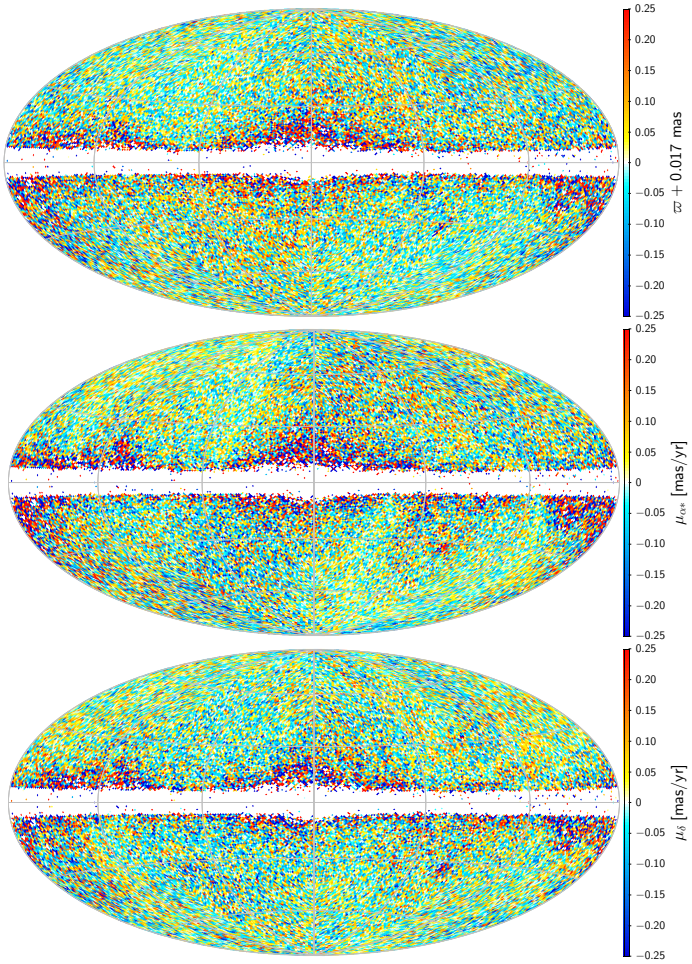

Fig. 9. Sky distributions of the parallax ϖ and the components of proper motion $\mu_{\alpha*}$ and μ_δ. Median values over the pixels of HEALPix level 6 (pixel size $\simeq$0.84 deg^2) are shown. These maps use a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left.

where the noise corrections N^t_ℓ and N^s_ℓ are obtained by the same formulas as C^t_ℓ and C^s_ℓ but using the uncertainties of the VSH coefficients instead of the coefficients themselves. Explicit computations for the present data show that $N^t_\ell \simeq N^s_\ell \simeq 3.44 - 0.07\,\ell - 0.01\,\ell^2$ μas^2 yr^{-2} ($2 \le \ell \le 12$).

Figure 10 shows the corrected estimates $\hat{C}^t_\ell$ and $\hat{C}^s_\ell$ for $\ell = 2, \ldots, 12$. The ragged shape of (especially) the toroidal spectrum is very real, as shown by the error bars. One sees that the

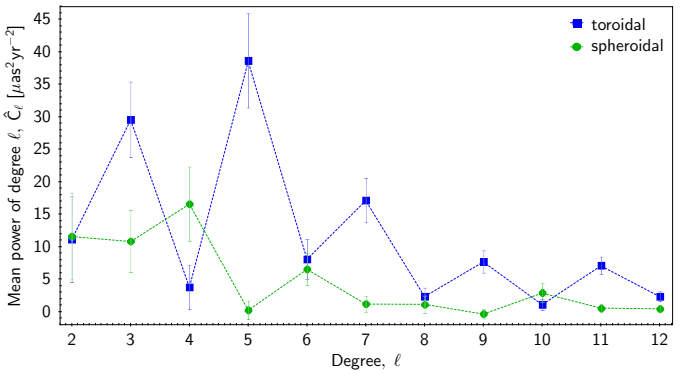

Fig. 10. Corrected mean toroidal and spheroidal powers $\hat{C}^t_\ell$ and $\hat{C}^s_\ell$ of the systematics in proper motions of the *Gaia*-CRF3 sources with five-parameter solutions. The error bars show the ±1 standard deviation of each point, calculated on the assumption that the $2\ell + 1$ contributing VSH coefficients have Gaussian errors.

toroidal powers are substantially larger than the spheroidal ones for the odd degrees, while they are more similar for even degrees. We have no explanation for this but presume it is related to the specific parameters of the scanning law and basic angle.

In analogy with Eq. (32) of Lindegren et al. (2021b), it is convenient to consider the cumulative characteristics

$$R^t(\ell_{\rm max}) = \sqrt{\frac{1}{4\pi}\sum_{\ell=2}^{\ell_{\rm max}}(2\ell+1)\,\hat{C}^t_\ell}\,, \tag{8}$$

$$R^s(\ell_{\rm max}) = \sqrt{\frac{1}{4\pi}\sum_{\ell=2}^{\ell_{\rm max}}(2\ell+1)\,\hat{C}^s_\ell}\,, \tag{9}$$

$$R(\ell_{\rm max}) = \sqrt{\frac{1}{4\pi}\sum_{\ell=2}^{\ell_{\rm max}}(2\ell+1)\left(\hat{C}^t_\ell + \hat{C}^s_\ell\right)} \tag{10}$$

quantifying the RMS variations of the toroidal, spheroidal and total proper motion systematics on angular scales $\gtrsim 180°/\ell_{\rm max}$ (Fig. 11). We note that the toroidal components completely dominate the systematics for $\ell \ge 5$.

Summarising all this information we can claim that the frame-independent RMS of the vector field representing systematic errors in proper motions is $\lesssim$12 μas yr^{-1} down to angular scales of ∼15°. This level of systematic errors is low compared to the random errors of individual sources given in Table 2 and made it possible to reliably measure the subtle physical effect of the acceleration of the Solar system in *Gaia* EDR3 (Gaia Collaboration 2021b).

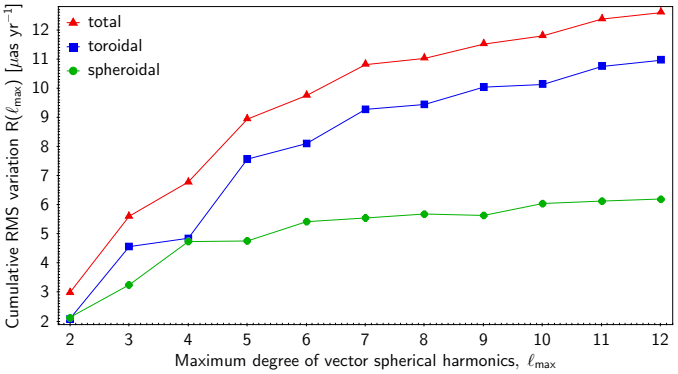

Fig. 11. Corrected cumulative angular power spectrum of the systematic in proper motions of the *Gaia*-CRF3 sources with five-parameter solutions.

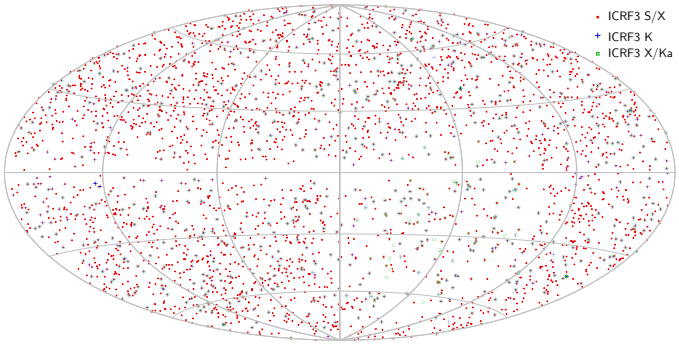

Fig. 12. Distribution of the ICRF3 sources in *Gaia*-CRF3 in Galactic coordinates. Sources from ICRF3 S/X are shown by red dots, from ICRF3 K by blue crosses, and from ICRF3 X/Ka by turquoise squares. This map uses a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left.

4. *Gaia*-CRF3 counterparts of ICRF3 sources

Radio-loud sources in *Gaia*-CRF3 deserve special attention as they provide a direct link between the ICRF implementations in the radio and optical regimes. In this section we focus on the properties of the *Gaia*-CRF3 sources matched to ICRF3 (Charlot et al. 2020) and a statistical comparison of the optical and radio positions.

4.1. Statistics of ICRF3 sources in Gaia-CRF3

Gaia-CRF3 contains a total of 4723 radio-loud sources, of which 3181 are optical counterparts of radio sources in ICRF3 (Charlot et al. 2020), while an additional 1542 sources are found only in OCARS. Because the ICRF3 sources have the best quality of the astrometry in the radio domain and represent the official realisation of the ICRF in the radio, we restrict the discussion in this section to the 3181 *Gaia*-CRF3 sources in common with ICRF3. This set is the union of 3142, 660, and 576 sources matching the ICRF3 sources in the S/X, K, and X/Ka bands, respectively (69.3, 80.1 and 85.0% of the corresponding catalogues). Most of the analysis below is further limited to the 3142 sources in ICRF3 S/X. Very few additional *Gaia* EDR3 sources are found only in the K or X/Ka bands of the ICRF3 (24 and 31 sources, respectively), and their overall properties are not too different from those of the S/X set.

From the sky distribution of the 3181 ICRF3 sources in *Gaia*-CRF3 (Fig. 12) we see that a band along the Galactic plane and the region with $\delta < -30°$ (part of the fourth quadrant of the map in the Galactic coordinates) are underpopulated. This reflects the strong optical extinction in the Galactic plane and the fact that ICRF3 itself has a lower density of sources for $\delta < -30°$.

The left panel in Fig. 13 shows the magnitude distribution of the 3142 *Gaia*-CRF3 sources in common with ICRF3 S/X and for the subset of 259 sources that are defining sources in ICRF3. As expected, there is no marked difference between the two sets, since the optical brightness plays no role in the ICRF3 selection of defining sources. The median $G = 18.9$ is about one magnitude brighter than for the whole *Gaia*-CRF3 (Table 2).

The central panel in Fig. 13 shows the distribution of the positional uncertainty $\sigma_{\rm pos,max}$ in *Gaia*-CRF3 for the same selections. The median uncertainty is 194 μas for the whole set of ICRF3 S/X sources, and 162 μas for the defining sources. These values are less than one half of the median positional uncertainty in the full *Gaia*-CRF3 sample, which can be attributed to their different distributions in G. Indeed, as shown by the middle panel of Fig. 8, the ICRF3 sources follow the same general variation of $\sigma_{\rm pos,max}$ with G as other *Gaia*-CRF3 sources. In other words, nothing in the *Gaia* astrometry suggests that the ICRF3 are better or worse than other *Gaia*-CRF3 sources of comparable magnitude.

The right panel in Fig. 13 shows the distribution of the positional uncertainty in the radio domain for the same sources. By coincidence, the median positional uncertainty of the 3142 sources in ICRF3, 197 μas, is practically the same as in *Gaia*-CRF3. For the subset of defining sources, however, the positional uncertainties in ICRF3 are much smaller (the median is 39 μas), as these sources were selected precisely for their high radio-astrometric quality. In *Gaia*-CRF3 the defining sources are only marginally more accurate than the non-defining sources, consistent with the former being, on average, about 0.2 mag brighter (cf. Fig. 13, left panel).

4.2. Positional differences

We now compare the individual positions of ICRF3 S/X sources in ICRF3 and *Gaia*-CRF3. Thanks to the high positional accuracy in both catalogues and the small cross-matching radius $\Delta_{\rm max} = 100$ mas used, cross-matching errors between the two catalogues are highly unlikely and we can assume that the optical and radio emissions being compared originate in broadly the same extragalactic object. Positional differences $\Delta\alpha* = (\alpha_{\rm EDR3} - \alpha_{\rm ICRF3})\cos\delta$ and $\Delta\delta = \delta_{\rm EDR3} - \delta_{\rm ICRF3}$ are computed after propagating the ICRF3 positions from the ICRF3 epoch (J2015.0) to the *Gaia*-CRF3 epoch (J2016.0), using the precepts described in Charlot et al. (2020, Sect. 5.3); this takes into account the small shifts ($\lesssim$5 μas) expected from the Galactic acceleration.

The top panel of Fig. 14 is a scatter diagram of the positional differences. Of the 3142 sources, 127 (4%) have a positional difference >4 mas in either coordinate and fall outside the boundaries of the plot. The median separation $\Delta = (\Delta\alpha*^2 + \Delta\delta^2)^{1/2}$ is 0.516 mas for all 3142 sources, and 0.292 mas for the 259 defining sources (Fig. 15).

The distributions of $\Delta\alpha*$ and $\Delta\delta$ are roughly symmetric around zero with a median offset of +4 μas in right ascension and −6 μas in declination. Similarly, the scatter diagrams of the positional differences versus right ascension and declination (Fig. 16) show no significant systematics or trends in either coordinate. The distributions are very regular with a pronounced central concentration of width below 1 mas, plus a scatter of several hundred outliers at several mas. Furthermore, an analysis of

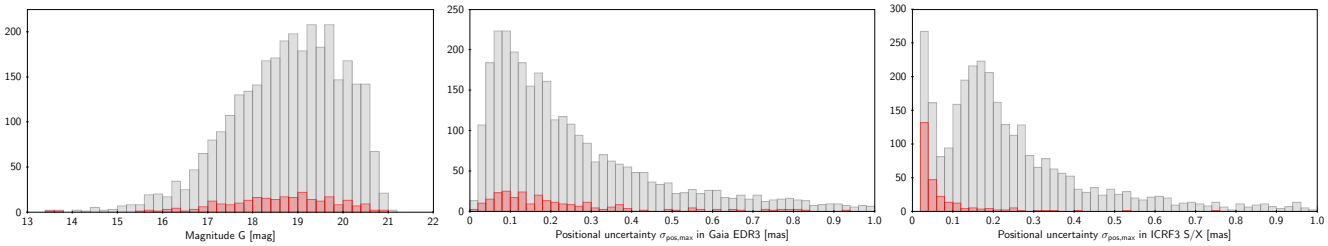

Fig. 13. Distribution of magnitudes and formal positional uncertainties for the sources common to *Gaia*-CRF3 and ICRF3 S/X. The full set of 3142 sources is shown by the grey histograms, and the 259 defining sources by the red histograms. *Left*: distribution of magnitudes G. *Middle*: distribution of positional uncertainties in *Gaia* EDR3. There are 126 sources with $\sigma^{\rm EDR3}_{\rm pos,max} > 1$ mas, of which eight are defining sources in ICRF3. *Right*: distribution of positional uncertainties in ICRF3 S/X. There are 155 sources with $\sigma^{\rm ICRF3}_{\rm pos,max} > 1$ mas, of which one is a defining source.

