Bootstrapping *Swarm* and observatory data to generate candidates for the DGRF and IGRF.

Pavón-Carrasco, F.J.[1,2*], Marsal, S.[3], Torta, J.M.[3], Martín-Hernández, F.[1,2], Catalán, M.[4], Tordesillas, J.M.[5].

[1] Universidad Complutense de Madrid, Madrid. *Team leader and corresponding author (fjpavon@ucm.es).
[2] Geoscience Institute IGEO (CSIC- UCM), Madrid.
[3] Observatori de l'Ebre (OE), Univ. Ramon Llull – CSIC, Roquetes, Tarragona.
[4] Real Observatorio Geofísico de la Armada, San Fernando, Cádiz.
[5] Observatorio Geofísico de Toledo, Instituto Geográfico Nacional, Toledo.

Technical Notes

1. Input data for DGRF and IGRF.

We have used *Swarm* satellite data of the European Space Agency (ESA) and geomagnetic observatory data.

1a) Satellite data.

DGRF-2015. To obtain the DGRF-2015 candidate we have generated a parent model (called "DGRF parent model") using *Swarm* magnetic data from 1st July 2014 to 30th June 2015 (1-yr time window centred around 1st January 2015). We have used the three satellites Alpha (A), Bravo (B) and Charlie (C). From the FTP server of ESA (*Swarm*-diss.eo.esa.int), we have downloaded all the *Swarm* magnetic data available for this time period corresponding to the L1b data product Mag-L (level version _0505_, i.e., the last version in the ESA server). We have used the scalar data (i.e., the total field) given by the Absolute Scalar Magnetometer (ASM) and the vector data (i.e., the XYZ or $B_\theta B_\lambda B_r$ elements in the NEC frame) from the Vector Field Magnetometer (VFM). Note that since 5th November 2014, *Swarm* C does not provide ASM scalar data and we have estimated this geomagnetic element using the vector data.

ESA provides two sampling frequencies for magnetic data: 1 Hz (low resolution, denoted as LR in the files) and 50 Hz (high resolution, HR). For the DGRF parent model we have used the low resolution (LR) data, i.e., 1 measurement per second. Daily magnetic data of *Swarm* A and C cover the whole time interval (i.e., 365 data files for *Swarm* A and C). However, no data are available for *Swarm* B on 11th January 2015 (364 data files for *Swarm* B).

IGRF-2020. The IGRF-2020 candidate is estimated from a parent model denoted as "IGRF parent model". This model is developed using the same type of *Swarm* data as the DGRF parent model, but downloading the *Swarm* A-B-C data from 1st September 2018 to 15th September 2019 (380 days). For *Swarm* A there are not data available for the days 10th and 11th June 2019 (378 data files). For *Swarm* B the 17th August 2019 is not available (379 data files). For *Swarm* C the days 29th and 30th April 2019, 1st May 2019, and 16th July 2019 are not available (376 data files).

1b) Observatory data.

DGRF-2015. To generate the DGRF parent model, we have used hourly mean values from a total of 159 geomagnetic observatories spanning the annual period 1st July 2014 – 30th June 2015. The data sets were obtained from the portal of the WDC for Geomagnetism in Edinburgh (www.wdc.bgs.ac.uk/dataportal/) and include definitive data only.

List of observatories (IAGA code) used for the DGRF parent model: *AAA, ABG, ABK, AIA, ALH, API, ARS, ASC, ASP, BDV, BEL, BFO, BGY, BJN, BLC, BMT, BOU, BOX, BRD, BRW, BSL, CBB, CBI, CDP, CKI, CLF, CMO, CNB, CNH, COI, CSY, CTA, CTS, CYG, CZT, DED, DLT, DOB, DOU, EBR, ELT, ESA, ESK, EYR, FCC, FRD, FRN, FUR, GAN, GCK, GDH, GLM, GNG, GUA, GUI, GUL, GZH, HAD, HBK, HER, HLP, HON, HRB, HRN, HUA, HYB, IPM, IQA, IRT, IZN, JAI, JCO, KAK, KDU, KEP, KHB, KIR, KIV, KMH, KNY,*

KNZ, KOU, KPG, KSH, LER, LIV, LON, LRM, LRV, LVV, LYC, LZH, MAB, MAW, MBO, MCQ, MEA, MGD, MIZ, MMB, MZL, NAQ, NCK, NEW, NGK, NGP, NMP, NUR, NVS, OTT, PAG, PET, PHU, PIL, PLR, PND, PPT, PST, QGZ, QIX, QZH, RES, RKT, SBA, SBL, SFS, SHL, SHU, SIL, SIT, SJG, SOD, SPG, SPT, STJ, SUA, TAM, TDC, THJ, THL, THY, TIR, TRO, TRW, TSU, TTB, TUC, UPS, VAL, VIC, VNA, VOS, VSK, VSS, WHN, WIK, WNG, YAK, YKC.

IGRF-2020. To generate the IGRF parent model, we have used the available 1-minute data from a total of 75 geomagnetic observatories spanning the annual period 1st September 2018 – 31st August 2019. The data sets were obtained from the INTERMAGNET portal (intermagnet.org/data-donnee/download-eng.php) and include the best available data type, either definitive or quasi-definitive (provisional and variation data have been excluded). Hourly mean values have been used throughout as a basis for our analysis concerning observatory data. These have been calculated from the 1-minute data in this case.

