Earth, Planets and Space

FULL PAPER

Open Access

Machine learning-based calibration of the GOCE satellite platform magnetometers

Kevin Styp-Rekowski[1*], Ingo Michaelis[2], Claudia Stolle[3], Julien Baerenzung[2], Monika Korte[2] and Odej Kao[1]

Abstract

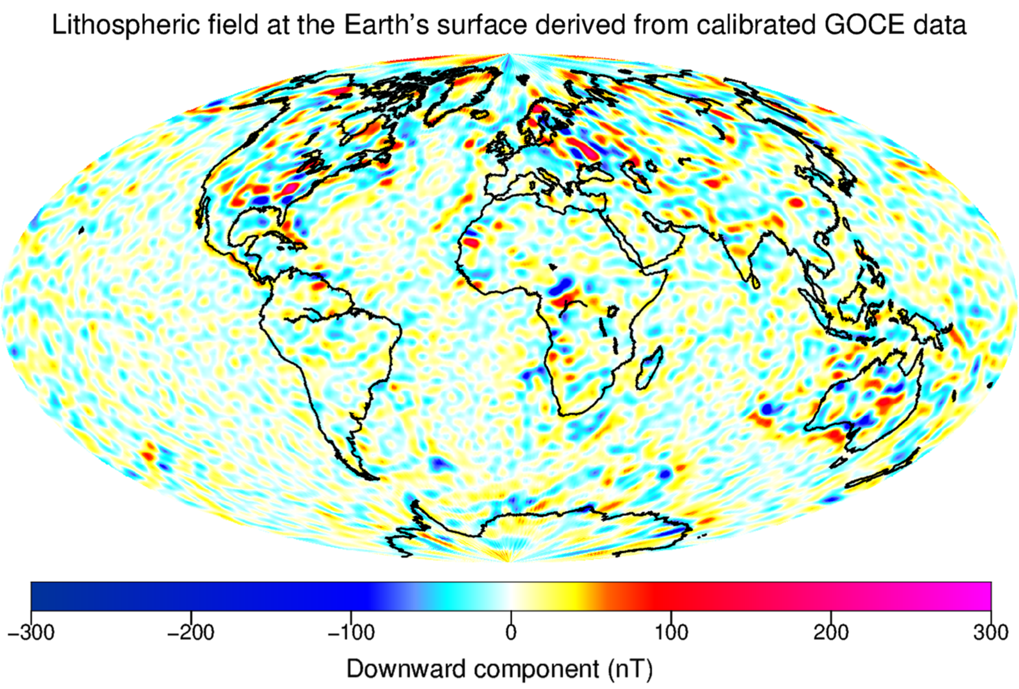
Additional datasets from space-based observations of the Earth's magnetic field are of high value to space physics and geomagnetism. The use of platform magnetometers from non-dedicated satellites has recently successfully provided additional spatial and temporal coverage of the magnetic field. The Gravity and steady-state Ocean Circulation Explorer (GOCE) mission was launched in March 2009 and ended in November 2013 with the purpose of measuring the Earth's gravity field. It also carried three platform magnetometers onboard. Careful calibration of the platform magnetometers can remove artificial disturbances caused by other satellite payload systems, improving the quality of the measurements. In this work, a machine learning-based approach is presented that uses neural networks to achieve a calibration that can incorporate a variety of collected information about the satellite system. The evaluation has shown that the approach is able to significantly reduce the calibration residual with a mean absolute residual of about 6.47nT for low- and mid-latitudes. In addition, the calibrated platform magnetometer data can be used for reconstructing the lithospheric field, due to the low altitude of the mission, and also observing other magnetic phenomena such as geomagnetic storms. Furthermore, the inclusion of the calibrated platform magnetometer data also allows improvement of geomagnetic field models. The calibrated dataset is published alongside this work.