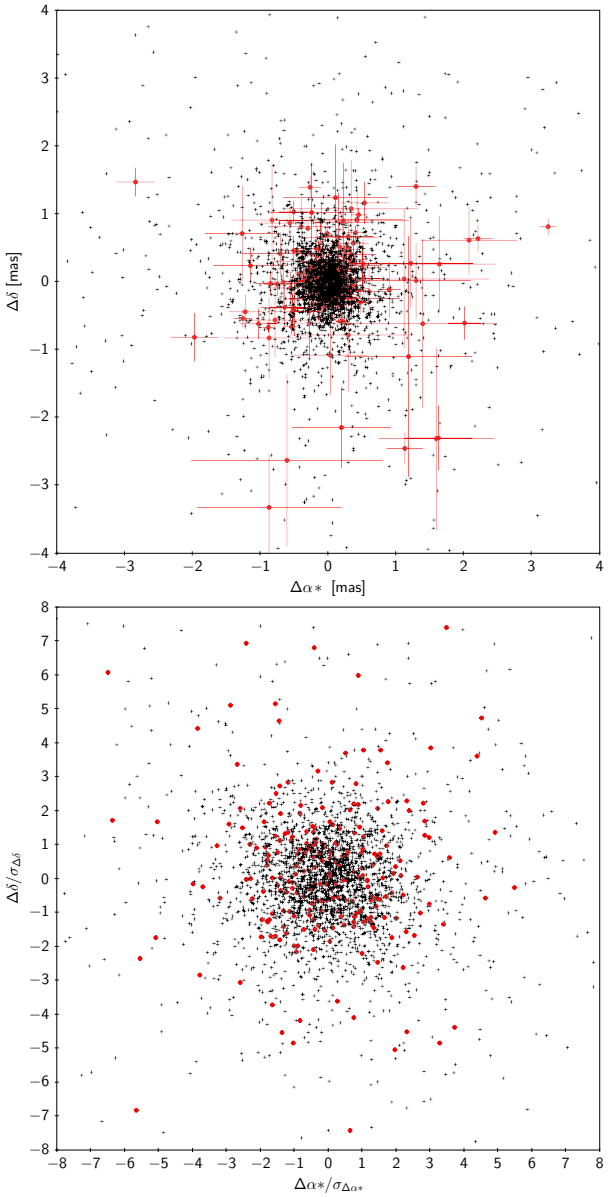

Fig. 14. Positional differences in α and δ between *Gaia*-CRF3 and ICRF3 S/X for the 3142 common sources. The 259 defining sources in ICRF3 are shown as filled red circles, other sources as black crosses. *Top*: differences in mas. Errors bars representing the quadratically combined uncertainties in the two catalogues are shown only for the defining sources. *Bottom*: same differences divided by the quadratically combined uncertainties in the two catalogues. No error bars are shown in this plot, as they would all be ±1 unit.

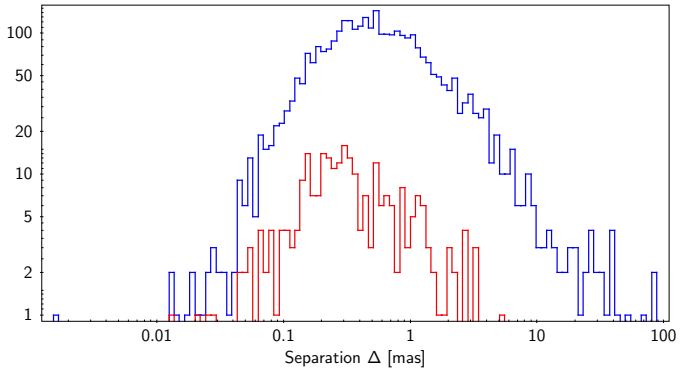

Fig. 15. Histograms of the separations Δ between *Gaia*-CRF3 and ICRF3 S/X for the full set of 3142 common sources (blue; the median is 0.516 mas) and the 259 defining sources (red; the median is 0.292 mas).

the positional differences using VSH does not reveal any significant systematic effects. The statistical significance of the fitted coefficients is low, while the largest effects represented by the VSH expansion are located in the areas with low source density (around the Galactic plane and for $\delta < -30°$).

The lower panel of Fig. 14 shows the same position differences as in the upper panel, but normalised by the quadratic combination of the uncertainties in the ICRF3 and *Gaia*-CRF3 catalogues. If the optical and radio centres of emission coincided, we would expect the normalised position differences to follow a normal distribution with standard deviation only slightly above 1.0. In fact, the distributions are significantly wider (the RSE[3] is 1.92 in right ascension and 1.80 in declination), with fatter tails than expected for a Gaussian: 225 sources fall outside the plotted frame, and more than 100 are outside $\pm 5 \times$ RSE in each coordinate.

Both in ICRF3 and in *Gaia*-CRF3 there are significant correlations between the errors in right ascension and declination, and a useful statistic that takes this into account is

$$X^2_\Delta \equiv \begin{bmatrix} \Delta\alpha* & \Delta\delta \end{bmatrix} \mathrm{Cov}(\mathbf{\Delta})^{-1} \begin{bmatrix} \Delta\alpha* \\ \Delta\delta \end{bmatrix}, \tag{11}$$

where $\mathrm{Cov}(\mathbf{\Delta})$ is the covariance matrix given by Eq. (E.7). This is completely analogous to the statistic X^2_μ for the proper motions, defined in Eq. (3), and like Δ it is independent of the

[3] The robust scatter estimate (RSE) is defined as $\left[2\sqrt{2}\,\mathrm{erf}^{-1}(4/5)\right]^{-1} \approx 0.390152$ times the difference between the 90th and 10th percentiles of the distribution of the variable. For a Gaussian distribution it equals the standard deviation. The RSE is widely used in *Gaia* as a standardised, robust measure of dispersion.

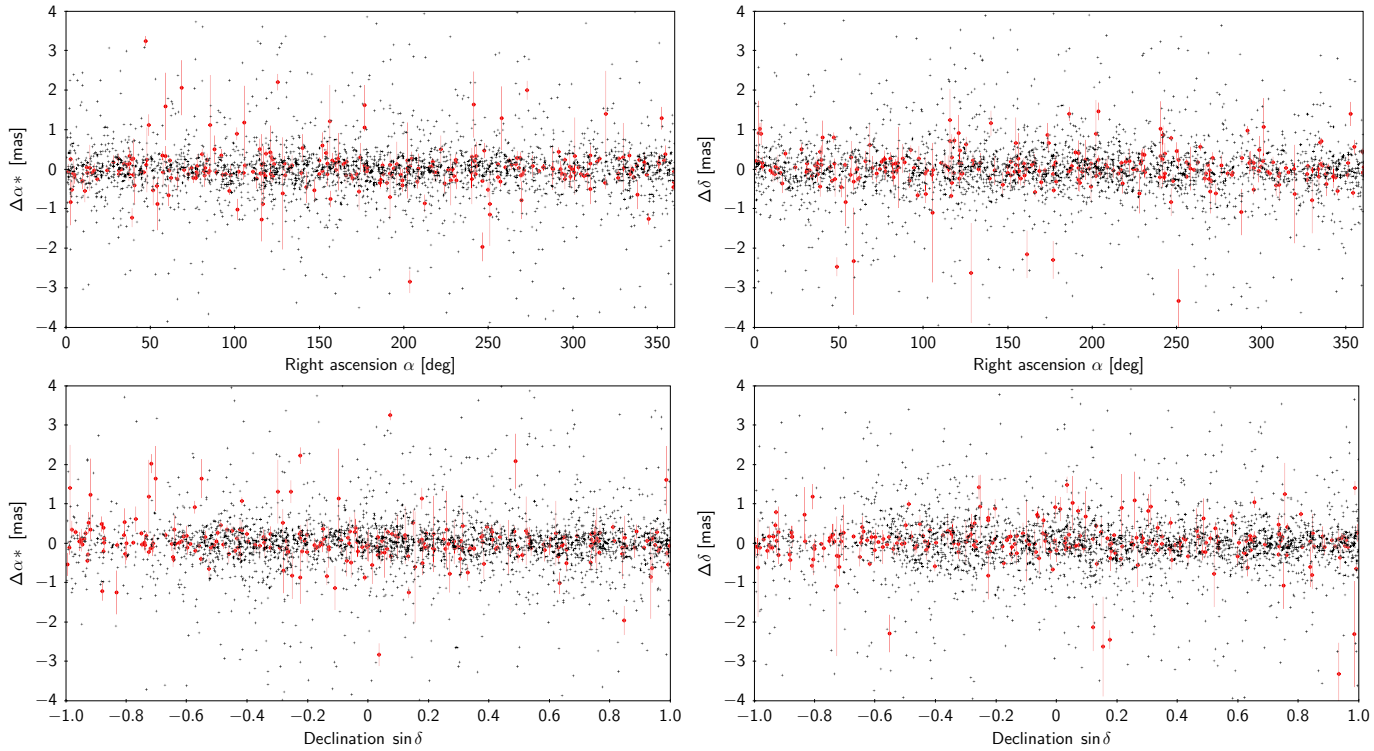

Fig. 16. Positional differences between *Gaia*-CRF3 and ICRF3 S/X versus α (*top*) and $\sin\delta$ (*bottom*). Defining sources in ICRF3 are shown as filled red circles, with errors bars representing the combined positional uncertainties in the two catalogues; other sources are shown as black crosses.

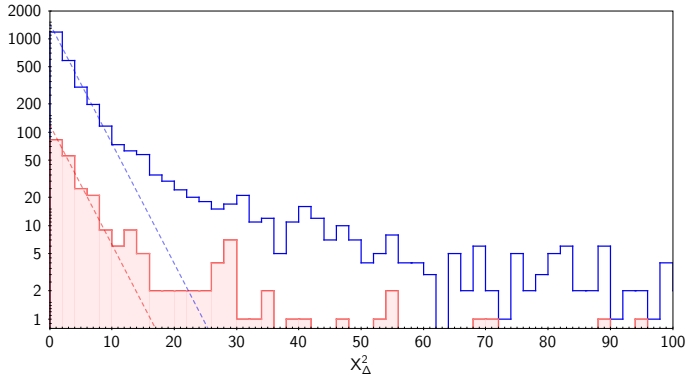

Fig. 17. Histograms of the normalised separations X^2_Δ between *Gaia*-CRF3 and ICRF3 S/X for the full set of 3142 common sources (blue) and the 259 defining sources (red). The dashed curves are the exponential functions described in the text. A total of 195 sources have $X_\Delta > 10$ and fall outside this plot (10 of them are defining sources in ICRF3).

reference frame used[4]. The distributions of X^2_Δ are shown in Fig. 17. For Gaussian errors we expect exponential distributions. The dashed lines are the functions $1450\exp[-\frac{1}{2}(X_\Delta/1.3)^2]$ and $120\exp[-\frac{1}{2}(X_\Delta/1.3)^2]$ fitted by eye to the inner parts of the histograms ($X^2_\Delta \lesssim 10$). The integrals of these exponential are 2450.5 and 202.8, that is 78% of respective sample. One possible interpretation of this is that the (combined) uncertainties are generally underestimated by a factor 1.3, and that about 22% of the objects in the samples have radio–optical positional offsets that are statistically significant at the current level of accuracy in the two catalogues; this fraction is, moreover, the same for the defining sources as for the full sample. Using a criterion such as $X^2_\Delta > 25$ to pick out the individual sources that have a significant offset yields 433 sources, 36 of which are defining ones, or 14% of the respective samples. The angular offsets of these sources range from 0.17 mas to tens of mas, with a median value of 2.3 mas for the 433 sources from the full sample, and 0.84 mas for the 36 defining sources.

We note that the median X^2_μ is slightly higher (2.36) for the 433 sources with $X^2_\Delta > 25$ than for the full sample of 3142 sources (1.82). This could indicate that the positional offsets are not constant over the time scale of the *Gaia* observations.

These results confirm and amplify the conclusion reached already with *Gaia* DR1 and *Gaia* DR2, namely that radio–optical offsets are significant for a large fraction of the ICRF3 sources. The correlation with jets has been well established, for example by Petrov et al. (2019) and Plavin et al. (2019), and the physical signature of offsets is also supported in the studies by Lambert et al. (2021) and Liu et al. (2021).

The question now is therefore not so much the existence of such offsets but rather what consequences they have for the relation between the radio and optical frames. A plain match between the two catalogues leaves sources with true differences in position, much larger than the expected random scatter. These differences will not decrease with future versions of the radio and optical frames. Although this does not impact the quality of either realisation, it could limit the attainable accuracy of their mutual alignment. To circumvent this problem, specific procedures may be required, such as the selection of a common set of reference sources for the alignment. More generally, the importance of physical deviations between the reference frames

[4] The quantity X^2_Δ is important also for another reason: as explained in Appendix E we consider the optical and radio frames to be aligned if the sum of X^2_Δ over the 'good' sources is minimal; cf. Eq. (E.18). The equivalent statistic was introduced by Mignard et al. (2016) already for the analysis of the reference frame in *Gaia* DR1; see their Eqs. (1) and (3)–(5).

at different wavelength bands needs to be better understood. This problem will be even more important for the future high-accuracy astrometric observations in the near infrared (Hobbs et al. 2021).

A closely related question concerns the interpretation and use of the reference frame at different epochs. ICRF3 is the first realisation of the VLBI frame that has an epoch (2015.0) and an official procedure to transform from this epoch to another (Charlot et al. 2020). The reason for the epoch transformation is the effect of the acceleration of the Solar system relative to the remote celestial sources, often referred to as 'Galactic acceleration', although it may contain other components as well. If ignored, the Galactic acceleration will slowly distort the frame at a rate of about 6 μas yr^{-1}. The effect should therefore be taken into account when comparing ICRF3 with the *Gaia*-CRF, as was indeed done for this paper. The transformation depends on three parameters that can be determined by observation; for ICRF3, however, a set of fixed values are specified. In case of *Gaia*, full astrometric solutions including parallaxes and proper motions are computed for all *Gaia*-CRF sources. Each version of *Gaia*-CRF has a particular epoch common to that of the astrometric catalogue of the corresponding *Gaia* DR; for *Gaia*-CRF3 this is 2016.0. No specific correction for the Galactic acceleration is applied or recommended for *Gaia*-CRF. However, the effect is implicitly contained in the measured proper motions and a normal epoch transformation would take also this effect into account, albeit with a positional accuracy that degrades with the epoch difference. In Gaia Collaboration (2021b) the components of the acceleration of the Solar system were determined from the proper motions of *Gaia*-CRF3. Although that determination differs from the ICRF3 prescription by less than 1 μas yr^{-1}, there thus is a conceptual difference between how the effect is treated in the two reference frames.

5. Use of *Gaia*-CRF3 sources in the frame rotator of AGIS 3.2

In the astrometric core solution for *Gaia* (the Astrometric Global Iterative Solution, AGIS; Lindegren et al. 2012), the frame rotator ensures that the reference frame of the astrometric solution complies, in a global sense, with the ICRS requirements. The frame rotator uses a pre-selected set of QSO-like sources, including optical counterparts of ICRF3 sources, to estimate the six rotation parameters (spin and orientation) for the solutions obtained in successive iterations, and corrects them in such a way that the QSO-like sources have no net rotation and are globally aligned with the ICRF3 sources.

The frame rotator sources are a subset of the much larger set of primary sources (about 14.3 million for EDR3) that contribute to the estimation of attitude, calibration, and global parameters in AGIS. Special selection rules apply to the primary sources; in particular, they must have colour information from *Gaia*'s BP and RP photometers of adequate quality for the calibration of colour-dependent instrumental effects. The selection of primary sources, including the frame rotator sources, is revised with each new release of astrometric data, but for logistic reasons it is mainly based on the astrometric and photometric data of the previous release. Thus, for *Gaia* EDR3 the selection of frame rotator sources was primarily based on the 556 869 QSO-like sources, including 2820 ICRF3 counterparts, that constitute the *Gaia*-CRF2 (Gaia Collaboration 2018). However, not all of them had adequate colour information or passed the more stringent astrometric criteria for primary sources in *Gaia* EDR3, and the number of frame rotator sources used for EDR3 is therefore substantially smaller (429 249 sources). All of them have five-parameter solutions in *Gaia* EDR3 and were filtered by a procedure similar to the one used for *Gaia*-CRF3 but using a preliminary version of the *Gaia* EDR3 catalogue known as AGIS 3.1 (Lindegren et al. 2021b).