List of observatories (IAGA code) used for the IGRF parent model: *ABG, ABK, ASC, ASP, BEL, BFO, BOU, BOX, BRW, CKI, CLF, CMO, CNB, CSY, CTA, CYG, DLT, DOU, EBR, ESK, EYR, FRD, FRN, FUR, GDH, GNG, GUA, HAD, HBK, HER, HLP, HON, HRN, HYB, IRT, IZN, JCO, KAK, KDU, KEP, KHB, KMH, KNY, KOU, LER, LON, LRM, LYC, LZH, MAB, MAW, MCQ, MGD, MMB, NAQ, NEW, NGK, NVS, PET, PST, SBA, SBL, SFS, SHU, SIT, SJG, SPT, SUA, TAM, THL, THY, TUC, UPS, WIC, WNG.*

2. Data selection and rejection criteria.

2a) *Swarm* data.

DGRF parent model. The data selection criteria are listed in Table 1. We have used the *OMNIWeb* site of NASA (https://omniweb.gsfc.nasa.gov/form/dx1.html) to get the geomagnetic indices and near-Earth solar wind magnetic field and plasma parameters for the period 1st July 2014 – 30th June 2015.

Table 1. Data selection criteria for Swarm data.

Swarm data selection
• Data from dark regions: Sun at least 10° below the horizon.
• $Vector_{VFM}$ data in non-polar regions: $\lvert QDL \rvert \leq 55°$ (QDL: Quasi-dipole latitude).
• $Scalar_{ASM}$ data in polar regions: $\lvert QDL \rvert > 55°$.
• $\lvert Dst \rvert < 30$ nT.
• $\lvert dDst/dt \rvert < 2$ nT/h for non-polar data, $\lvert dDst/dt \rvert < 5$ nT/h for polar data.
• $a_p < 10$; a_p (3-hr before) < 12; a_p (3-hr after) < 12.
• $\lvert IMF\ B_y \rvert < 8$ nT (IMF: Interplanetary Magnetic Field).
• $-2 < IMF\ B_z < 6$ nT.
• $E_m < 0.8$ mV/m for polar data (E_m: Merging electric field at the magnetopause. We have used the expression given in Finlay et al., 2015).
• $\lvert Scalar_{ASM} - Scalar_{VFM} \rvert < 3$ nT.
• $\lvert Vector_{VFM} - Vector_{CHAOS-6} \rvert < 500$ nT (CHAOS-6, Finlay et al., 2015).
• $\lvert Scalar_{ASM} - Scalar_{CHAOS-6} \rvert < 100$nT.

After applying the selection criteria, we are left with **3605739 *Swarm* A**, **3650850 *Swarm* B** and **3549601 *Swarm* C** data for the total time interval. Figure 1 shows the spatial and temporal distribution of *Swarm* A data (for the time-axis, we have used the modified Julian days referred to 2015.0). For *Swarm* B and *Swarm* C, both spatial and temporal distributions are similar.
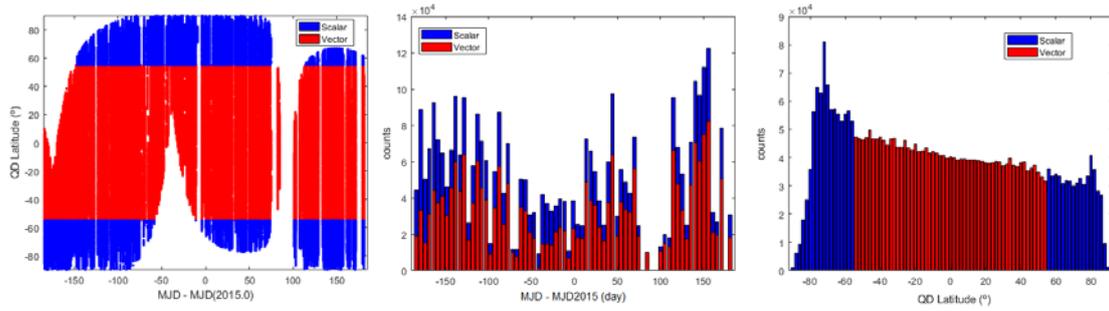
Figure 1. a) Spatial distribution of Swarm A data according to the quasi-dipole latitude and time in modified Julian days. b) Temporal histogram of the number of data (scalar and vector) using bins of 5 days. c) Number of Swarm A data as a function of the quasi-dipole latitude (bins every 2 °).

IGRF parent model. We have used the data selection criteria of Table 1. However, there are no indices available in the *OMNIWeb* site for the period from 8th August 2019 to 15th September 2019, so we have resorted to the provisional Dst and ap indices from the World Data Center for Geomagnetism, Kyoto (http://wdc.kugi.kyoto-u.ac.jp/kp/index.html). The IMF and E_m thresholds have not been used for that time interval. In addition, the last two selection criteria related to the comparison with the CHAOS-6 model is only applied up to April 2019 (after this date, CHAOS-6 is not available).