Keywords: Machine learning, Calibration, Platform magnetometer, GOCE satellite, Magnetic field model

*Correspondence: styp-rekowski@tu-berlin.de

[1] Distributed and Operating Systems, Technical University of Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Full list of author information is available at the end of the article

Graphical Abstract


Lithospheric field at the Earth's surface derived from calibrated GOCE data
−300
−200
−100
0
100
200
300
Downward component (nT)

Introduction

Space-based observations of the Earth's magnetic field from low Earth orbit are of high value for space physics and geomagnetism due to their potential global coverage. High-precision magnetic satellite missions carrying a magnetometer package to provide absolute vector data of the geomagnetic field have revolutionized our knowledge of its distribution and variations (see, e.g., Olsen and Stolle (2012)). The most prominent missions with decade-long continuous time series are the CHAllenging Minisatellite Payload (CHAMP) (Reigber et al. 2002) and Swarm missions (Olsen et al. 2013). The measurements of these missions have led to high-quality models, describing different components of the geomagnetic field, e.g., the core, the lithospheric, and the large-scale magnetospheric field, e.g., (Finlay et al. 2020; Baerenzung et al. 2020).

In addition to high-precision magnetic mission data, the spatiotemporal coverage of magnetic measurements can be enhanced by applying observations of so-called platform magnetometers which are mounted on many satellites primarily for attitude control. Usually, they do not provide absolute data, have coarse sampling rates of one to several seconds, and are mounted with low attention to magnetic cleanliness. However, based on high-quality magnetic field models and information on the satellite attitude and on-board operations and control describing possible artificial magnetic signals, measurements of platform magnetometers can be carefully calibrated. Previously, platform magnetometer data from the GRACE, GRACE-FO, Cryosat-2, and DMSP missions have been calibrated and made publicly available (Olsen 2021; Stolle et al. 2021a; Olsen et al. 2020; Alken et al. 2020). A summary of these calibrations and examples of scientific applications in both geomagnetism and space science is given by Stolle et al. (2021b). Especially between the end of the CHAMP mission in September 2010 and the launch of the Swarm mission in November 2013, these data have had the potential to enhance magnetic field models.

In this work, we present a machine learning (ML) algorithm to calibrate platform magnetometer data of the Gravity and steady-state Ocean Circulation Explorer (GOCE) (Floberghagen et al. 2011; Drinkwater et al. 2003). The satellite was launched in March 2009 and ended in November 2013, i.e., being another satellite with the potential of bridging the previously mentioned gap of high-precision measurements. With its low altitude of about 255 km, the magnetic part of the GOCE mission may be especially interesting in detecting the lithospheric field. The GOCE satellite flies in a polar orbit of 98° inclination and the mission follows a sun-synchronous dawn–dusk orbit at local times of about 6 and 18 LT for the descending and ascending orbits, respectively. The spacecraft carries three 3-axis magnetometers of the type Billingsley TFM100-S measuring the Earth's magnetic field at a rate of 16 s (Billingsley 2020). More details on the mission and especially its magnetometer package is given by Michaelis et al. (2022).

All above-mentioned data from platform magnetometers including the dataset of GOCE by Michaelis et al. (2022) have applied analytical approaches solving a least-square problem. The purpose of this work is to develop and present a ML technique to calibrate magnetometer data with the aim to provide a method that does not need preselection of parameters that describe potential artificial magnetic disturbances. As an example, parameters like magnetorquer activations, battery currents, or sensor temperatures are known to contribute to such disturbances. Rather, all available parameters are fed into the process and the ML algorithm itself identifies relevant properties for the calibration. In addition, timing issues of the satellite clock or non-linear relationships between parameters are automatically identified. By that, we aim at providing a calibration tool for platform magnetometers that is easily applicable to other missions as well. A similar calibration has been presented earlier for the GRACE-FO satellite mission (Styp-Rekowski et al. 2021).

In the following, the Chapter "Datasets and preprocessing" describes the datasets used and the application of different preprocessing steps. Afterward, the Chapter "Machine learning-based calibration" describes the ML approach and its application to the data. The assessment of the results as well as examples of geophysical applications are presented in the Chapter "Results and discussion". Finally, our findings are concluded in the Chapter "Conclusion".

Datasets and preprocessing

The GOCE satellite produces data in an interval of 16 s. In a normal month containing 30 days, this means a total of 162000 data points. With an inclination of 98° degrees, the GOCE mission has a 61-day orbit periodicity, meaning that after this timespan it has covered the whole earth in a uniform pattern.