Details of the frame rotator as used for *Gaia* EDR3 are given in Appendix E. As described in the appendix, the orientation ($\boldsymbol{\varepsilon}$) and spin ($\boldsymbol{\omega}$) are estimated by separate least-squares solutions to the position differences of the ICRF3 sources (for the orientation) and to the proper motions of the many more QSO-like sources (for the spin). The algorithm includes an iterative procedure to identify and eliminate outliers, that is sources that deviate very significantly from the overall fit in one or both coordinates. Thus, we need to distinguish between the sources that were 'considered' by the algorithm, and the subset of non-outliers that was actually 'used' to determine the orientation or spin. The auxiliary table `gaiaedr3.frame_rotator_source` in the *Gaia* Archive lists the 429 249 frame rotator sources for *Gaia* EDR3 and contains flags to indicate which of them were considered and used for the orientation and spin solutions. As specified in the table, 428 034 of the sources considered for the spin were actually used to determine the spin in the final iteration of the primary solution, while 1215 (0.3%) were rejected as outliers. Of the 2269 sources with counterparts in ICRF3 S/X considered for the orientation, 2007 sources were actually used to determine the orientation, while 262 (11.5%) were rejected as outliers. Eight of the sources were rejected in both the spin and orientation solutions.

Because the preliminary AGIS 3.1 astrometry had to be used to select the frame rotator sources, it turns out that 46 of the latter do not pass the stricter criteria for the *Gaia*-CRF3 selection described in Sect. 2. These sources are thus listed in `gaiaedr3.frame_rotator_source` although they are not part of the *Gaia*-CRF3. This slight inconsistency has minimal impact on the resulting reference frame, as only two of the 46 sources were accepted by the frame rotator algorithm for the determination of the spin, and none of them for the determination of the orientation.

If the frame rotator algorithm is applied to this subset of sources, using the astrometric data as given in *Gaia* EDR3, the resulting estimates of the orientation and spin vectors are not strictly zero, as might be expected. For the spin, we obtain $|\boldsymbol{\omega}| \simeq 0.8$ μas yr^{-1} with an uncertainty of about 0.4 μas yr^{-1} in each axis; for the orientation, the result is $|\boldsymbol{\varepsilon}(2016.0)| \simeq 6.2$ μas with an uncertainty of about 7 μas per axis. These deviations from zero can be traced to small differences in the positions and proper motions between the last iteration of the primary solution and the secondary source updates, from which the final data are taken. The reason for this (undesirable) effect is currently under investigation.

If the same algorithm is instead applied to the full set of *Gaia*-CRF3 sources, the result for the spin is

$$\boldsymbol{\omega} = \begin{bmatrix} -3.44 \pm 0.30 \\ +1.57 \pm 0.28 \\ -1.24 \pm 0.32 \end{bmatrix} \mu\text{as yr}^{-1}, \tag{12}$$

in which 1 607 289 (99.6%) of the sources are accepted; for the orientation correction the result is

$$\boldsymbol{\varepsilon}(2016.0) = \begin{bmatrix} +3.41 \pm 6.71 \\ +8.99 \pm 6.50 \\ -1.47 \pm 6.04 \end{bmatrix} \mu\text{as}, \tag{13}$$

in which 2738 (87.1%) of the 3142 ICRF3 S/X sources are accepted. Generally speaking, the result is sensitive to the precise selection of sources at the level of a few μas yr^{-1} in spin and several μas in the orientation.

6. Conclusions and prospects

In this paper we have reported on the construction and properties of the *Gaia*-CRF3, a celestial reference frame materialised in the optical domain resulting from the *Gaia* all-sky survey. Although built on only 33 months of data (from August 2014 to May 2017), the astrometric accuracy is already very similar to what was expected for the 5-yr nominal mission, and with a larger set of QSOs than originally anticipated.

The *Gaia*-CRF3 consists of an astrometric catalogue of more than 1.6 million QSOs selected using a number of existing catalogues, with further filtering based on the *Gaia* data to ensure that the sample is as much as possible free of stellar contaminants. The G magnitude extends from ∼ 13 to ∼ 21 mag, with a peak density at $G \simeq 20.5$ mag. The formal uncertainty is primarily ruled by the G magnitude, with a typical precision of 1 mas at $G = 20.6$, 0.4 mas at $G = 20$ and 0.1 mas at $G = 17.9$. There are 32 000 sources with formal position uncertainty < 0.1 mas and 210 000 with <0.2 mas. *Gaia*-CRF3 is aligned to the radio ICRF3 by minimising the position differences for about 2000 common sources with high-quality solutions.

The comparison with ICRF3 shows that the floor formal accuracies are comparable in the optical and radio domains and better than 0.05 mas for the best measured sources. However the normalised positional differences are not compatible with a standard normal distribution and an estimated ∼20% of the common sources have statistically significant positional radio–optical offsets. For many sources, if not all, these offsets are probably real, although this is hard to prove without a detailed investigation at the source level (Petrov et al. 2019; Plavin et al. 2019). The radio and optical CRFs are both internally very consistent and materialise the ICRS in their respective wavelength range, but in view of the detected offsets it remains unclear how much their mutual consistency can be improved without a better knowledge of the physical processes responsible for the light and radio emissions. It is hoped that high-resolution observations and improved radio and optical positions will further characterise these offsets and lead to a better understanding of the underlying mechanisms (Johnson et al. 2019).

It should be understood that *Gaia*-CRF3 is entirely defined by the *Gaia* EDR3 data for the ∼1.6 million QSO-like sources discussed in this paper. These data include not only the barycentric positions of the sources at the reference epoch 2016.0, but also their proper motions (and parallaxes). Although individually insignificant in the vast majority of cases, these data provide invaluable information on the quality of the data and on possible systematic effects in the reference frame, both of instrumental and astrophysical origin. The reference frame defined by the proper motions of the *Gaia*-CRF3 sources is globally non-rotating at a level of a few μas yr^{-1}. It is not possible to quantify more precisely than this how well the *Gaia*-CRF3 complies with the 'no rotation' condition, because different subsets of the *Gaia*-CRF3 lead to determinations of the spin that differ on this level, which is also consistent with the estimated RMS level of systematics on large angular scales (low values of $\ell_{\max}$) depicted in Fig. 11. As mentioned in Sect. 1, the secondary frame defined by the positions and proper motions of stars in *Gaia* EDR3, in particular at bright magnitudes, may have significantly higher systematics.

Gaia-CRF3 is by far the best realisation to date of the celestial reference frame in the visible. It meets the ICRS requirements for a realisation based on extragalactic sources. The IAU, during its XXXIst General Assembly in August 2021, resolved (Resolution B3) that ICRF3 at radio wavelengths and *Gaia*-CRF3 in the optical are realisations (ICRF) of the International Celestial Reference System (ICRS) to be used as a fundamental standard in their respective domain.

A new version of the *Gaia*-CRF comes out with every release of a new astrometric solution based on a larger volume of *Gaia* data. The current version, *Gaia*-CRF3, is common to *Gaia* EDR3 and *Gaia* DR3, for which the basic astrometric data are identical. Two more releases are expected, *Gaia* DR4 based on 66 months of observations, and *Gaia* DR5 using up to 120 months of data. The positional accuracy improves roughly as $T^{-1/2}$, where T is the time baseline, leading to an improvement by almost a factor two for the final version compared with *Gaia*-CRF3.

It is expected that the systematics will also be improved in the future releases, although this cannot easily be quantified. A dramatic illustration of the improvement from *Gaia*-CRF2 to *Gaia*-CRF3 is the successful use of the latter to determine the acceleration of the Solar System (Gaia Collaboration 2021b). Of paramount importance for the quality of the reference frame is the accuracy of proper motions, which impacts both the astrometric filtering of the sources (thus reducing stellar contamination) and the accuracy of the positions at earlier and later epochs. Proper motions generally improve as $T^{-3/2}$, which could give a factor seven more precise proper motions in the final release. Corresponding improvements can be expected for the determination of the reference frame (especially, its spin) as well as for the various physical effects that may be seen in the proper motions (e.g. the solar system acceleration; see Sect. 8 in Gaia Collaboration 2021b for a brief discussion of other effects).

As mentioned in Sect. 3, up to 4 million QSOs could be recognised by *Gaia*. In future releases the identification of the QSOs will be refined using a combination of all kinds of *Gaia* data: an improved astrometry, spectrophotometry, and variability analysis. Special efforts will be made to better identify QSOs in the crowded areas and to improve the homogeneity of the sky coverage (a reduced source density at the Galactic plane will however persist due to the effects of Galactic extinction). The prospect for the future versions of the *Gaia*-CRF therefore lies not only in a further improvement of positional accuracy, but even more in the number of sources, a better homogeneity of the source distribution of the sky, the purity of the selection and in the overall consistency of the solution.

Acknowledgements. The acknowledgements are available in Appendix A.

References

Andrei, A. H., Souchay, J., Zacharias, N., et al. 2009, A&A, 505, 385
Arias, E. F., Charlot, P., Feissel, M., & Lestrade, J.-F. 1995, A&A, 303, 604
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Bailer-Jones, C. A. L., Andrae, R., Arcay, B., et al. 2013, A&A, 559, A74
Bailer-Jones, C. A. L., Fouesneau, M., & Andrae, R. 2019, MNRAS, 490, 5615
Björck, Å. 1996, Other Titles in Applied Mathematics, Numerical Methods for Least Squares Problems (Philadelphia, PA: Society for Industrial and Applied Mathematics), 102
Bonato, M., Liuzzo, E., Herranz, D., et al. 2019, MNRAS, 485, 1188
Cantat-Gaudin, T., & Brandt, T. D. 2021, A&A, 649, A124
Chang, Y. L., Arsioli, B., Giommi, P., & Padovani, P. 2017, A&A, 598, A17
Charlot, P., Jacobs, C. S., Gordon, D., et al. 2020, A&A, 644, A159
Croom, S. M., Smith, R. J., Boyle, B. J., et al. 2004, MNRAS, 349, 1397
de Bruijne, J. H. J., Allen, M., Azaz, S., et al. 2015, A&A, 576, A74
Delchambre, L. 2018, MNRAS, 473, 1785

ESA. 1997, The Hipparcos and Tycho Catalogues, ESA SP-1200
Flesch, E. W. 2019, ArXiv e-prints [arXiv:1912.05614]
Fricke, W. 1985, Celest. Mech., 36, 207
Fu, Y., Wu, X.-B., Yang, Q., et al. 2021, ApJS, 254, 6
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Mignard, F., et al.) 2018, A&A, 616, A14
Gaia Collaboration (Brown, A. G. A., et al.) 2021a, A&A, 649, A1 (*Gaia* EDR3 SI)
Gaia Collaboration (Klioner, S. A., et al.) 2021b, A&A, 649, A9 (*Gaia* EDR3 SI)
Gaia Collaboration (Luri, X., et al.) 2021c, A&A, 649, A7 (*Gaia* EDR3 SI)
Heintz, K. E., Fynbo, J. P. U., & Høg, E. 2015, A&A, 578, A91
Heintz, K. E., Fynbo, J. P. U., Høg, E., et al. 2018, A&A, 615, L8
Heintz, K. E., Fynbo, J. P. U., Geier, S. J., et al. 2020, A&A, 644, A17
Hobbs, D., Brown, A., Høg, E., et al. 2021, Exp. Astron., 51, 783
Johnson, M., Schinzel, F., Darling, J., et al. 2019, BAAS, 51, 273
Lambert, S., Liu, N., Arias, E. F., et al. 2021, A&A, 651, A64
Lawson, C., & Hanson, R. 1995, Classics in Applied Mathematics, Solving Least Squares Problems (Philadelphia, PA: Society for Industrial and Applied Mathematics), 15
Lindegren, L., Lammers, U., Hobbs, D., et al. 2012, A&A, 538, A78
Lindegren, L., Bastian, U., Biermann, M., et al. 2021a, A&A, 649, A4
Lindegren, L., Klioner, S. A., Hernández, J., et al. 2021b, A&A, 649, A2
Liu, N., Lambert, S. B., Charlot, P., et al. 2021, A&A, 652, A87
Lunz, S., Anderson, J., Xu, M. H., et al. 2021, in EGU General Assembly Conference Abstracts, EGU21–8604
Malkin, Z. 2018, ApJS, 239, 20
Massaro, E., Maselli, A., Leto, C., et al. 2015, Ap&SS, 357, 75
Mignard, F., & Klioner, S. 2012, A&A, 547, A59
Mignard, F., Klioner, S., Lindegren, L., et al. 2016, A&A, 595, A5
Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51
Perryman, M. A. C., Lindegren, L., Kovalevsky, J., et al. 1997, A&A, 500, 501
Petrov, L., Kovalev, Y. Y., & Plavin, A. V. 2019, MNRAS, 482, 3023
Plavin, A. V., Kovalev, Y. Y., & Petrov, L. Y. 2019, ApJ, 871, 143
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Schiff, L. I. 1964, Rev. Mod. Phys., 36, 510
Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Shu, Y., Koposov, S. E., Evans, N. W., et al. 2019, MNRAS, 489, 4741
Souchay, J., Gattano, C., Andrei, A. H., et al. 2019, A&A, 624, A145
Taris, F., Damljanovic, G., Andrei, A., et al. 2018, A&A, 611, A52
van Leeuwen, F. 2007, A&A, 474, 653
Yao, S., Wu, X.-B., Ai, Y. L., et al. 2019, ApJS, 240, 6

[1] Lohrmann Observatory, Technische Universität Dresden, Mommsenstraße 13, 01062 Dresden, Germany
e-mail: sergei.klioner@tu-dresden.de
[2] Lund Observatory, Department of Astronomy and Theoretical Physics, Lund University, Box 43, 22100 Lund, Sweden
[3] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[4] European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[5] Telespazio UK S.L. for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[6] Astronomisches Rechen-Institut, Zentrum für Astronomie der Universität Heidelberg, Mönchhofstr. 12-14, 69120 Heidelberg, Germany
[7] HE Space Operations BV for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[8] INAF – Osservatorio Astrofisico di Torino, via Osservatorio 20, 10025 Pino Torinese (TO), Italy
[9] Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[10] DAPCOM for Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, 08028 Barcelona, Spain
[11] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[12] ATG Europe for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[13] Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands
[14] INAF – Osservatorio astronomico di Padova, Vicolo Osservatorio 5, 35122 Padova, Italy
[15] European Space Agency (ESA), European Space Research and Technology Centre (ESTEC), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[16] GEPI, Observatoire de Paris, Université PSL, CNRS, 5 Place Jules Janssen, 92190 Meudon, France
[17] Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
[18] Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France
[19] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[20] Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
[21] Aurora Technology for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[22] CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[23] Institut d'Astronomie et d'Astrophysique, Université Libre de Bruxelles CP 226, Boulevard du Triomphe, 1050 Brussels, Belgium
[24] F.R.S.-FNRS, Rue d'Egmont 5, 1000 Brussels, Belgium
[25] INAF – Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, 50125 Firenze, Italy
[26] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
[27] European Space Agency, Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands (retired)
[28] University of Turin, Department of Physics, Via Pietro Giuria 1, 10125 Torino, Italy
[29] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, via Piero Gobetti 93/3, 40129 Bologna, Italy
[30] Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium
[31] Observational Astrophysics, Division of Astronomy and Space Physics, Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden
[32] ALTEC S.p.a, Corso Marche, 79, 10146 Torino, Italy
[33] Sednai Sàrl, Geneva, Switzerland
[34] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
[35] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[36] Gaia DPAC Project Office, ESAC, Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[37] SYRTE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, LNE, 61 avenue de l'Observatoire 75014 Paris, France
[38] National Observatory of Athens, I. Metaxa and Vas. Pavlou, Palaia Penteli, 15236 Athens, Greece
[39] IMCCE, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Univ. Lille, 77 av. Denfert-Rochereau, 75014 Paris, France
[40] Serco Gestión de Negocios for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
[41] Institut d'Astrophysique et de Géophysique, Université de Liège, 19c, Allée du 6 Août, 4000 Liège, Belgium
[42] CRAAG - Centre de Recherche en Astronomie, Astrophysique et Géophysique, Route de l'Observatoire Bp 63 Bouzareah 16340 Algiers, Algeria
[43] RHEA for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain

[44] CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Campus de Elviña s/n, A Coruña 15071, Spain
[45] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 11 rue de l'Université, 67000 Strasbourg, France
[46] Kavli Institute for Cosmology Cambridge, Institute of Astronomy, Madingley Road, Cambridge, CB3 0HA, UK
[47] Leibniz Institute for Astrophysics Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
[48] CENTRA, Faculdade de Ciências, Universidade de Lisboa, Edif. C8, Campo Grande, 1749-016 Lisboa, Portugal
[49] Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, 5226 Donald Bren Hall, 92697-3440 CA Irvine, USA
[50] INAF – Osservatorio Astrofisico di Catania, via S. Sofia 78, 95123 Catania, Italy
[51] Dipartimento di Fisica e Astronomia "Ettore Majorana", Università di Catania, Via S. Sofia 64, 95123 Catania, Italy
[52] INAF – Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monte Porzio Catone (Roma), Italy
[53] Space Science Data Center - ASI, Via del Politecnico SNC, 00133 Roma, Italy
[54] Department of Physics, University of Helsinki, PO Box 64, 00014 Helsinki, Finland
[55] Finnish Geospatial Research Institute FGI, Geodeetinrinne 2, 02430 Masala, Finland
[56] Institut UTINAM CNRS UMR6213, Université Bourgogne Franche-Comté, OSU THETA Franche-Comté Bourgogne, Observatoire de Besançon, BP1615, 25010 Besançon Cedex, France
[57] HE Space Operations BV for European Space Agency (ESA), Keplerlaan 1, 2201AZ, Noordwijk, The Netherlands
[58] Dpto. de Inteligencia Artificial, UNED, c/ Juan del Rosal 16, 28040 Madrid, Spain
[59] Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Eötvös Loránd Research Network (ELKH), MTA Centre of Excellence, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[60] ELTE Eötvös Loránd University, Institute of Physics, 1117, Pázmány Péter sétány 1A, Budapest, Hungary
[61] Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[62] Department of Astrophysics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands
[63] University of Vienna, Department of Astrophysics, Türkenschanzstraße 17, A1180 Vienna, Austria
[64] Institute of Physics, Laboratory of Astrophysics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, 1290 Versoix, Switzerland
[65] Kapteyn Astronomical Institute, University of Groningen, Landleven 12, 9747 AD Groningen, The Netherlands
[66] School of Physics and Astronomy / Space Park Leicester, University of Leicester, University Road, Leicester LE1 7RH, UK
[67] Thales Services for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[68] Depto. Estadística e Investigación Operativa. Universidad de Cádiz, Avda. República Saharaui s/n, 11510 Puerto Real, Cádiz, Spain
[69] Center for Research and Exploration in Space Science and Technology, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore MD, USA
[70] GSFC – Goddard Space Flight Center, Code 698, 8800 Greenbelt Rd, 20771 MD Greenbelt, USA
[71] EURIX S.r.l., Corso Vittorio Emanuele II 61, 10128 Torino, Italy
[72] Porter School of the Environment and Earth Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[73] Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS 15, Cambridge, MA 02138, USA
[74] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[75] LFCA/DAS, Universidad de Chile, CNRS, Casilla 36-D, Santiago, Chile
[76] SISSA - Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, 34136 Trieste, Italy
[77] Telespazio for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[78] University of Turin, Department of Computer Sciences, Corso Svizzera 185, 10149 Torino, Italy
[79] Dpto. de Matemática Aplicada y Ciencias de la Computación, Univ. de Cantabria, ETS Ingenieros de Caminos, Canales y Puertos, Avda. de los Castros s/n, 39005 Santander, Spain
[80] Centro de Astronomía - CITEVA, Universidad de Antofagasta, Avenida Angamos 601, Antofagasta 1270300, Chile
[81] DLR Gesellschaft für Raumfahrtanwendungen (GfR) mbH Münchener Straße 20 , 82234 Weßling, Germany
[82] Centre for Astrophysics Research, University of Hertfordshire, College Lane, AL10 9AB, Hatfield, UK
[83] University of Turin, Mathematical Department "G.Peano", Via Carlo Alberto 10, 10123 Torino, Italy
[84] INAF – Osservatorio Astronomico d'Abruzzo, Via Mentore Maggini, 64100 Teramo, Italy
[85] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão, 1226, Cidade Universitaria, 05508-900 São Paulo, SP, Brazil
[86] APAVE SUDEUROPE SAS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[87] Mésocentre de calcul de Franche-Comté, Université de Franche-Comté, 16 route de Gray, 25030 Besançon Cedex, France
[88] ATOS for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[89] School of Physics and Astronomy, Tel Aviv University, Tel Aviv 6997801, Israel
[90] Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK
[91] Centre de Données Astronomique de Strasbourg, Strasbourg, France
[92] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Géoazur, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[93] Institute for Computational Cosmology, Department of Physics, Durham University, Durham DH1 3LE, UK
[94] European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching, Germany
[95] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85748 Garching, Germany
[96] Data Science and Big Data Lab, Pablo de Olavide University, 41013, Seville, Spain
[97] Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell 1-3, 08034-Barcelona, Spain
[98] ETSE Telecomunicación, Universidade de Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Galicia, Spain
[99] Asteroid Engineering Laboratory, Space Systems, Luleå University of Technology, Box 848, 981 28 Kiruna, Sweden
[100] Vera C Rubin Observatory, 950 N. Cherry Avenue, Tucson, AZ 85719, USA
[101] Department of Astrophysics, Astronomy and Mechanics, National and Kapodistrian University of Athens, Panepistimiopolis, Zografos, 15783 Athens, Greece
[102] TRUMPF Photonic Components GmbH, Lise-Meitner-Straße 13, 89081 Ulm, Germany
[103] IAC – Instituto de Astrofisica de Canarias, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain
[104] Department of Astrophysics, University of La Laguna, Via Láctea s/n, 38200 La Laguna S.C., Tenerife, Spain
[105] Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
[106] Radagast Solutions, Simon Vestdijkpad 24, 2321WD Leiden, The Netherlands

[107] Laboratoire Univers et Particules de Montpellier, CNRS Université Montpellier, Place Eugène Bataillon, CC72, 34095 Montpellier Cedex 05, France
[108] Université de Caen Normandie, Côte de Nacre Boulevard Maréchal Juin, 14032 Caen, France
[109] LESIA, Observatoire de Paris, Université PSL, CNRS, Sorbonne Université, Université de Paris, 5 Place Jules Janssen, 92190 Meudon, France
[110] SRON Netherlands Institute for Space Research, Niels Bohrweg 4, 2333 CA Leiden, The Netherlands
[111] Astronomical Observatory, University of Warsaw, Al. Ujazdowskie 4, 00-478 Warszawa, Poland
[112] Scalian for CNES Centre Spatial de Toulouse, 18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
[113] Université Rennes, CNRS, IPR (Institut de Physique de Rennes) - UMR 6251, 35000 Rennes, France
[114] INAF – Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[115] Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, PR China
[116] University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Shijingshan District, Beijing 100049, PR China
[117] Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen Ø, Denmark
[118] DXC Technology, Retortvej 8, 2500 Valby, Denmark
[119] Las Cumbres Observatory, 6740 Cortona Drive Suite 102, Goleta, CA 93117, USA
[120] CIGUS CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Paseo de Ronda 51, 15071, A Coruña, Spain
[121] Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool L3 5RF, UK
[122] IPAC, Mail Code 100-22, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA
[123] IRAP, Université de Toulouse, CNRS, UPS, CNES, 9 Av. colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France
[124] MTA CSFK Lendület Near-Field Cosmology Research Group, Konkoly Observatory, MTA Research Centre for Astronomy and Earth Sciences, Konkoly Thege Miklós út 15-17, 1121 Budapest, Hungary
[125] Departmento de Física de la Tierra y Astrofísica, Universidad Complutense de Madrid, 28040 Madrid, Spain
[126] Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[127] Villanova University, Department of Astrophysics and Planetary Science, 800 E Lancaster Avenue, Villanova PA 19085, USA
[128] INAF – Osservatorio Astronomico di Brera, via E. Bianchi, 46, 23807 Merate (LC), Italy
[129] STFC, Rutherford Appleton Laboratory, Harwell, Didcot, OX11 0QX, UK
[130] Charles University, Faculty of Mathematics and Physics, Astronomical Institute of Charles University, V Holesovickach 2, 18000 Prague, Czech Republic
[131] Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Rehovot 7610001, Israel
[132] Department of Astrophysical Sciences, 4 Ivy Lane, Princeton University, Princeton NJ 08544, USA
[133] Departamento de Astrofísica, Centro de Astrobiología (CSIC-INTA), ESA-ESAC. Camino Bajo del Castillo s/n. 28692 Villanueva de la Cañada, Madrid, Spain
[134] naXys, University of Namur, Rempart de la Vierge, 5000 Namur, Belgium
[135] CGI Deutschland B.V. & Co. KG, Mornewegstr. 30, 64293 Darmstadt, Germany
[136] Institute of Global Health, University of Geneva, Geneva Switzerland
[137] Astronomical Observatory Institute, Faculty of Physics, Adam Mickiewicz University, Poznań, Poland
[138] H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[139] Department of Physics and Astronomy G. Galilei, University of Padova, Vicolo dell'Osservatorio 3, 35122 Padova, Italy
[140] CERN, Esplanade des Particules 1, P.O. Box 1211, Geneva, Switzerland
[141] Applied Physics Department, Universidade de Vigo, 36310 Vigo, Spain
[142] Association of Universities for Research in Astronomy, 1331 Pennsylvania Ave. NW, Washington, DC 20004, USA
[143] European Southern Observatory, Alonso de Córdova 3107, Casilla 19, Santiago, Chile
[144] Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, 98bis bd. Arago, 75014 Paris, France
[145] Faculty of Mathematics and Physics, University of Ljubljana, Jadranska ulica 19, 1000 Ljubljana, Slovenia

Appendix A: Acknowledgments

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 https://archives.esac.esa.int/gaia.

The *Gaia* mission and data processing have financially been supported by, in alphabetical order by country:

the Algerian Centre de Recherche en Astronomie, Astrophysique et Géophysique of Bouzareah Observatory; the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF) Hertha Firnberg Programme through grants T359, P20046, and P23737; the BELgian federal Science Policy Office (BELSPO) through various PROgramme de Développement d'Expériences scientifiques (PRODEX) grants and the Polish Academy of Sciences - Fonds Wetenschappelijk Onderzoek through grant VS.091.16N, and the Fonds de la Recherche Scientifique (FNRS), and the Research Council of Katholieke Universiteit (KU) Leuven through grant C16/18/005 (Pushing AsteRoseismology to the next level with TESS, GaiA, and the Sloan DIgital Sky SurvEy – PARADISE); the Brazil-France exchange programmes Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) - Comité Français d'Evaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB); the Chilean Agencia Nacional de Investigación y Desarrollo (ANID) through Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) Regular Project 1210992 (L. Chemin); the National Natural Science Foundation of China (NSFC) through grants 11573054, 11703065, and 12173069, the China Scholarship Council through grant 201806040200, and the Natural Science Foundation of Shanghai through grant 21ZR1474100; the Tenure Track Pilot Programme of the Croatian Science Foundation and the École Polytechnique Fédérale de Lausanne and the project TTP-2018-07-1171 'Mining the Variable Sky', with the funds of the Croatian-Swiss Research Programme; the Czech-Republic Ministry of Education, Youth, and Sports through grant LG 15010 and INTER-EXCELLENCE grant LTAUSA18093, and the Czech Space Office through ESA PECS contract 98058; the Danish Ministry of Science; the Estonian Ministry of Education and Research through grant IUT40-1; the European Commission's Sixth Framework Programme through the European Leadership in Space Astrometry (ELSA) Marie Curie Research Training Network (MRTN-CT-2006-033481), through Marie Curie project PIOF-GA-2009-255267 (Space AsteroSeismology & RR Lyrae stars, SAS-RRL), and through a Marie Curie Transfer-of-Knowledge (ToK) fellowship (MTKD-CT-2004-014188); the European Commission's Seventh Framework Programme through grant FP7-606740 (FP7-SPACE-2013-1) for the *Gaia* European Network for Improved data User Services (GENIUS) and through grant 264895 for the *Gaia* Research for European Astronomy Training (GREAT-ITN) network; the European Cooperation in Science and Technology (COST) through COST Action CA18104 'Revealing the Milky Way with *Gaia* (MW-Gaia)'; the European Research Council (ERC) through grants 320360, 647208, and 834148 and through the European Union's Horizon 2020 research and innovation and excellent science programmes through Marie Skłodowska-Curie grant 745617 (Our Galaxy at full HD – Gal-HD) and 895174 (The build-up and fate of self-gravitating systems in the Universe) as well as grants 687378 (Small Bodies: Near and Far), 682115 (Using the Magellanic Clouds to Understand the Interaction of Galaxies), 695099 (A sub-percent distance scale from binaries and Cepheids – CepBin), 716155 (Structured ACCREtion Disks – SACCRED), 951549 (Sub-percent calibration of the extragalactic distance scale in the era of big surveys – UniverScale), and 101004214 (Innovative Scientific Data Exploration and Exploitation Applications for Space Sciences – EXPLORE); the European Science Foundation (ESF), in the framework of the *Gaia* Research for European Astronomy Training Research Network Programme (GREAT-ESF); the European Space Agency (ESA) in the framework of the *Gaia* project, through the Plan for European Cooperating States (PECS) programme through contracts C98090 and 4000106398/12/NL/KML for Hungary, through contract 4000115263/15/NL/IB for Germany, and through PROgramme de Développement d'Expériences scientifiques (PRODEX) grant 4000127986 for Slovenia; the Academy of Finland through grants 299543, 307157, 325805, 328654, 336546, and 345115 and the Magnus Ehrnrooth Foundation; the French Centre National d'Études Spatiales (CNES), the Agence Nationale de la Recherche (ANR) through grant ANR-10-IDEX-0001-02 for the 'Investissements d'avenir' programme, through grant ANR-15-CE31-0007 for project 'Modelling the Milky Way in the *Gaia* era' (MOD4Gaia), through grant ANR-14-CE33-0014-01 for project 'The Milky Way disc formation in the *Gaia* era' (ARCHEOGAL), through grant ANR-15-CE31-0012-01 for project 'Unlocking the potential of Cepheids as primary distance calibrators' (UnlockCepheids), through grant ANR-19-CE31-0017 for project 'Secular evolution of galxies' (SEGAL), and through grant ANR-18-CE31-0006 for project 'Galactic Dark Matter' (GaDaMa), the Centre National de la Recherche Scientifique (CNRS) and its SNO *Gaia* of the Institut des Sciences de l'Univers (INSU), its Programmes Nationaux: Cosmologie et Galaxies (PNCG), Gravitation Références Astronomie Métrologie (PNGRAM), Planétologie (PNP), Physique et Chimie du Milieu Interstellaire (PCMI), and Physique Stellaire (PNPS), the 'Action Fédératrice *Gaia*' of the Observatoire de Paris, the Région de Franche-Comté, the Institut National Polytechnique (INP) and the Institut National de Physique nucléaire et de Physique des Particules (IN2P3) co-funded by CNES; the German Aerospace Agency (Deutsches Zentrum für Luft- und Raumfahrt e.V., DLR) through grants 50QG0501, 50QG0601, 50QG0602, 50QG0701, 50QG0901, 50QG1001, 50QG1101, 50QG1401, 50QG1402, 50QG1403, 50QG1404, 50QG1904, 50QG2101, 50QG2102, and 50QG2202, and the Centre for Information Services and High Performance Computing (ZIH) at the Technische Universität Dresden for generous allocations of computer time; the Hungarian Academy of Sciences through the Lendület Programme grants LP2014-17 and LP2018-7 and the Hungarian National Research, Development, and Innovation Office (NKFIH) through grant KKP-137523 ('SeismoLab'); the Science Foundation Ireland (SFI) through a Royal Society - SFI University Research Fellowship (M. Fraser); the Israel Ministry of Science and Technology through grant 3-18143 and the Tel Aviv University Center for Artificial Intelligence and Data Science (TAD) through a grant; the Agenzia Spaziale Italiana (ASI) through contracts I/037/08/0, I/058/10/0, 2014-025-R.0, 2014-025-R.1.2015, and 2018-24-HH.0 to the Italian Istituto Nazionale di Astrofisica (INAF), contract 2014-049-R.0/1/2 to INAF for the Space Science Data Centre (SSDC, formerly known as the ASI Science Data Center, ASDC), contracts I/008/10/0, 2013/030/I.0, 2013-030-I.0.1-2015, and 2016-17-I.0 to the Aerospace