After applying the selection criteria, we are left with **5875141** ***Swarm*** **A**, **5806766** ***Swarm*** **B** and **5957684** ***Swarm*** **C** data for the total time interval. Figure 2 shows the spatial and temporal distribution of *Swarm* A data (for the time-axis, we have used the modified Julian days referred to 1st March 2019, approx. 2019.16). For *Swarm* B and *Swarm* C, both spatial and temporal distributions are similar.
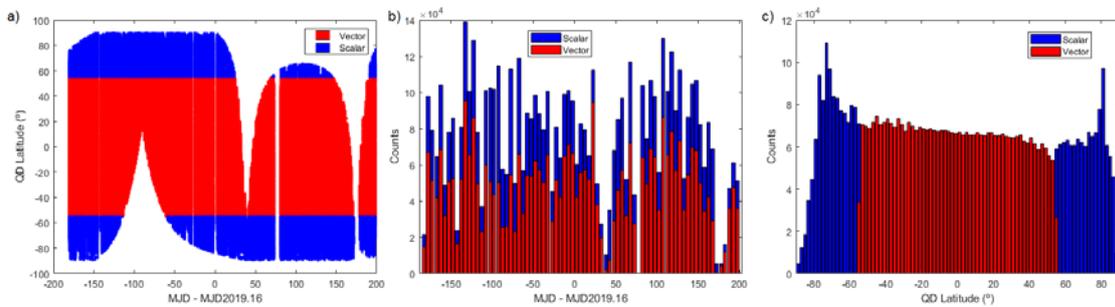

Figure 2. a) Spatial distribution of Swarm A data according to the quasi-dipole latitude and time in modified Julian days. b) Temporal histogram of the number of data (scalar and vector) using bins of 5 days. c) Number of Swarm A data as a function of the quasi-dipole latitude (bins every 2 °).
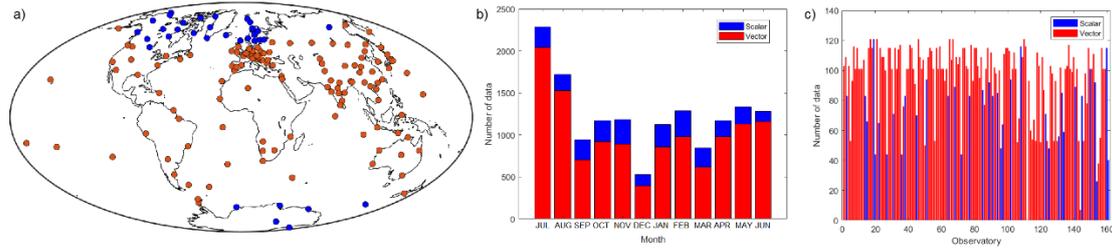
2b) Observatory data.

After a thorough revision of the observatory data sets and rejection of inconsistent or suspicious values, we have applied the following criteria for the selection of quiet-time intervals for both DGRF and IGRF parent models (Table 2).

Figure 3 shows the location of the observatories and the temporal distribution according to the scalar and vector data for the DGRF and IGRF parent models.

Table 2. Data selection criteria for observatory data.

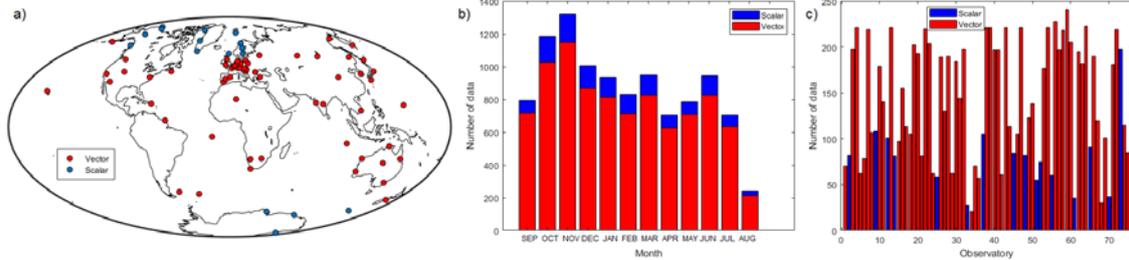| Observatory data selection |
|---|
| • Vector data in non-polar regions, i.e. $|GL| \leq 59°$ (GL: geomagnetic latitude). |
| • Scalar data in polar regions, i.e. $|GL| > 59°$. |
| • Local midnight hourly values: 01 – 02 Local Time. |
| • $Kp \leq 1^+$ ($a_p \leq 5$ nT) for observatories in non-polar regions. |
| • $AE \leq 50$ nT for observatories in polar regions for the DGRF parent model. |
| • $Kp \leq 0^+$ ($a_p \leq 2$ nT) for observatories in polar regions for the IGRF parent model (AE indices are not available for this period). |

Figure 3. a) Spatial distribution of the observatories. b) Number of observatory data per month. c) Number of data per observatory (ordered in alphabetical order according to their IAGA code). Upper panel: data for DGRF parent model. Lower panel: data for IGRF parent model.

3. Weighting scheme.

Satellite and observatory data have been weighted according to their spatial distribution using a Kernel function that takes into account the spatial distribution of the data and the geocentric angular distance between them.

For observatory data, this weighting scheme is important to avoid possible biases due to the high concentration of observatories in some regions (e.g. Europe). Using the Kernel function we have established the weight for the observatories between 1 (corresponding to a mean angular distance of 50°) and 2.4 (mean angular distance of 150°) depending on the geocentric location. Figure 4 shows the weights of the 159 observatories for the DGRF parent model and 75 observatories for the IGRF parent model.

For *Swarm* data, we have used the same kernel function. However, our modelling approach (see section 5) imposes a homogeneous distribution for the satellite data and then all satellite data have the same weight fixed as 1.8 (corresponding to the maximum of the kernel function for an angular distance over the sphere of 90°).