Dataset

The data used in this work consist of the same input dataset as described in Michaelis et al. (2022). The magnetometer measurements, the supporting housekeeping data, the available telemetry data, and the CHAOS7 data were interpolated onto the same timestamps because of different subsystems of the satellite measuring at different timestamps and time intervals. Contrary to Michaelis et al. (2022), all available data were used instead of selecting a subset. The measurements are recorded in the satellite frame and are provided by the European Space Agency (ESA). Therefore, calibration is performed in the same frame in which the instruments are mounted on the satellite since the available information affects the magnetometer measurements in this way. The CHAOS7 reference model was therefore also rotated into the satellite frame.

The available information was collected, interpolated, and finally merged into our final dataset. With available position and attitude information provided by the combined product of the Electrostatic Gravity Gradiometer and Star Trackers of the GOCE mission, the rotations were calculated such that the final calibration dataset of this study will be available in the satellite frame as well as the Earth-fixed North–East–Center (NEC) frame.

In this work, the measured properties or attributes by the satellite are referred to as features, which is a common term in ML. Features of the dataset used include the magnetometer readings, the magnetorquer activations, the solar array as well as battery and other available currents, temperatures, thruster activations, and also the available telemetry data of the satellite which includes a multitude of properties like status variables, flags, and others. They also carry a variety of physical units that are not considered further. Overall, there are 975 of the available 2233 features taken into account for the calibration after the application of the following preprocessing steps. The final list of used features can be found in the 'feature_list.csv' file published together with the dataset. The ML approach does not differentiate these features and will inherently select the most relevant features for the calibration. The magnetometer readings before the proposed calibration originate from the L1b product and have been pre-calibrated to a bias below 500nT.

Feature preprocessing

As previously mentioned, a multitude of features is collected to automatically identify relevant features for the calibration of the platform magnetometers. The more information can be collected and included about the satellite as a system, the better the calibration can potentially become. Before the ML approach is applied, the available data need to be preprocessed. While each new line in the data represents a new data record with an assigned unique timestamp, each of the columns or recorded measurements associated with this timestamp are referred to as features of the measurement for the rest of the work. In the first step of the analysis, these features will be converted, added, or filtered with a variety of goals.

One-hot encoding

As all available data of the GOCE mission are used, a lot of telemetry data are included as well which is partly delivered as textual information, e.g., certain states of systems being encoded with the literals 'AVAILABLE' or 'NOT_AVAILABLE'. As some features have more than 2 possible literals, a sophisticated method is needed to convert this textual information into numerical information which is usable by statistical models for the calibration. For each of the possible literals of a feature available, another new feature is generated which encodes a 1 if the feature equals this literal and 0 otherwise, this is also called one-hot encoding or dummy coding. Thus, if a certain literal that was recorded in a feature has an influence on the calibration, the calibration can utilize the associated feature. In addition, with this technique every resulting new feature has the same euclidean distance contrary to assigning number to every different literal available within the feature value range.

Removed features

Several features are explicitly not used for the calibration to perform well. These include timestamps, magnetic local time (MLT), and positional data like the latitude, longitude, or radius as these have the potential to encode positional information which needs to be avoided to not misguide the model calibration. When training against the reference model, the objective is not to learn an efficient mapping of position data to reference model values, but to model the magnetic conditions of the satellite system itself under different system activations. For the same reason, features like recorded measurements of the star trackers are removed as they inherit a potential to encode positional information about the satellite.

Additional features

The solar activity, given by the F10.7 index, an 81-day moving average of it as well as the day of the year were added as additional features (Tapping 2013). These were added to correct for potential influences from the angle of the satellite towards the sun or dependencies on the solar activity. Likewise, the 3-1-3-Euler angles were calculated using the quaternions and added as additional features to give the calibration model the possibility to adjust for inaccuracies in their estimation.

Missing data

Data gaps occur several times throughout the duration of the mission. These gaps are not part of the calibration because the data are unavailable. In addition, some positional data are either missing or measured incorrectly. For a minority of the data, there is no information about the position of the satellite, these data points have been removed. This was especially the case near data gaps like in July 2010. Gaps are caused either by missing data from the magnetometers or by lack of essential information such as position or attitude Also, data where the interpolation distance during the data gathering process described in Michaelis et al. (2022) was larger than 16 s are removed from the calibration process and flagged in the error flag "B_FLAG" as this data is considered similar to missing data because the data has to be considered uncertain.