Logistics Technology Engineering Company (ALTEC S.p.A.), INAF, and the Italian Ministry of Education, University, and Research (Ministero dell'Istruzione, dell'Università e della Ricerca) through the Premiale project 'MIning The Cosmos Big Data and Innovative Italian Technology for Frontier Astrophysics and Cosmology' (MITiC); the Netherlands Organisation for Scientific Research (NWO) through grant NWO-M-614.061.414, through a VICI grant (A. Helmi), and through a Spinoza prize (A. Helmi), and the Netherlands Research School for Astronomy (NOVA); the Polish National Science Centre through HARMONIA grant 2018/30/M/ST9/00311 and DAINA grant 2017/27/L/ST9/03221 and the Ministry of Science and Higher Education (MNiSW) through grant DIR/WK/2018/12; the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through national funds, grants SFRH/BD/128840/2017 and PTDC/FIS-AST/30389/2017, and work contract DL 57/2016/CP1364/CT0006, the Fundo Europeu de Desenvolvimento Regional (FEDER) through grant POCI-01-0145-FEDER-030389 and its Programa Operacional Competitividade e Internacionalização (COMPETE2020) through grants UIDB/04434/2020 and UIDP/04434/2020, and the Strategic Programme UIDB/00099/2020 for the Centro de Astrofísica e Gravitação (CENTRA); the Slovenian Research Agency through grant P1-0188; the Spanish Ministry of Economy (MINECO/FEDER, UE), the Spanish Ministry of Science and Innovation (MICIN), the Spanish Ministry of Education, Culture, and Sports, and the Spanish Government through grants BES-2016-078499, BES-2017-083126, BES-C-2017-0085, ESP2016-80079-C2-1-R, ESP2016-80079-C2-2-R, FPU16/03827, PDC2021-121059-C22, RTI2018-095076-B-C22, and TIN2015-65316-P ('Computación de Altas Prestaciones VII'), the Juan de la Cierva Incorporación Programme (FJCI-2015-2671 and IJC2019-04862-I for F. Anders), the Severo Ochoa Centre of Excellence Programme (SEV2015-0493), and MICIN/AEI/10.13039/501100011033 (and the European Union through European Regional Development Fund 'A way of making Europe') through grant RTI2018-095076-B-C21, the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia 'María de Maeztu') through grant CEX2019-000918-M, the University of Barcelona's official doctoral programme for the development of an R+D+i project through an Ajuts de Personal Investigador en Formació (APIF) grant, the Spanish Virtual Observatory through project AyA2017-84089, the Galician Regional Government, Xunta de Galicia, through grants ED431B-2021/36, ED481A-2019/155, and ED481A-2021/296, the Centro de Investigación en Tecnologías de la Información y las Comunicaciones (CITIC), funded by the Xunta de Galicia and the European Union (European Regional Development Fund – Galicia 2014-2020 Programme), through grant ED431G-2019/01, the Red Española de Supercomputación (RES) computer resources at MareNostrum, the Barcelona Supercomputing Centre - Centro Nacional de Supercomputación (BSC-CNS) through activities AECT-2017-2-0002, AECT-2017-3-0006, AECT-2018-1-0017, AECT-2018-2-0013, AECT-2018-3-0011, AECT-2019-1-0010, AECT-2019-2-0014, AECT-2019-3-0003, AECT-2020-1-0004, and DATA-2020-1-0010, the Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya through grant 2014-SGR-1051 for project 'Models de Programació i Entorns d'Execució Parallels' (MPEXPAR), and Ramon y Cajal Fellowship RYC2018-025968-I funded by MICIN/AEI/10.13039/501100011033 and the European Science Foundation ('Investing in your future'); the Swedish National Space Agency (SNSA/Rymdstyrelsen); the Swiss State Secretariat for Education, Research, and Innovation through the Swiss Activités Nationales Complémentaires and the Swiss National Science Foundation through an Eccellenza Professorial Fellowship (award PCEFP2_194638 for R. Anderson); the United Kingdom Particle Physics and Astronomy Research Council (PPARC), the United Kingdom Science and Technology Facilities Council (STFC), and the United Kingdom Space Agency (UKSA) through the following grants to the University of Bristol, the University of Cambridge, the University of Edinburgh, the University of Leicester, the Mullard Space Sciences Laboratory of University College London, and the United Kingdom Rutherford Appleton Laboratory (RAL): PP/D006511/1, PP/D006546/1, PP/D006570/1, ST/I000852/1, ST/J005045/1, ST/K00056X/1, ST/K000209/1, ST/K000756/1, ST/L006561/1, ST/N000595/1, ST/N000641/1, ST/N000978/1, ST/N001117/1, ST/S000089/1, ST/S000976/1, ST/S000984/1, ST/S001123/1, ST/S001948/1, ST/S001980/1, ST/S002103/1, ST/V000969/1, ST/W002469/1, ST/W002493/1, ST/W002671/1, ST/W002809/1, and EP/V520342/1.

Further acknowledgments can be found at https://gea.esac.esa.int/archive/documentation/GEDR3/Miscellaneous/sec_acknowl/.

Appendix B: Source distributions in external catalogues

Figure B.1 shows the distribution of the sources in the 17 external catalogues used for *Gaia*-CRF3 as discussed in Sect. 2 and listed in Table 1. The maps show all the sources in the respective catalogue, irrespective of whether they are also in *Gaia*-CRF3. For the VLBI and blazar catalogues (ICRF3 S/X, ICRF3 K, ICRF3 X/Ka, Roma-BZCAT, 2WHSP, and the ALMA calibrators) that have a small source density on the sky, individual sources are plotted. All three parts of ICRF3 are shown on one plot using different colours. For the other catalogues, the maps show the density of sources per square degree, computed from the source counts per pixel using HEALPix of level 6 (pixel size $\simeq 0.84\,\mathrm{deg}^2$). One sees the strong inhomogeneity of the input catalogues caused by the special observational programmes not covering the whole sky (e.g. SDSS and LAMOST), by the observational schedule of the WISE satellite (AllWISE, R90, and C75), and by the artificial de-selection of certain crowded areas applied by the authors of some catalogues (e.g. the Galactic plane and the LMC and SMC areas in R90, C75, and Gaia-unWISE). It is also obvious from the maps that not all catalogues are independent. For example, one directly sees that LQRF and LQAC-5 considered the results of 2QZ and that SDSS DR14Q was used in LQAC-5. Detailed information on all these aspects can be found in the corresponding original publications cited in Table 1. All these inhomogeneities contribute to the resulting inhomogeneity of *Gaia*-CRF3, the density of which is shown in Fig. 4.

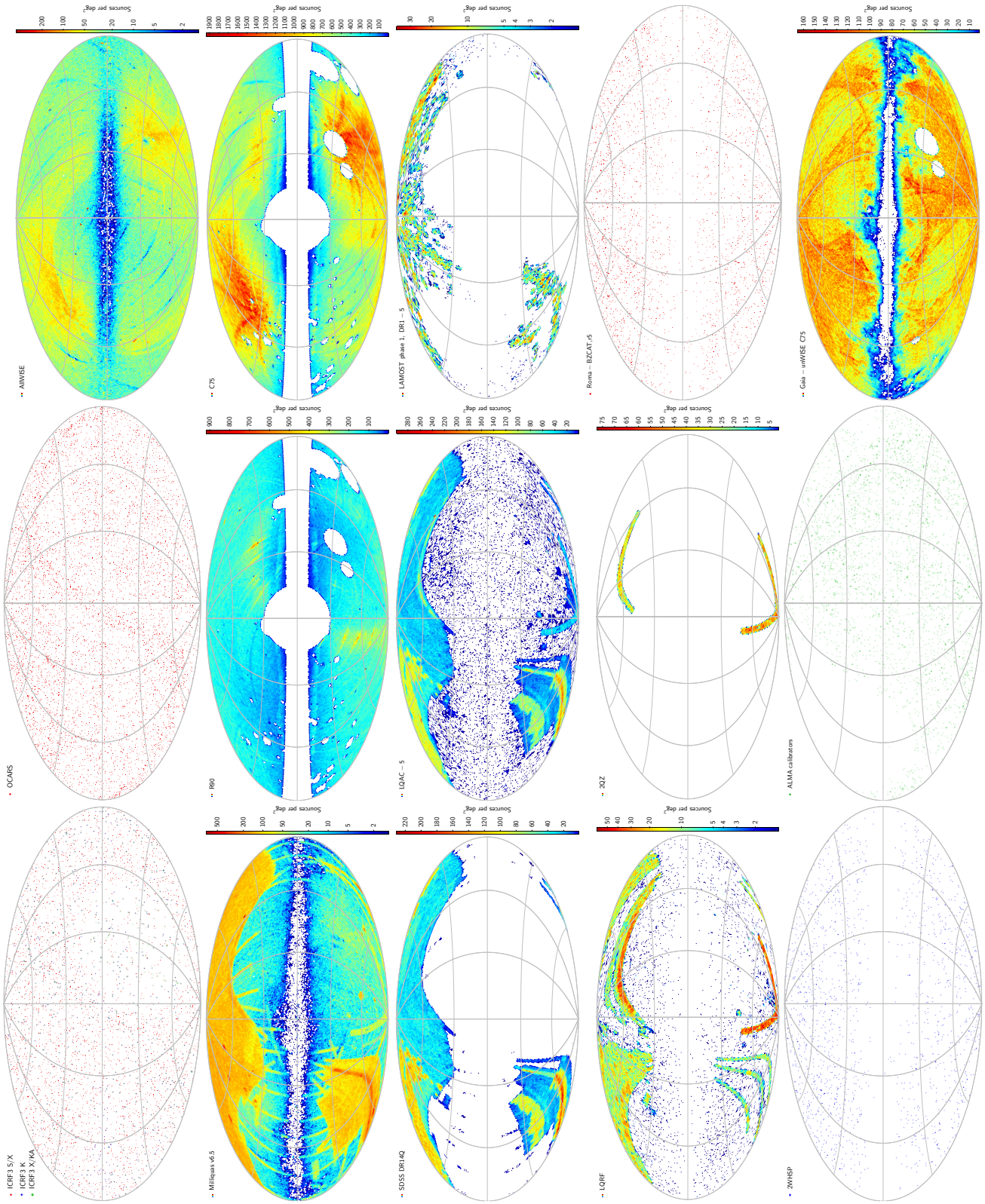

Fig. B.1. Distribution of the sources in the external catalogues listed in Table 1. These maps use a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left.

Appendix C: Access to the *Gaia*-CRF3 data

Information on the *Gaia*-CRF3 sources is distributed over several tables in the *Gaia* Archive.[5] Table `gaiaedr3.agn_cross_id` lists all sources (`source_id`) of *Gaia*-CRF3 together with the name of one external catalogue where the source can also be found as well as the name of the source in that catalogue. Table `gaiaedr3.frame_rotator_source` contains information on the sources used by the frame rotator (see Sect. 5) to define the orientation and spin of the *Gaia* EDR3 catalogue. Together with the *Gaia* EDR3 source identifier `source_id`, it contains the boolean flags `considered_for_reference_frame_orientation` and `considered_for_reference_frame_spin` to indicate if the source was considered by the frame rotator algorithm to fix the orientation and spin of the *Gaia* EDR3 catalogue, respectively. The frame rotator algorithm has a mechanism to detect outliers (see Appendix E), and another pair of boolean flags, `used_for_reference_frame_orientation` and `used_for_reference_frame_spin`, to specify which sources were used to fix, respectively, the orientation and rotation of the *Gaia* EDR3 catalogue. The sources that were considered but not used are the ones identified as outliers by the frame rotator algorithm.

Using these two tables and the main *Gaia* EDR3 astrometric table `gaiaedr3.gaia_source`, one can retrieve astrometric data of the *Gaia*-CRF3 sources and their subsets. We give here a few useful examples of ADQL queries that can be directly used in the *Gaia* Archive. The query

```
SELECT agn.source_name_in_catalogue,
       agn.catalogue_name, edr3.*
FROM gaiaedr3.agn_cross_id AS agn
INNER JOIN gaiaedr3.gaia_source AS edr3
USING (source_id)
```

gives a table containing astrometric data for all 1 614 173 *Gaia*-CRF3 sources. Adding a conditional clause at the end of the same query,

```
SELECT agn.source_name_in_catalogue,
       agn.catalogue_name, edr3.*
FROM gaiaedr3.agn_cross_id AS agn
INNER JOIN gaiaedr3.gaia_source AS edr3
USING (source_id)
WHERE edr3.astrometric_params_solved = 31
```

allows one to retrieve the astrometric data for only the 1 215 942 *Gaia*-CRF3 sources that have five-parameter astrometric solutions. This dataset was used for example in Gaia Collaboration (2021b) to measure the acceleration of the Solar System. Changing '31' to '95' in the last line, one gets instead the 398 231 *Gaia*-CRF3 sources with six-parameter astrometric solutions.

For the sources used to define the orientation and spin of the *Gaia* EDR3 catalogue the following two queries are useful. The query

```
SELECT sub.source_name_in_catalogue,
       sub.catalogue_name, edr3.*
FROM (
SELECT frame.source_id, agn.source_name_in_catalogue,
       agn.catalogue_name
FROM gaiaedr3.frame_rotator_source AS frame
LEFT OUTER JOIN gaiaedr3.agn_cross_id AS agn
USING (source_id)
WHERE frame.used_for_reference_frame_orientation='true'
) AS sub
INNER JOIN gaiaedr3.gaia_source AS edr3
USING (source_id)
```

gives data for the 2007 sources that were used to define the orientation of the *Gaia* EDR3 catalogue, while

```
SELECT sub.source_name_in_catalogue,
       sub.catalogue_name, edr3.*
FROM (
SELECT frame.source_id, agn.source_name_in_catalogue,
       agn.catalogue_name
FROM gaiaedr3.frame_rotator_source AS frame
LEFT OUTER JOIN gaiaedr3.agn_cross_id AS agn
USING (source_id)
WHERE frame.used_for_reference_frame_spin = 'true'
) AS sub
INNER JOIN gaiaedr3.gaia_source AS edr3
USING (source_id)
```

gives data for the 428 034 sources used to define the rotation (spin) of *Gaia* EDR3. The field names `used_for_...` can be replaced by `considered_for_...` to get the data of all sources considered by the reference frame algorithm. As discussed in Sect. 5, all sources that were considered by the frame rotator algorithm for fixing the orientation of the *Gaia* EDR3 catalogue also belong to *Gaia*-CRF3. However, among the sources that were considered for fixing the spin, there are 46 sources that are not part of *Gaia*-CRF3. Two of them were used to fix the spin. For those 46 or 2 sources, the respective queries return empty names of the source and the external catalogue.