In terms of time, no weighting scheme has been applied. However, for the observatory data, we have selected the data following a homogenous distribution in time with one value per month (see section 5).
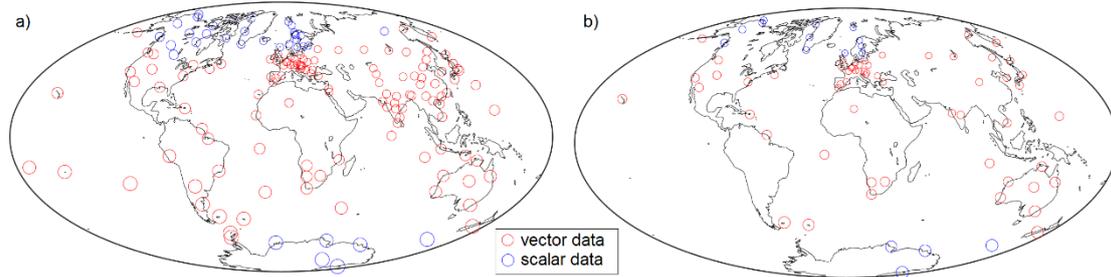

Figure 4. Assigned weights for the observatories (the smallest circles correspond to weight 1 and the biggest to weight 2.4). Observatories for the a) DGRF and b) IGRF parent models.

4. Estimated sources and models.

We have only pre-estimated the crustal field and removed it from the satellite data using the crustal model LCS-1 (http://www.spacecenter.dk/files/magnetic-models/LCS-1/, Olsen et al., 2017). No crustal field is extracted from the observatory data since we have used these data just to constrain the secular variation and consequently the crustal field is automatically removed. The core and magnetospheric (and their Earth-induced counterparts) fields are directly modelled from the input data to provide both DGRF and IGRF parent models (more details in section 5).

5. Modelling approach, starting model, scalar linearization, regularization, least squares equations, number of iterations, and Gauss coefficient uncertainties.

We have used a **_bootstrap method_** to resample both satellite and observatory data. This approach provides robust outputs (**ensemble models**) to better estimate the Gauss coefficients and to have more reliable and realistic uncertainties (despite these uncertainties can be higher than those given by a single-inversion model). This is important for the IGRF-13 product, where an artificial extrapolation to 1st January 2020 must be applied.

Below, we list the steps to jointly model both vector and scalar data (the same procedure has been used for DGRF and IGRF parent models):

Step 1. Satellite input data.

- A grid of 1000 points homogeneously distributed over the sphere is used to resample the satellite data. Each point in this grid represents a spherical cap of approx. 3.3° semi-angle over the sphere (see Figure 5). We use different grids for *Swarm* A, B, C satellites.
- We have used these 1000 spherical caps to separate the *Swarm* data in areas homogeneously distributed over the sphere. Figure 5 shows the spatial distribution, in terms of spherical caps of 3.3°, of the *Swarm* A data for the time window of the DGRF parent model (this figure is similar for *Swarm* B and C, and for the *Swarm* data for the IGRF parent model).
- Then, we select a sub-dataset of *Swarm* data. To do that, we randomly pick a fix number of *Swarm* data for each spherical cap: 2 data for *Swarm* A and 2 for *Swarm* C (they fly approximately at the same altitude) and 4 data for *Swarm* B. Consequently, the random

sub-dataset contains 2000 *Swarm* A data, 2000 *Swarm* B data and 4000 *Swarm* C data (in total, 8000 data). We use this type of random dataset to generate the parent models following the modelling approach detailed below.

- We repeat this procedure 1000 times obtaining 1000 different random datasets of 8000 *Swarm* data each. Note that for each dataset, the 8000 *Swarm* random data are homogeneously distributed over the sphere.
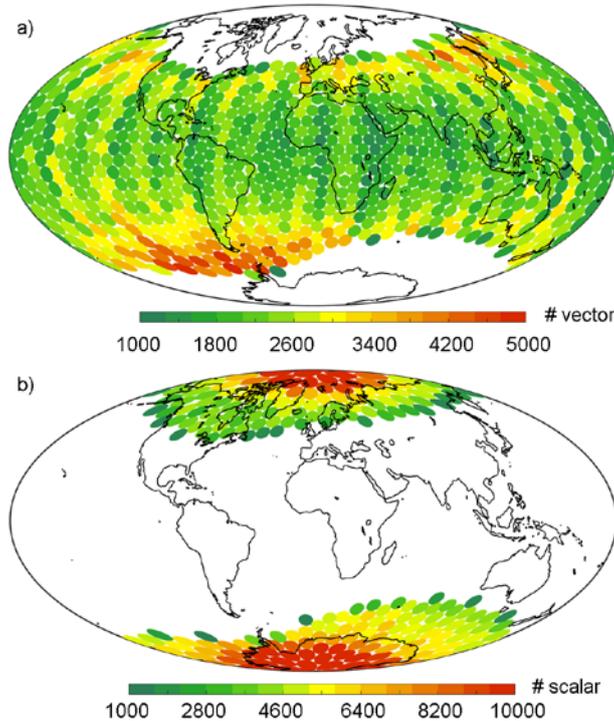

Figure 5. a) Number of vector Swarm A data within the spherical caps of 3.3° semi-angle. b) Number of scalar Swarm A data within the spherical caps of 3.3° semi-angle.

Step 2. Observatory input data.

- We have randomly resampled the observatory data keeping 1 value per month for each observatory (when available). Consequently, we have 12xD observatory data per year, where D is the number of observatories used (D = 159 for the DGRF parent model and D = 75 for the IGRF parent model).
- We repeat this procedure 1000 times and again, we obtain 1000 different random datasets of observatory data.

Step 3. Modelling approach.

- Using the first 8000-dataset of *Swarm* data and the first 12xD-dataset of observatories, we generate a first parent model for both DGRF and IGRF following the weighted least-square inversion.
- For the parent model, we have estimated the core field using the classical expansion of the geomagnetic potential in spherical harmonic functions. We have used an expansion from degree n =1 to 20 with Gauss coefficients g_n^m (and h_n^m) linearly depending on time:

$$g_n^m(t) = g_n^m|_{t_0} + \dot{g}_n^m \cdot (t - t_0)$$

where t_0 is a reference date for each parent model ($t_0 = 2015.0$ for DGRF and 2019.16 for IGRF). This means an estimation of 880 parameters representing the core field.

- To model the external field, we follow the methodology of the last CHAOS model (Finlay et al., 2015) using the external potential expression:

$$V^{ext} = a \sum_{n=1}^{2} \sum_{m=0}^{n} \left(\frac{r}{a}\right)^n P_n^m(cos\theta_d)(q_n^m cosmT_d + s_n^m sinmT_d) +$$

$$+ a \sum_{n=1}^{2} q_{GSM,n}^0 R_n^0(r,\theta,\lambda).$$

The first addend represents the near magnetospheric sources using a harmonic expansion in solar magnetic coordinates. For this term, the degree-1 coefficients (q_1^0, q_1^1, s_1^1) depend on time as a function of both induced and external Dst index (or magnetospheric ring current index, RC, see Olsen et al., 2014) as follows (same for s_1^1):

$$q_1^m(t) = \hat{q}_1^m \left[E(t) + I(t) \left(\frac{a}{r}\right)^3\right] + \Delta q_1^m(t)$$

where *E(t)* and *I(t)* represent the external and induced Dst or RC indices. $\hat{q}_1^m$ is a constant parameter to be determined and $\Delta q_1^m(t)$ represent a set of temporal baseline corrections homogeneously distributed within the considered time window. Following Finlay et al. (2015), we estimated $\Delta q_1^0(t)$ in bins of 5 days and $\Delta q_1^1(t)$ and $\Delta s_1^1(t)$ in bins of 30 days for both DGRF and IGRF parent models. The second addend represents the remote magnetospheric currents by using a spherical harmonic expansion in terms of the geocentric solar magnetospheric coordinates (GSM). According to this potential expansion, the external field is estimated by 10 constant parameters (3 $\hat{q}_1^m$, 5 q_2^m, and 2 $q_{GSM,n}^0$) plus 99 baselines (73 for Δq_1^0, 13 Δq_1^1, 13 Δs_1^1) for the DGRF parent model and 102 baselines (76 for Δq_1^0, 13 Δq_1^1, 13 Δs_1^1) for the IGRF parent model.

- For the **DGRF** parent model we have to estimate a total of **989 parameters** (880 for the core field and 109 for the external field).
- For the **IGRF** parent model we have to estimate **992 parameters** (880 for the core field and 112 for the external field).
- In order to jointly model both vector (non-polar areas) and scalar (polar areas) data, we have applied a linearization approach for the scalar element that depends on the matrix of spatial and temporal parameters of both internal and external spherical harmonic expansion. This linearization involves the use of an iterative approach using an initial model. We have used a constant axial dipole field of -30000 nT as the g_1^0 Gauss coefficient for the initial model (a null starting external field was considered). The inversion problem is carried out using the iterative least-square method using the weight matrix W described in section 3:

$$m_{i+1} = m_i + (A_i' \cdot W \cdot A_i)^{-1} A_i' \cdot W \cdot (A_i' \cdot m_i - \delta)$$

where m_i is the vector containing both core and external coefficients and baselines for the iteration i. A_i is the matrix of parameters calculated by using the Fréchet derivative around the iteration i, and δ is the vector with the input data.

- The parent models converge after 3 iterations. Figure 6 shows the differences between consecutive iterations for the estimated parameters contained in the vector m (core and external Gauss coefficients and external baselines). After 3 iterations, no differences are observed.
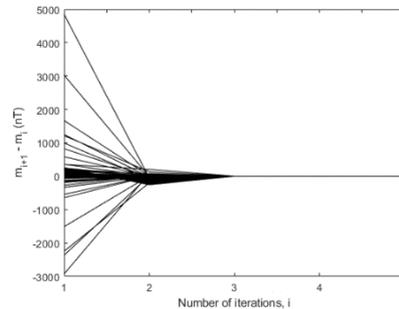

Figure 6. Difference between consecutive model coefficients as a function of the number of iterations for one of the DGRF parent model.

- No type of regularization has been applied in our modelling inversion.
- To estimate the core field (degree n=1 to 20), both satellite and observatory data have been used. However, the observatory data only constrain the secular variation, since we have used the differences between two time-consecutive data.
- To estimate the external field, we have used both satellite and observatory data. For the DGRF parent model we have used the RC index of Finlay et al. (2015) (http://www.spacecenter.dk/files/magnetic-models/CHAOS-6/) and for the IGRF parent model the DST index obtained from the NOAA database (https://www.ngdc.noaa.gov/geomag/est_ist.shtml).

Step 4. Bootstrap to generate an ensemble of 1000 parent models.