As described in Sect. 2, the *Gaia*-CRF3 sources resulted from a filtered cross-match between the *Gaia* EDR3 catalogue and a number of external AGN/QSO catalogues. The information on which sources were found in the different external catalogues is given in table `gaiadr3.gaia_crf3_xm`. We note that `gaiaedr3.agn_cross_id` cannot be used to retrieve all *Gaia*-CRF3 sources from a given external catalogue, because a *Gaia*-CRF3 source may be found in several external catalogues but `gaiaedr3.agn_cross_id` only gives its identifier in one of them. To obtain a list of the *Gaia*-CRF3 sources that were selected from a specific external catalogue one can use the query:[6]

```
SELECT xm.$$$_name, edr3.*
FROM gaiaedr3.gaia_source AS edr3
INNER JOIN gaiadr3.gaia_crf3_xm AS xm
USING (source_id)
WHERE xm.$$$ = 'true'
```

where '$$$' should be replaced by the catalogue code from column 6 of Table 1. For the sources in the three ICRF3 catalogues (the catalogue codes `icrf3sx`, `icrf3k`, `icrf3xka`) `gaiadr3.gaia_crf3_xm` provides two source designations as also given in the original ICRF3 catalogues: the ICRF and IERS designations as `icrf_name` and `iers_name`, respectively. Additionally, the above ADQL query with the catalogue code 'b19'

[5] https://gea.esac.esa.int/archive/

[6] Table `gaiadr3.gaia_crf3_xm` will be published in the *Gaia* Archive together with *Gaia* DR3. Alternatively, the table can be downloaded from https://www.cosmos.esa.int/web/gaia/gaia-crf3-cross-match-table and uploaded to the *Gaia* Archive as a user table. In that case the table name in the ADQL query should be modified accordingly.

can be used to see the 920 599 *Gaia*-CRF3 sources that were also found in the catalogue of Bailer-Jones et al. (2019).

We note that when considering an external catalogue '`$$$`' some sources were rejected only because a lower quality of astrometry used in that catalogue resulted in the cross-match distance Δ being too large to satisfy Eq. (5) in Sect. 2. For those sources the field `gaiadr3.gaia_crf3_xm.$$$` is set to false. In some cases, however, the same *Gaia* source was found also in other external catalogues and eventually selected for *Gaia*-CRF3. In those cases, the name `gaiadr3.gaia_crf3_xm.$$$_name` of that *Gaia* source in catalogue '`$$$`' is also given. To obtain all *Gaia*-CRF3 sources that appear in an external catalogue one should replace the last line in the previous ADQL query by '`WHERE xm.$$$_name IS NOT NULL`'. This helps to mitigate the effect of a lower quality astrometry in some of the external catalogues and improve the overall cross-match quality for those catalogues.

Appendix D: Confusion sources in *Gaia* EDR3

In this appendix we give the number and distributions of sources in *Gaia* EDR3 that satisfy Eqs. (1)–(3), ensuring that the *Gaia* astrometry is statistically compatible with the hypothesis that a source is extragalactic with zero parallax and proper motion. (As discussed in Sect. 2, we ignore the small effects of the Galactic acceleration and possible apparent proper motions induced by variable source structure, but take into account an overall parallax offset of −0.017 mas.) In this paper we call these the 'confusion sources', because most of them are in fact distant stars in our Galaxy or its satellite galaxies, and stellar contaminants resulting from erroneous cross-matching with external QSO catalogue (or from stellar contamination in these catalogues) can be expected to have similar characteristics as the confusion sources.

We give separate statistics for confusion sources with five- and six-parameter solutions in *Gaia* EDR3. For the five-parameter solutions, an explicit query of the *Gaia* EDR3 catalogue[7]

```
SELECT * FROM gaiaedr3.gaia_source AS g
WHERE g.astrometric_params_solved = 31
  AND abs((g.parallax + 0.017)/g.parallax_error) < 5
  AND ( power(g.pmra/g.pmra_error,2)
       +power(g.pmdec/g.pmdec_error,2)
       -2*g.pmra_pmdec_corr*
         g.pmra/g.pmra_error*
         g.pmdec/g.pmdec_error)/
      (1-power(g.pmra_pmdec_corr,2)) < 25
```

results in 30 723 995 sources satisfying Eqs. (1)–(3). Their sky density is shown in the top panel of Fig. D.1. Of those, 7 090 844 sources are in the Galactic zone with $|\sin b| < 0.1$, thus violating the fourth criterion in Eq. (4). The remaining 23 633 151 confusion sources with five-parameter astrometric solutions in *Gaia* EDR3 satisfy all four criteria and could, based on their astrometry alone, be potential candidates for the *Gaia*-CRF.

A corresponding query of *Gaia* EDR3 for confusion sources with six-parameter solutions (obtained by changing the second line of the query to '`WHERE g.astrometric_params_solved = 95`') returns 182 814 959 sources. Their sky density is shown in the lower panel of Fig. D.1). 90 979 403 of the sources have $|\sin b| > 0.1$ and thus satisfy all four criteria in Eqs. (1)–(4).

Among the many features seen in Fig. D.1 we recognise the expected density distribution due to the overall distribution of Galactic stars and a strong over-density in the areas of the LMC and SMC. There are several more compact high-density areas outside the Galactic plane. A detailed (but not exhaustive) inspection of the density map revealed about 20 stellar (mostly globular) clusters and 10 (dwarf) galaxies on the maps. Not counting the LMC and SMC areas, the total number of sources in these clusters and galaxies is however only about 5000, which is ignored in the statistical considerations below. We note that the Sagittarius stream is clearly visible (mainly in the top panel), as well as numerous streaks of under-density related to the scanning law.

A prominent feature in Fig. D.1 is the relatively low density of sources with five-parameter solutions along the Galactic equator and in the centres of the Magellanic Clouds. This is caused by the difficulty to obtain reliable colour information from the BP and RP spectra on faint sources in these very crowded areas (Riello et al. 2021). As a result, the astrometric solution did not use a spectrophotometrically determined colour (effective wavenumber) for many of these sources, but instead determined a pseudo-colour as the sixth astrometric parameter.

Because the confusion sources in the LMC and SMC tend to have distinctly different kinematics from the Galactic stars, it is advantageous to separate out the sources in the LMC and SMC areas in the further analysis. Guided by the density maps, we take a circle of radius 9° centred on $(\alpha, \delta) = (81.3°, -68.7°)$ to represent the LMC, and a circle of radius 6° centred on $(\alpha, \delta) = (16.0°, -72.8°)$ to represent the SMC. These areas contain, respectively, 2 657 059 and 496 520 confusion sources with five-parameter solutions (6 817 915 and 1 307 591 with six-parameter solutions). This simplistic selection does not pretend to any deeper physical significance, and only partially coincides with the carefully selected samples of *Gaia* EDR3 sources in the LMC and SMC discussed in Gaia Collaboration (2021c).

Figure D.2 shows the distribution of the confusion sources in galactic coordinates and magnitude, split by the kind of solution (five or six parameters) and further divided into four disjoint groups: sources in the LMC and SMC areas defined above, and for the remaining sources according to $|\sin b| \gtrless 0.1$. The confusion sources outside the LMC and SMC areas are called 'Galactic' below, although it is clear that this set contains also many extragalactic objects.

Figures D.3 and D.4 show the distributions of the normalised parallaxes and proper motions of the confusion sources separately for the Galactic sources (split by galactic latitude) and the LMC and SMC areas. The distributions are strongly non-Gaussian (with the possible exception of parallaxes in the LMC/SMC areas) and very different from those of the *Gaia*-CRF3 sources in Fig. 3. In Sect. 2, this difference was used to check the purity of the filtered sources in the external QSO catalogues and in the final *Gaia*-CRF3. The proper motions give the strongest indication of impurity. Interestingly, for the Galactic confusion sources there is no clear dependence according to $|\sin b| \gtrless 0.1$. On the other hand, the confusion sources in the LMC and SMC areas give rather different and distinct distributions. The distributions for the five- and six-parameter solutions are not drastically different and mainly reflect the generally higher uncertainties of the six-parameter sources, which makes their distributions slightly less non-Gaussian.

Selection of QSO-like objects in the *Gaia* catalogue, based exclusively on parallaxes and proper motions, is an interesting prospect already discussed in the literature (Heintz et al. 2015; 2018). Among the *Gaia* EDR3 sources with five-parameter solutions, there are 23.6 million satisfying Eqs. (1)–(4), of which 1.2 million (5.1%) are also in *Gaia*-CRF3. Allowing that the actual number of QSOs among the five-parameter sources in *Gaia* EDR3 could be a factor two higher (cf. Sect. 3), we conclude that the purity of such a selection from *Gaia* EDR3 would be at most about 10% – too low to be of practical use, but still interesting and promising. In certain areas on the sky the percentage of true QSO-like sources in the sample is much higher. For example, excluding $|b| < 20°$ and the LMC/SMC areas, there are 5 767 163 confusion sources, of which 1 410 314 (24.5%) are *Gaia*-CRF3 sources. The purity of such selections in future *Gaia* data releases will improve drastically thanks to the higher accuracy of parallaxes and (especially) proper motions (cf. Sect. 2.5).

[7] The queries discussed in this Appendix are costly and result in very big tables. To cope with the technical limitations, one can either split the results in chunks by using some conditions on '`random_index`' (the random index assigned for each source in the *Gaia* Archive) or execute the equivalent queries offline using a local copy of the whole *Gaia* EDR3 catalogue.

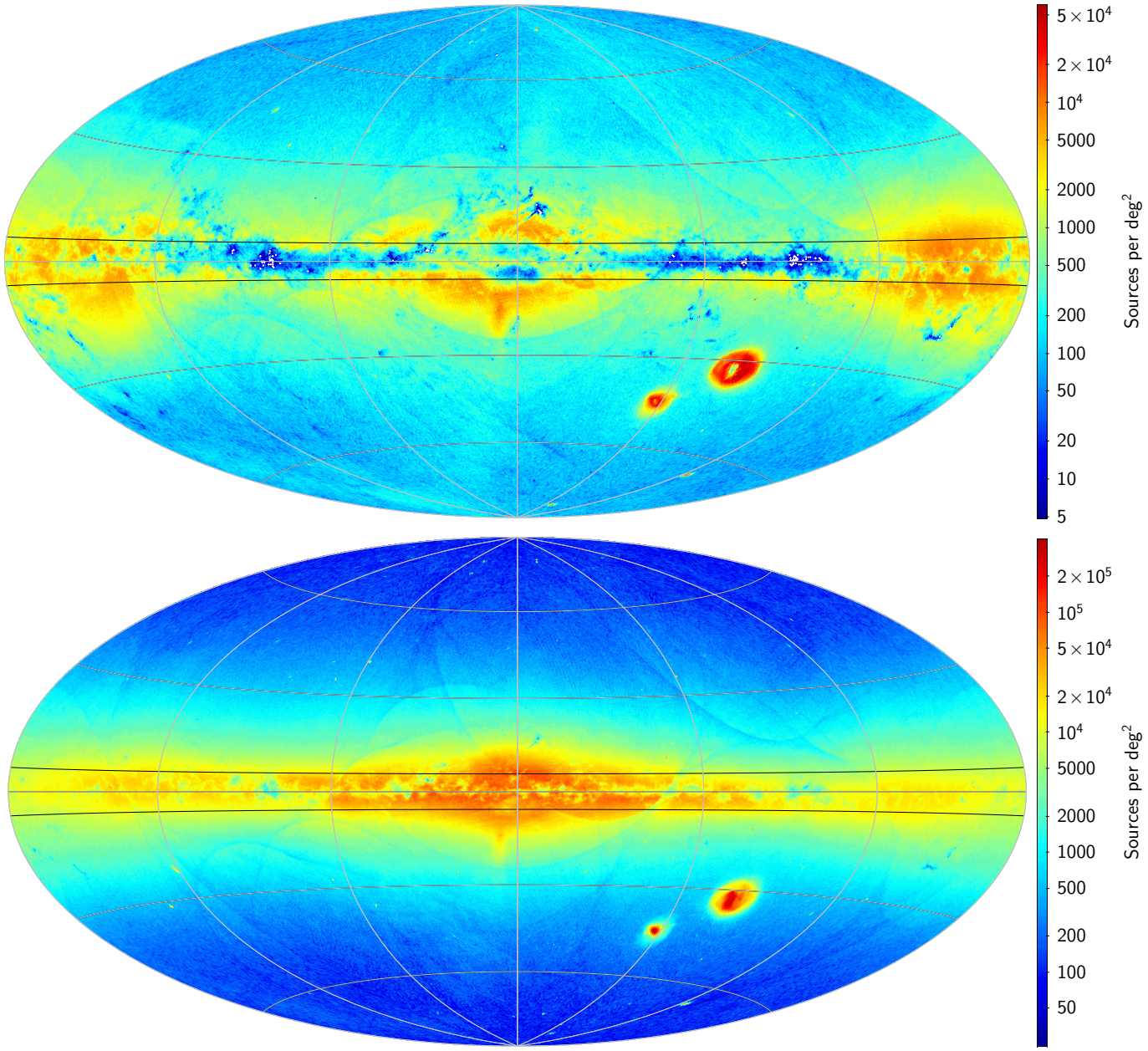

Fig. D.1. Distribution of the confusion sources in *Gaia* EDR3. These maps use a Hammer–Aitoff projection in galactic coordinates, with $l = b = 0$ at the centre, north up, and l increasing from right to left. The black lines bound the Galactic zone $|\sin b| < 0.1$ where sources are normally excluded by Eq. (4). *Top:* sources with five-parameter solutions. *Bottom:* sources with six-parameter solutions.