- Then, we have repeated the previous steps, but now using the second 8000-dataset of *Swarm* data and the second 12x*D*-dataset of observatory data obtaining a new set of coefficients and baselines for the parent model. Successively and using the 1000 sub-datasets, we obtain an ensemble of 1000-parent models. The ensemble of 1000-parent models provide a set of Gauss coefficients (and baselines) that follows pretty well a Gaussian or normal distribution. Using this normal distribution, we can estimate the mean value and its uncertainty for each coefficient (baseline). This approach is useful to infer the effect of the reliability of the spatial and temporal distribution of the input data to the final set of modelled parameters. Figure 7 shows the histograms of some Gauss coefficients (and baselines) for the ensemble of 1000-parent models for the DGRF and IGRF candidates.
- The DGRF candidate is obtained from the ensemble of 1000-parent models from July 1st 2014 to June 30th 2015 for a time t = 2015.0 and using the degrees 1 to 13 (core field).
- The IGRF candidate is obtained from the 1000-ensemble parent models from September 1st 2018 to September 15th 2019 using a linear extrapolation up to t = 2020.0. The core field is given for degrees 1 to 13 and its secular variation for degrees 1 to 8.
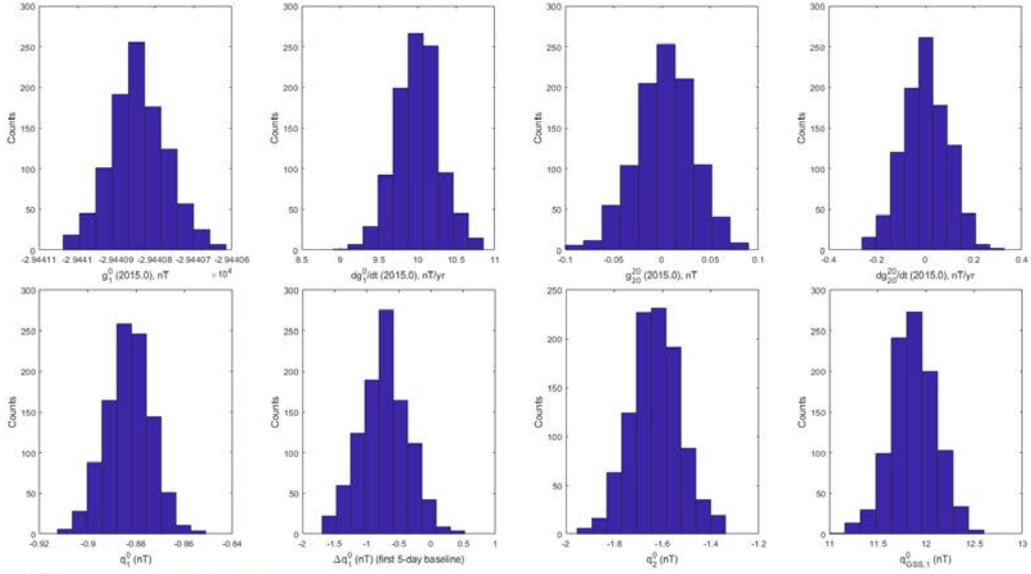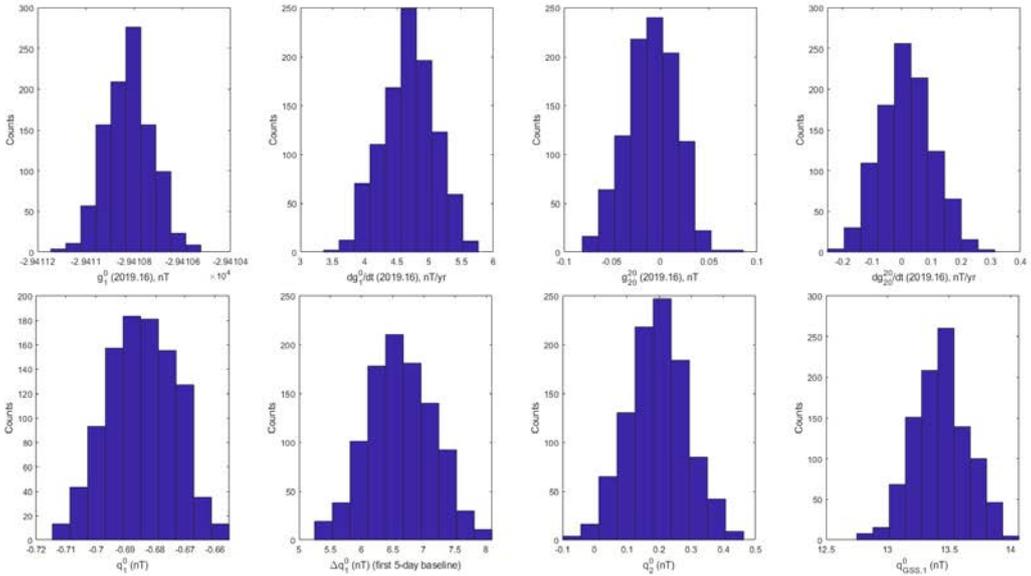
Figure 7. Some Gauss coefficients and baseline corrections for both DGRF and IGRF parent models.

6. Extrapolation of IGRF to 2020.0. Prediction of secular variation from 2020.0 to 2025.0.

The IGRF-2020 (core field and secular variation) has been extracted from the 1000-ensemble parent model for the time period 1st September 2018 to 15th September 2019 by a linear extrapolation:

$$g_n^m(2020.0) = g_n^m|_{2019.16} + \dot{g}_n^m \cdot (2020.0 - 2019.16),$$
$$\dot{g}_n^m(2020.0) = \dot{g}_n^m.$$

Figure 8 shows this extrapolation for the first Gauss coefficient (dashed lines indicate the Gauss coefficient uncertainty).
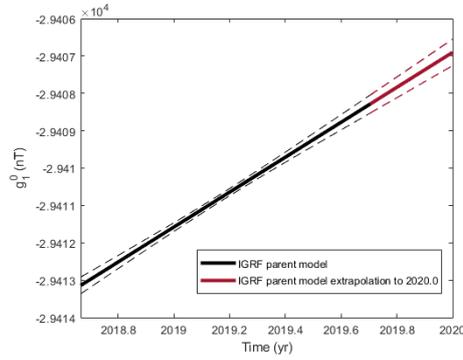
Figure 8. Extrapolation of the IGRF parent model to 2020.0 for g_1^0.

7. Data fitting and residuals.

The following tables and figures show the data misfit (in terms of means and RMS) for Swarm and observatory data for the DGRF and IGRF parent models.

7a) DGRF parent model.

***Swarm* data** (approx. 6500 vector data and 1500 scalar data for each ensemble model. Residuals are calculated for the 1000 different dataset).

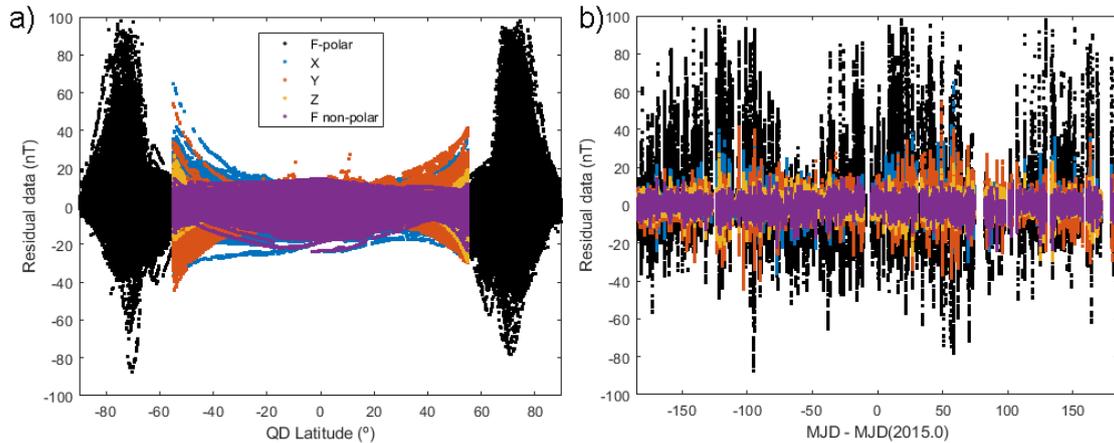| Geomagnetic Element	Mean (nT)	RMS (nT)
X (B_θ)	0.01	3.57
Y (B_λ)	-0.05	3.25
Z (B_r)	-0.03	2.30
F (non-polar)	0.01	2.76
F (polar)	0.14	10.12

Figure 9. Residuals vs a) quasi-dipole latitude and b) modified Julian days (centred at January 1st 2015) of the Swarm data for the DGRF parent model.

Observatory data (1380 vector data and 412 scalar data for each ensemble model. Residuals are calculated using the total set of observatory data). "dX" means the difference for the element X in two time-consecutive measurements (estimation of the secular variation). "X − mean(X)" represents the difference between the X element for an observatory and the mean value of that element (since we do not model the anomaly biases of observatories, it is the only way to estimate the residual).

Geomagnetic Element	Mean (nT)	RMS (nT)
dX	-0.01	2.92
dY	0.01	3.41
dZ	0.01	1.90
dF (non-polar)	0.01	2,24
dF (polar)	0.08	11.95

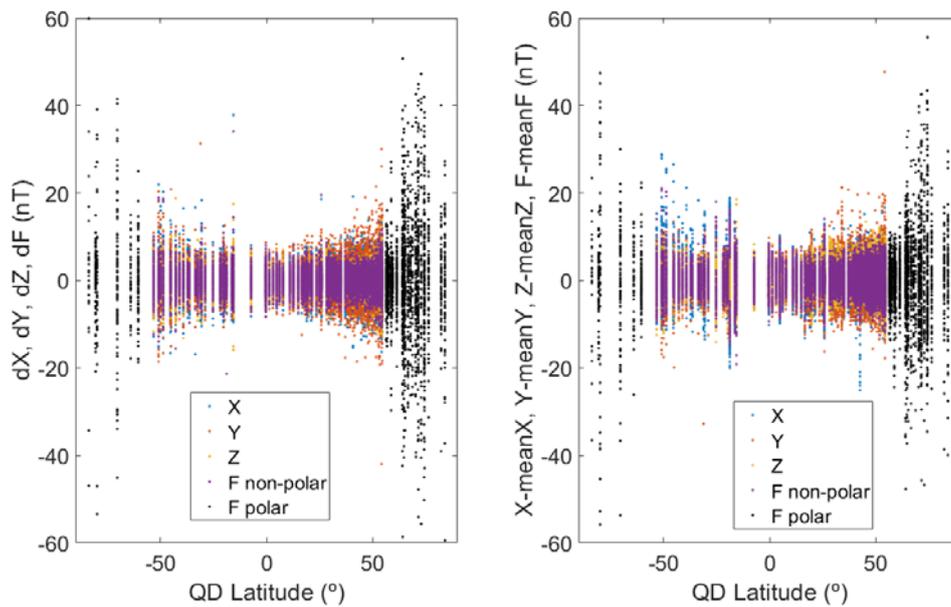| Geomagnetic Element	Mean (nT)	RMS (nT)
X – mean(X)	-	3.64
Y – mean(Y)	-	3.48
Z – mean(Z)	-	3.26
F – mean(F) (non-polar)	-	4.91
F – mean (F) (polar)	-	10.27

Figure 10. Residuals of the observatory data for the DGRF parent model. a) Residuals using the secular variation "dX". b) Residuals using the difference "X – mean(X)".