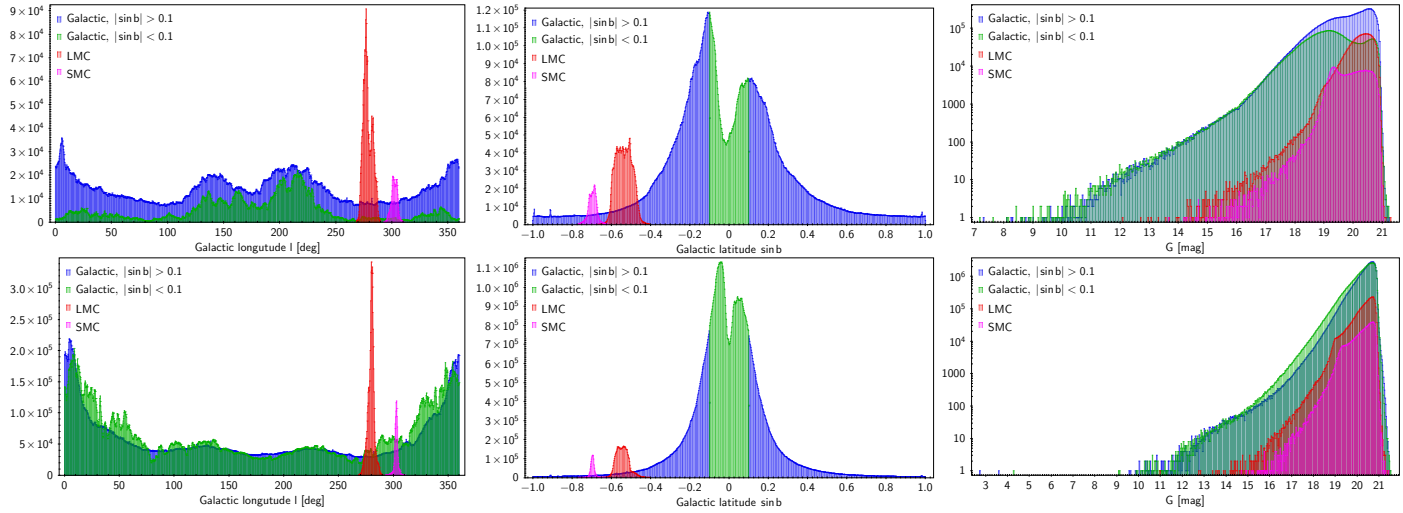

Fig. D.2. Distribution of confusion sources in (from *left to right*) galactic longitude, sine of galactic latitude, and magnitude. *Top:* sources with five-parameter solutions. *Bottom:* sources with six-parameter solutions. The selections indicated in the legend are further explained in the text.

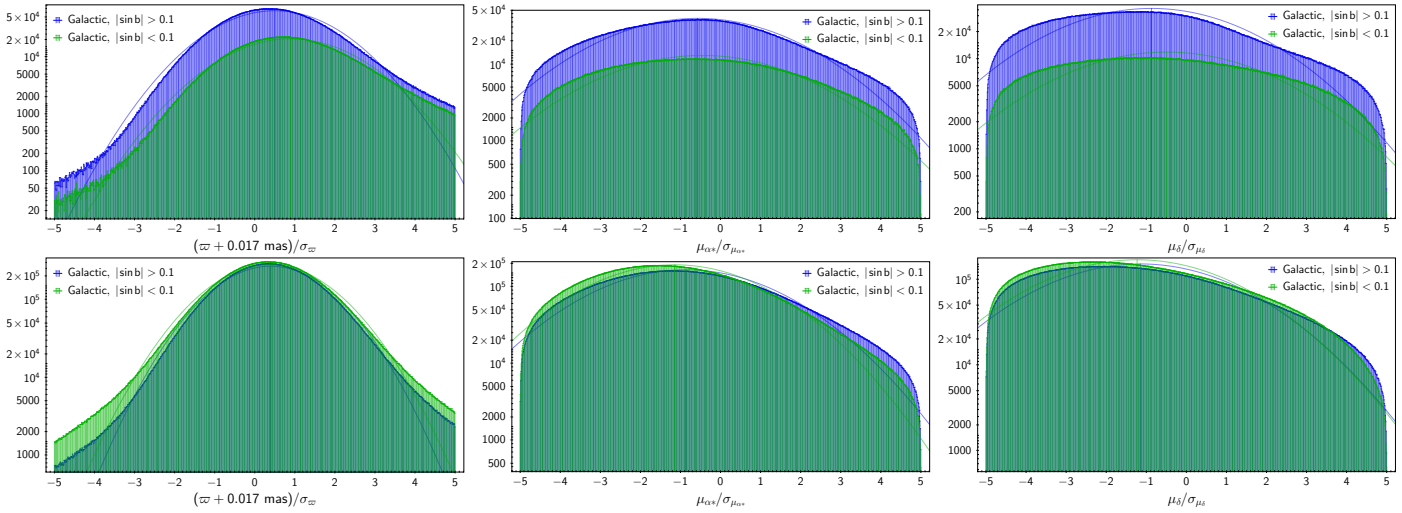

Fig. D.3. Histograms of the normalised parallaxes and proper motion components for the Galactic confusion sources. *Top:* sources with five-parameter solutions. *Bottom:* sources with six-parameter solutions. The different colours show the selections indicated in the legend.

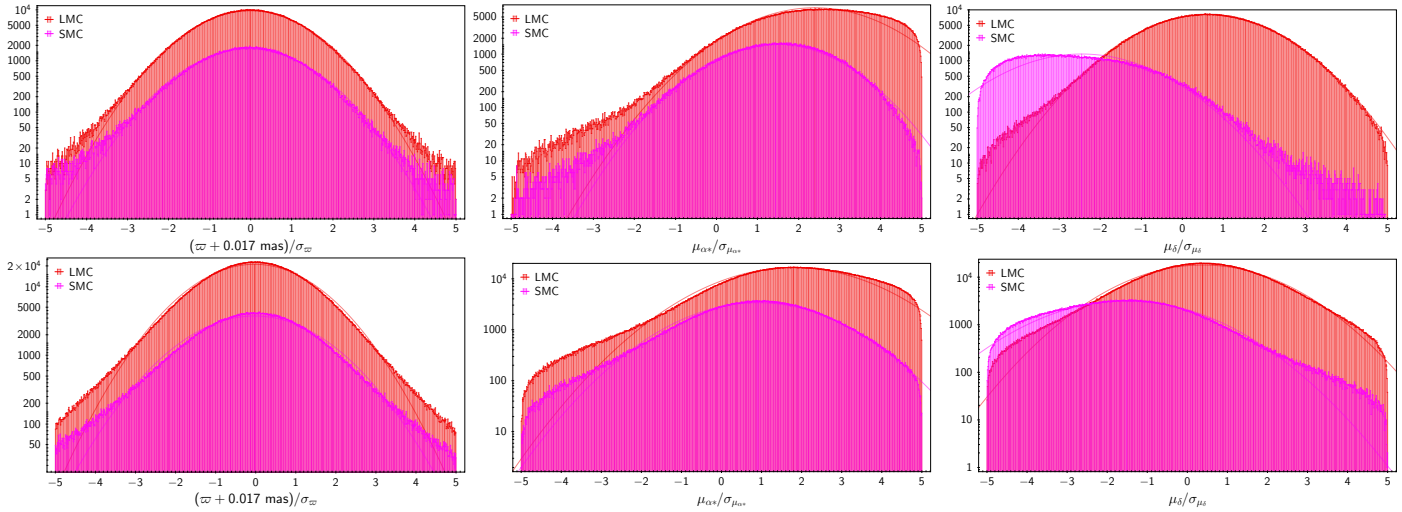

Fig. D.4. Histograms of the normalised parallaxes and proper motion components for the confusion sources in the LMC and SMC areas. *Top:* sources with five-parameter solutions. *Bottom:* sources with six-parameter solutions.

Appendix E: Estimating the spin and orientation corrections

As described in Sect. 6.1 of Lindegren et al. (2012), the frame rotator is an integral part of AGIS with three main functions: (i) to estimate the spin and orientation corrections needed to bring the current astrometric solution onto the ICRS; (ii) to apply the estimated corrections to the astrometric parameters of the primary sources; and (iii) to apply the corresponding corrections to the attitude parameters. This appendix describes the algorithm used for the first function in the construction of *Gaia* EDR3.

The 2012 paper gave a stringent mathematical framework for all three functions and outlined how they could be implemented in AGIS. The frame rotator used for the first two releases of *Gaia* data followed these prescriptions quite rigorously, which however resulted in code that was unnecessarily complex and not very transparent. For example, it permitted the use of different reference epochs for the AGIS solution, comparison catalogue, and orientation parameters, while in practice this was never needed. It also foresaw the use of radio stars observed by VLBI to help align the reference frame, which is no longer considered necessary or even desirable. On the other hand, the details of the numerical algorithm used for the robust least-squares estimation (LSE) of the rotation parameters were not at all described in Lindegren et al. (2012), making it very difficult for outside users to reproduce how the frame rotator worked in those releases.

In 2018 it was decided to implement a new version of the frame rotator to be used for the third and subsequent data releases. The main changes with respect to the earlier version are as follows: only two kinds of sources are used for estimating the rotation parameters, namely QSO-like sources with high-accuracy external positions (the VLBI positions of ICRF3 S/X in *Gaia* EDR3), used for both orientation and spin, and QSO-like sources without accurate external positions, used only for the spin. The same reference epoch is used for rotation parameters as for the astrometric data (2016.0 for (E)DR3), which permits to separate completely the solution for the orientation from that of the spin. Furthermore, the observation equations are written directly in terms of the proper motions and position differences (Sect. E.1), rather than using the rigorous but less transparent formalism of Lindegren et al. (2012). The solution takes into account the correlation between the components in right ascension and declination of both *Gaia* and external data (Sect. E.2), and has a transparent scheme for the detection and treatment of outliers (Sect. E.3). Finally, the solution provides reasonable estimates of the statistical uncertainties of the rotation parameters also when the uncertainties of the input data are underestimated (Sect. E.4), and may optionally include additional terms, such as a glide in the proper motions or more general expansions of the proper motions and position differences (Sect. E.5).

All these changes concern the estimation algorithm, that is the first function of the frame rotator. The remaining two functions, that is the application of the corrections to the source and attitude data, are still done as described in Sects. 6.1.2 and 6.1.3 of the 2012 paper and are not further discussed here.

We denote by Q_0 the set of QSO-like sources considered for the determination of the spin correction, of which the subset Q was actually used after elimination of outliers; similarly, $\mathcal{R}_0$ and $\mathcal{R}$ are the set of sources considered and used for the determination of the orientation. Although not required by the algorithm, we normally have $\mathcal{R}_0 \subseteq Q_0$. The number of sources in each set is denoted $n(Q_0)$, etc. We use $\sigma(x)$, $\rho(x, y)$, and $\mathrm{Cov}(x, y)$ for the uncertainty, correlation coefficient, and covariance of the arbitrary parameters x and y. The asterisk in α^* indicates that the $\cos\delta$ factor is implicit, as in $\mu_{\alpha^*} = \dot{\alpha}\cos\delta$. To distinguish between data in the current astrometric solution and the corresponding data in the external catalogue, we put an overline ($\bar{}$) on the latter. The prime $'$ denotes the matrix transpose.

Appendix E.1. Observation equations

Let $\boldsymbol{\omega} = [\omega_X,\ \omega_Y,\ \omega_Z]'$ be the spin correction of the current solution, expressed by its components along the principal axes of the ICRS. (As indicated by the prime, $\boldsymbol{\omega}$ is a column matrix.) In matrix form, the observation equations in proper motion for source i are

$$\boldsymbol{A}_i\boldsymbol{\omega} = \boldsymbol{\mu}_i\,, \tag{E.1}$$

with coefficient matrix

$$\boldsymbol{A} = \begin{bmatrix} -\cos\alpha_i\sin\delta_i & +\sin\alpha_i\sin\delta_i & +\cos\delta_i \\ +\sin\alpha_i & -\cos\alpha_i & 0 \end{bmatrix} \tag{E.2}$$

and right-hand side

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{\alpha^*i} \\ \mu_{\delta i} \end{bmatrix}. \tag{E.3}$$

Here, α_i, δ_i, μ_{α^*i} and $\mu_{\delta i}$ are the position and proper motion components of the source in the current solution, with reference epoch t_{ref} (= 2016.0 for *Gaia* EDR3). For the weight normalisation and decorrelation (Sect. E.2), we also need the covariance of the right-hand side in Eq. (E.3),

$$\mathrm{Cov}(\boldsymbol{\mu}_i) = \begin{bmatrix} \sigma(\mu_{\alpha^*i})^2 & \rho(\mu_{\alpha^*i},\mu_{\delta i})\,\sigma(\mu_{\alpha^*i})\,\sigma(\mu_{\delta i}) \\ \rho(\mu_{\alpha^*i},\mu_{\delta i})\,\sigma(\mu_{\alpha^*i})\,\sigma(\mu_{\delta i}) & \sigma(\mu_{\delta i})^2 \end{bmatrix}. \tag{E.4}$$

Similarly, let $\boldsymbol{\varepsilon} = [\varepsilon_X,\ \varepsilon_Y,\ \varepsilon_Z]'$ be the orientation correction at the reference epoch t_{ref}, expressed by its components in the ICRS. The linearised observation equations in position for source i are

$$\boldsymbol{A}_i\boldsymbol{\varepsilon} = \boldsymbol{\Delta}_i\,, \tag{E.5}$$

with the same coefficient matrix $\boldsymbol{A}_i$ as before. The right-hand side,

$$\boldsymbol{\Delta}_i \equiv \begin{bmatrix} \Delta\alpha^*_i \\ \Delta\delta_i \end{bmatrix} = \begin{bmatrix} (\alpha_i - \bar{\alpha}_i)\cos\delta_i \\ \delta_i - \bar{\delta}_i \end{bmatrix}, \tag{E.6}$$

is the difference between the position $(\alpha_i,\ \delta_i)$ in the current solution and the position $(\bar{\alpha}_i,\ \bar{\delta}_i)$ according to the external catalogue. We note that the external positions at t_{ref} (= 2016.0 for EDR3) should be used here. In particular, for *Gaia* EDR3 the galactic acceleration should be taken into account with the ICRF3 S/X positions as prescribed in Sect. 5.3 of Charlot et al. (2020). The covariance of $\boldsymbol{\Delta}_i$ in Eq. (E.6) is computed on the assumption that the positional errors of the current solution and of the external catalogue are independent:

$$\mathrm{Cov}(\boldsymbol{\Delta}_i) = \mathrm{Cov}(\alpha^*_i,\ \delta_i) + \mathrm{Cov}(\bar{\alpha}^*_i,\ \bar{\delta}_i)\,, \tag{E.7}$$

where

$$\mathrm{Cov}(\alpha^*_i,\ \delta_i) = \begin{bmatrix} \sigma(\alpha^*_i)^2 & \rho(\alpha^*_i,\delta_i)\,\sigma(\alpha^*_i)\,\sigma(\delta_i) \\ \rho(\alpha^*_i,\delta_i)\,\sigma(\alpha^*_i)\,\sigma(\delta_i) & \sigma(\delta_i)^2 \end{bmatrix}, \tag{E.8}$$

and correspondingly for the external data. We note that

$$\sigma(\Delta\alpha_i^*) = \sqrt{\sigma(\alpha_i^*)^2 + \sigma(\bar{\alpha}_i^*)^2}\,, \tag{E.9}$$

$$\sigma(\Delta\delta_i) = \sqrt{\sigma(\delta_i)^2 + \sigma(\bar{\delta}_i)^2}\,, \tag{E.10}$$

$$\rho(\Delta\alpha_i^*, \Delta\delta_i) = \frac{\rho(\alpha_i^*, \delta_i)\,\sigma(\alpha_i^*)\,\sigma(\delta_i) + \rho(\bar{\alpha}_i^*, \bar{\delta}_i)\,\sigma(\bar{\alpha}_i^*)\,\sigma(\bar{\delta}_i)}{\sigma(\Delta\alpha_i^*)\,\sigma(\Delta\delta_i)}\,. \tag{E.11}$$

The total number of observation equations considered for $\boldsymbol{\omega}$ is $2n(\mathcal{Q}_0)$, and for $\boldsymbol{\varepsilon}$ it is $2n(\mathcal{R}_0)$. A feature of the algorithm described here is that the observation equations always come in pairs, as in Eqs. (E.1) and (E.5), and that the two components in a pair are always treated together. This simplifies the decorrelation of data (see below), and affects the way outliers are detected and removed. The resulting solution and statistics are independent of the coordinate system in which they are calculated.

Appendix E.2. Weight normalisation and decorrelation

Because the observation equations have different uncertainties (heteroscedasticity) and are pairwise correlated, they should be solved by the generalised least-squares method for optimality. Equivalently, the ordinary (unweighted) least-squares method can be applied to the normalised equations,

$$\boldsymbol{K}_i\boldsymbol{A}_i\boldsymbol{\omega} = \boldsymbol{K}_i\boldsymbol{\mu}_i\,, \tag{E.12}$$

$$\boldsymbol{L}_i\boldsymbol{A}_i\boldsymbol{\varepsilon} = \boldsymbol{L}_i\boldsymbol{\Delta}_i\,, \tag{E.13}$$

where $\boldsymbol{K}_i$ and $\boldsymbol{L}_i$ are 2×2 matrices chosen in such a way that the transformed right-hand sides are uncorrelated and of unit variance. This condition does not define $\boldsymbol{K}_i$ and $\boldsymbol{L}_i$ uniquely, since any orthogonal transformation of them will give the same result; the form used here is the inverse lower Cholesky factor of the covariance matrix, or

$$\boldsymbol{K}_i = \begin{bmatrix} \frac{1}{\sigma(\mu_{\alpha^*i})} & 0 \\ \frac{-\rho(\mu_{\alpha^*i},\,\mu_{\delta i})}{\sigma(\mu_{\alpha^*i})\sqrt{1-\rho(\mu_{\alpha^*i},\,\mu_{\delta i})^2}} & \frac{1}{\sigma(\mu_{\delta i})\sqrt{1-\rho(\mu_{\alpha^*i},\,\mu_{\delta i})^2}} \end{bmatrix} \tag{E.14}$$

and

$$\boldsymbol{L}_i = \begin{bmatrix} \frac{1}{\sigma(\Delta\alpha_i^*)} & 0 \\ \frac{-\rho(\Delta\alpha_i^*,\,\Delta\delta_i)}{\sigma(\Delta\alpha_i^*)\sqrt{1-\rho(\Delta\alpha_i^*,\,\Delta\delta_i)^2}} & \frac{1}{\sigma(\Delta\delta_i)\sqrt{1-\rho(\Delta\alpha_i^*,\,\Delta\delta_i)^2}} \end{bmatrix}. \tag{E.15}$$

It is readily verified that, as required, $\boldsymbol{K}_i\mathrm{Cov}(\boldsymbol{\mu}_i)\boldsymbol{K}_i' = \boldsymbol{I}$ and $\boldsymbol{L}_i\mathrm{Cov}(\boldsymbol{\Delta}_i)\boldsymbol{L}_i' = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Given the subsets $\mathcal{Q}$, $\mathcal{R}$ of the accepted sources, the generalised least-squares solutions are obtained by minimising the sum of squared normalised residuals,

$$X^2(\boldsymbol{\omega}, \mathcal{Q}) = \sum_{i\in\mathcal{Q}} \left|\boldsymbol{K}_i\boldsymbol{\mu}_i - \boldsymbol{K}_i\boldsymbol{A}_i\boldsymbol{\omega}\right|^2, \tag{E.16}$$

$$X^2(\boldsymbol{\varepsilon}, \mathcal{R}) = \sum_{i\in\mathcal{R}} \left|\boldsymbol{L}_i\boldsymbol{\Delta}_i - \boldsymbol{L}_i\boldsymbol{A}_i\boldsymbol{\varepsilon}\right|^2. \tag{E.17}$$

This can be done by a range of standard numerical techniques, including orthogonal transformations and the use of normal equations (e.g. Lawson & Hanson 1995, Björck 1996). For the frame rotator, the normalised observation equations in Eqs. (E.12) and (E.13) are solved by QR decomposition, using Householder transformations.

Appendix E.3. Residual statistics and treatment of outliers

Standard least-squares fitting is sensitive to outliers, that is data points that do not reasonably fit the observation equations, considering the uncertainties and the typical scatter of residuals. Outliers must be expected among the external counterparts, as well as among the other QSO-like sources, for a variety of reasons. It is therefore essential that the fit is robust against outliers. The method adopted for the frame rotator is to identify the outliers that must be removed from the originally considered samples ($\mathcal{Q}_0$, $\mathcal{R}_0$) to create the subsets $\mathcal{R}$ and $\mathcal{Q}$ used in Eqs. (E.17) and (E.16). Two things should be noted concerning this process. Firstly, outliers are identified at source level, that is, a given source is either rejected in both coordinates (α and δ, or μ_{α^*} and μ_δ), or accepted in both coordinates. This makes sense from a physical viewpoint, because a 'bad' source is probably bad (at some level) in both coordinates. Secondly, because $\boldsymbol{\omega}$ and $\boldsymbol{\varepsilon}$ are treated separately, it happens that a source rejected in $\mathcal{R}$ (for the orientation) is accepted in $\mathcal{Q}$ (for the spin), or vice versa. Again, this could be motivated on physical grounds.

Details of the subsequent algorithm are exactly the same for Eqs. (E.16) and (E.17), and are therefore described in terms of the generic function

$$X^2(\boldsymbol{x}, \mathcal{S}) = \sum_{i\in\mathcal{S}} X_i^2(\boldsymbol{x})\,, \tag{E.18}$$

where

$$X_i(\boldsymbol{x}) = \left|\,\boldsymbol{d}_i - \boldsymbol{D}_i\boldsymbol{x}\,\right|. \tag{E.19}$$

Here, $\boldsymbol{x}$ could be either $\boldsymbol{\omega}$ or $\boldsymbol{\varepsilon}$, with $\mathcal{S}$, $\boldsymbol{d}$, and $\boldsymbol{D}$ correspondingly defined according to Eq. (E.16) or (E.17).

We note that the functions $X_i(\boldsymbol{x})$, which can be computed for all the sources in $\mathcal{S}_0$, are a natural measure of the discrepancy of source i from the model for given parameter vector $\boldsymbol{x}$. A useful overall statistic is the median discrepancy

$$X_{0.5}(\boldsymbol{x}) = \operatorname*{med}_{i\in\mathcal{S}_0}\left(X_i(\boldsymbol{x})\right), \tag{E.20}$$

and a robust estimator could be to simply minimise this quantity. However, if the errors are approximately Gaussian, with a moderate fraction of outliers, robust least-squares estimation is more accurate and therefore preferred for the frame rotator. If $\boldsymbol{x}$ is the true parameter vector, and all the errors are Gaussian with nominal covariances, $X_{0.5}(\boldsymbol{x})$ is expected to be around $\sqrt{\ln 4} \simeq 1.1774$ (which is the median of the chi-distribution with two degrees of freedom).

The adopted solution $\hat{\boldsymbol{x}}$ and accepted subset $\mathcal{S}$ simultaneously satisfy the three conditions

$$\left.\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmin}_{\boldsymbol{x}} \sum_{i\in\mathcal{S}} X_i^2(\boldsymbol{x})\,, \\ \forall i \in \mathcal{S} &: X_i(\hat{\boldsymbol{x}}) \le \kappa\, X_{0.5}(\hat{\boldsymbol{x}})\,, \\ \forall i \in \mathcal{S}_0\backslash\mathcal{S} &: X_i(\hat{\boldsymbol{x}}) > \kappa\, X_{0.5}(\hat{\boldsymbol{x}}) \end{aligned}\right\} \tag{E.21}$$

for the chosen clip limit $\kappa \ge 1$ (see below). It is not obvious that these conditions lead to a unique solution; indeed, it is easy to construct examples where this is not the case, for example if $\mathcal{S}_0$ consists of disjoint subsets with distinctly different rotation parameters. However, these are somewhat contrived examples and in practice the non-uniqueness is hardly an issue. More relevant is the question whether a solution always exists – to which we do not know the answer. We take the pragmatic view that

a unique solution 'close enough' to satisfying Eq. (E.21) can always be found by the following algorithm.

Start by provisionally accepting all the sources in $\mathcal{S}_0$, compute the least-squares estimate $\hat{\boldsymbol{x}}_0$ by minimising $X^2(\boldsymbol{x}, \mathcal{S}_0)$, and hence obtain the statistic $X_{0.5}(\hat{\boldsymbol{x}}_0)$. In the next iteration, let $\mathcal{S}_1$ be the subset of sources satisfying $X_i(\hat{\boldsymbol{x}}_0) \leq \kappa X_{0.5}(\hat{\boldsymbol{x}}_0)$, leading to the estimate $\hat{\boldsymbol{x}}_1$ minimising $X^2(\boldsymbol{x}, \mathcal{S}_1)$ and the new statistic $X_{0.5}(\hat{\boldsymbol{x}}_1)$, from which the subset $\mathcal{S}_2$ is constructed, and so on. This gives a sequence of subsets $\mathcal{S}_k$, $k = 0, 1, 2, \ldots$ that may converge to a stable subset (i.e. $\mathcal{S}_{k+1} = \mathcal{S}_k$ for some k). If so, the stable subset and the corresponding estimate $\hat{\boldsymbol{x}}_k$ is the desired solution. But even if the sequence does not converge, it will sooner or later be found that the same subsets reappear cyclically in the sequence, that is $\mathcal{S}_{k+p} = \mathcal{S}_k$, where $p > 1$ is the cycle period. In such a case any of the solutions in the cycle might be accepted. However, by adopting the rule that the sequence is stopped at the lowest k where $\mathcal{S}_k = \mathcal{S}_{k-p}$ for some $p > 0$, we arrive at a unique result both for a converging sequence ($p = 1$) and in the cyclic case ($p > 1$).

To avoid having to save and compare the (potentially many and large) subsets $\mathcal{S}_k$, we assign a fixed random number r_i to each source in $\mathcal{S}_0$, and save from previous iterations only the outlier 'checksum' $o_k = \sum_{i \in \mathcal{S}_0 \setminus \mathcal{S}_k} r_i$. The iteration stops as soon as $o_k = o_{k-p}$ for some $p > 0$, at which point we adopt $\mathcal{S} = \mathcal{S}_k$ and $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_k$ as the solution.

The clip limit κ is a dimensionless number, typically in the range 2–5, to be chosen by the researcher. A value $\kappa \geq 1$ ensures that at most 50% of the sources in $\mathcal{S}_0$ are rejected as outliers, but in practice a higher value should always be used. In principle, an optimum clip level that minimises the expected uncertainty of $\hat{\boldsymbol{x}}$ can be estimated by careful modelling of the actual error distribution in $\mathcal{S}_0$, but for the frame rotator we always use a fixed $\kappa = 3$. In the nominal case of purely Gaussian errors and no actually bad data, this will (erroneously) reject about 1% of the sources. This appears to be an acceptable price to pay for an estimation algorithm that is both efficient and very robust.

Appendix E.4. Uncertainty of the estimates

The covariance of $\hat{\boldsymbol{x}}$ is estimated as

$$\mathrm{Cov}\,(\hat{\boldsymbol{x}}) = \left[\sum_{i \in \mathcal{S}} \boldsymbol{D}_i' \boldsymbol{D}_i\right]^{-1} \times f\,, \tag{E.22}$$

where the first factor (the inverse of the weighted normal matrix) is the nominal covariance obtained on the assumption that the data uncertainties in $\mathrm{Cov}(\boldsymbol{\Delta}_i)$ and $\mathrm{Cov}(\boldsymbol{\mu}_i)$ are correctly estimated, and that no outlier rejection is applied. An empirical correction factor $f > 0$ takes into account that the data uncertainties may be systematically wrong (usually underestimated), and that the nominal covariance estimate is biased by the outlier rejection process even if the data uncertainties are correctly estimated.

In practice, the correction factor f must be derived from the residuals of the least-squares solution, or in our case from the statistics of $X_i(\hat{\boldsymbol{x}})$. An obvious possibility is to use the reduced chi-square (the square of the unit weight error u),

$$u^2 = \frac{X^2(\hat{\boldsymbol{x}}, \mathcal{S})}{2n(\mathcal{S}) - m}\,, \tag{E.23}$$

as an estimate of f, where $n(\mathcal{S})$ is the number of accepted sources and m is the dimension of $\boldsymbol{x}$ (here, $m = 3$). This is accurate if $n(\mathcal{S})$ is large (as it usually is in our applications), the fraction of outliers is very small, and the data uncertainties are biased by an approximately constant factor. However, when a significant fraction of the sources are rejected as outliers, the use of $f = u^2$ tends to severely underestimate the uncertainties of $\hat{\boldsymbol{x}}$. A more reasonable correction factor should therefore consider the error distribution in the full sample $\mathcal{S}_0$, not just in the accepted subset $\mathcal{S}$. Discrepancy quantiles such as $X_{0.5}$ in Eq. (E.20) provide robust characterisation of the full error distribution.

Based on numerical simulations, we have adopted the following recipe for the correction factor to be used in Eq. (E.22):

$$f = \max\left[u^2,\ X_{0.5}^2/(\ln 4)\right]\,. \tag{E.24}$$

In the ideal case of Gaussian errors, correctly estimated data uncertainties, and a negligible fraction of outliers, we expect $u^2 \simeq 1$ and $X_{0.5}^2 \simeq \ln 4$, and hence $f \simeq 1$. If the data uncertainties are biased by a certain factor, both u and $X_{0.5}$ scale by the same factor, and the recipe still works. For samples yielding a non-negligible fraction of outliers we normally have $u^2 < X_{0.5}^2/(\ln 4)$, and the higher factor then tends to give a more accurate estimate of the uncertainties, hence Eq. (E.24).

Appendix E.5. Optional parameters

The estimation as described above has only three unknowns, that is the components of $\boldsymbol{\omega}$ or $\boldsymbol{\varepsilon}$. The algorithm, including weight normalisation, decorrelation, and outlier treatment, is readily adapted to more general linear models, for example using VSH to represent higher-order distortions of the proper motions and position differences (Mignard & Klioner 2012), or including terms depending on the colour and magnitude of the source. However, while such more complex models are useful for investigating the properties of the final catalogue, they should normally not be used in the frame rotator.

One possible exception is the glide in proper motion, which is expected to be present on physical grounds. Already in *Gaia* EDR3 this effect is very clear, and found to be consistent with the expected Galactic acceleration of the solar system barycentre (Gaia Collaboration 2021b). The proper motion model implicit in Eq. (E.1) neglects this effect, which was deemed acceptable for the frame rotator in the first three releases of *Gaia* astrometry. For future releases this may not be accurate enough (since the effects of rotation and glide can be correlated and mutually disturb each other). One possible remedy is then to correct the observed proper motions in Eq. (E.3) for the assumed Galactic glide. However, a better option may be to estimate the glide $\boldsymbol{g}$ together with $\boldsymbol{\omega}$, using the augmented observation equations

$$\boldsymbol{A}_i\boldsymbol{\omega} + \boldsymbol{B}_i\boldsymbol{g} = \boldsymbol{\mu}_i\,. \tag{E.25}$$

The estimation follows the generic scheme outlined above, minimising $X^2 = \sum_{i \in Q} |\,\boldsymbol{d}_i - \boldsymbol{D}_i\boldsymbol{x}\,|^2$ with

$$\boldsymbol{x} = \begin{bmatrix}\boldsymbol{\omega}\\ \boldsymbol{g}\end{bmatrix}, \quad \boldsymbol{d}_i = \boldsymbol{L}_i\boldsymbol{\mu}_i\,, \quad \text{and} \quad \boldsymbol{D}_i = \boldsymbol{L}_i\begin{bmatrix}\boldsymbol{A}_i & \boldsymbol{B}_i\end{bmatrix}. \tag{E.26}$$

It should be noted that the application of the spin and orientation corrections to the source and attitude data, that is functions (ii) and (iii) of the frame rotator, is in any case restricted to the pure rotations given by $\boldsymbol{\omega}$ and $\boldsymbol{\varepsilon}$. In particular, the published source parameters are never corrected for the (assumed or estimated) glide in proper motion.