7b) IGRF parent model.

Swarm **data** (approx. 6500 vector data and 1500 scalar data for each ensemble model. Residuals are calculated for the 1000 different dataset).

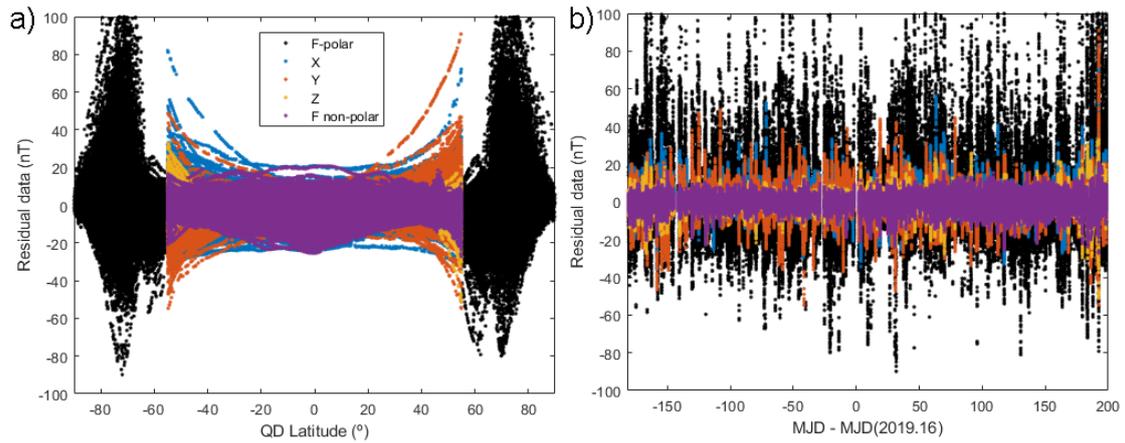| Geomagnetic Element	Mean (nT)	RMS (nT)
X (B_θ)	0.03	3.60
Y (B_λ)	0.16	3.26
Z (B_r)	-0.16	1.99
F (non-polar)	-0.03	2.86
F (polar)	0.00	11.00

Figure 11. Residuals vs a) quasi-dipole latitude and b) modified Julian days (centred at March 1st 2019) of the Swarm data for the IGRF parent model.

Observatory data (523 vector data and 148 scalar data for each ensemble model. Residuals are calculated using the total set of observatory data). "dX" means the difference for the element X in two time-consecutive measurements (estimation of the secular variation). "X – mean(X)" represents the difference between the X element for an observatory and the mean value of that element (since we do not model the anomaly biases of observatories, it is the only way to estimate the residual).

Geomagnetic Element	Mean (nT)	RMS (nT)
dX	0.05	3.44
dY	0.00	3.08
dZ	-0.01	1.82
dF (non-polar)	0.01	2,17
dF (polar)	-0.11	13.87

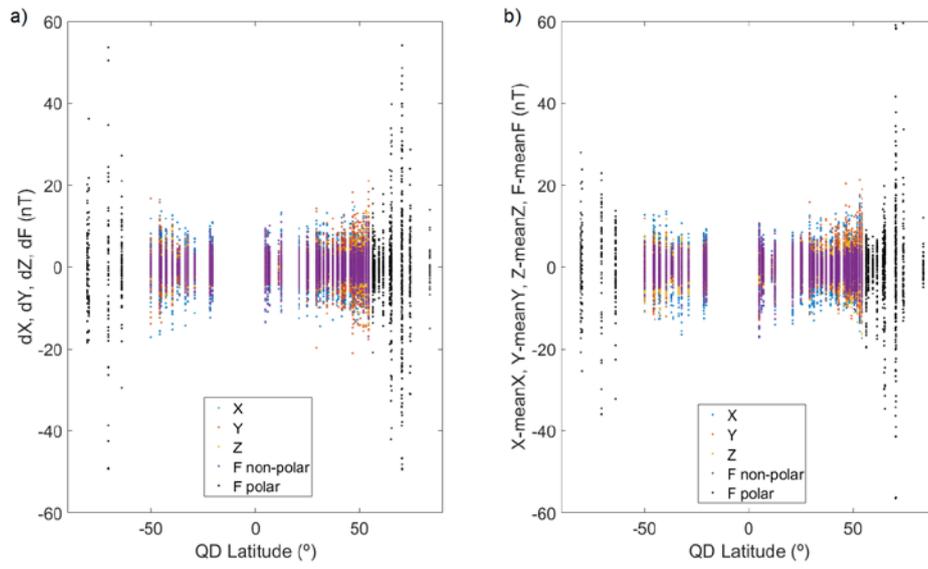| Geomagnetic Element	Mean (nT)	RMS (nT)
X – mean(X)	-	3.66
Y – mean(Y)	-	2.90
Z – mean(Z)	-	2.67
F – mean(F) (non-polar)	-	4.35
F – mean (F) (polar)	-	10.48

Figure 12. Residuals of the observatory data for the IGRF parent model. a) Residuals using the secular variation "dX". b) Residuals using the difference "X – mean(X)".