NaN-filling

If only small shares of the additional housekeeping data are missing, a filling strategy is used to make these data usable again. For each feature, the share of missing data is calculated. If the share of missing data within a feature is smaller than 20% of the available data, the information situation is considered to be well enough to take these features into account as the features deliver potentially beneficial information in at least 80% of the cases which constitutes the majority. For these features, the missing share is filled up with the mean of the present values. The data points are flagged with an introduced "NaN-Flag" which indicates that data have been substituted with non-original measurements, 1 indicating that at least one feature value has been filled, 0 indicating no manipulation of the data.

Magnetic quiet time filtering

As the satellite is calibrated post-launch, the data need to be filtered for calibration to ensure that only the magnetic system of the satellite itself is modeled. Therefore, natural disturbances need to be removed before calibration, as they shall remain in the calibrated measurements of the satellite. The Kp and Dst indices are good indicators for magnetic activity. The data are filtered for values of $Kp \leq 2$ (Matzka et al. 2021) and $Dst \leq 30$ (Nose et al. 2015) which is considered to contain magnetic quiet times without omitting too much data for the training. Thus, the calibration shall model the underlying process of the satellite. Data filtered out like that is flagged with the "KP_Dst_Flag", 1 indicating a too high magnetic activity, 0 indicating low magnetic activity.

Outlier removal

Despite careful selection of the data and filtering for missing positional information as well as magnetic active times, some data points deviate strongly from their expected value. It is suspected that this behavior originates in a bad attitude estimation of the satellite but the underlying reasons can be manifold. To identify such data, the raw measurements of the magnetometers are compared to the reference model without any further calibration. The raw data follow the rough pre-calibration of the platform magnetometers as used during the mission. From the remaining residuals, the mean $\overline{x}_j$ is calculated as well as its standard deviation (STD). Then, it is checked whether each data point falls within the range of three standard deviations around the mean:

$$\text{with } \overline{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{i,j} \tag{1}$$

$$|x_{i,j}| \leq |\overline{x}_j| + 3 * \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{i,j} - \overline{x}_j)^2} \tag{2}$$

for all data points i and all magnetometer measurements j. This means at least 99.7% of data lies within the defined range if the data are Gaussian distributed which holds true for the GOCE dataset. If at least one of the magnetometer measurements deviates stronger from the mean, the whole orbit is considered to be an outlier and removed from the training data of the calibration, because analysis has shown that the whole orbit of the satellite behaves unusual when outliers are detected within the orbit. In addition, this data is flagged within the error flag ”B_Flag” and is part of the published dataset.

CHAOS-7 model

As the reference model for the calibration, the CHAOS-7.8 model has been used (Finlay et al. 2020). Utilizing a variety of magnetic measurements, including ground station observations as well as satellite data from current and previous high-precision magnetic satellite missions, the CHAOS-7 model is able to precisely model the core, the lithospheric, and the external field. This reference model is especially needed as the platform magnetometers onboard the GOCE satellite mission do not have an absolute component, but rather measure the Earth’s magnetic field relatively. Thereby, similar to Michaelis et al. (2022) the reference model has been evaluated at each position of the satellite and was then rotated into the satellite frame for calibration purposes.

Machine learning-based calibration

The proposed method utilizes neural networks (NNs) to train a ML-based model for the calibration and characterization of the platform magnetometer measurements. In the proposed approach, the calibration and characterization are accomplished in one method and thus referred to as calibration in the remainder of this work. Contrary to previous work done on the calibration of platform magnetometers like for the GRACE, GRACE-FO, or Cryosat-2 satellites, there are several differences in the approach used within this work. In the proposed approach, the features are not selected manually based on experience about their relevance on improving the calibration result, rather as much information as possible about the satellite as a magnetic system is collected and presented to the calibration model which is able to choose the relevant features. Also, no interactions between the features need to be hand-crafted, e.g., polynomial combinations of the magnetometer axes, since the calibration method used has inherent non-linear modeling capabilities. Another difference lies in the usage of data from all available latitudes, this also includes the high-latitude area containing field-aligned currents (FACs).

Utilizing a large number of collected features as described in the data preprocessing step, renders linear models like linear regression unsuitable as the underlying statistical problem is over-parameterized and thus the calibrated model overfits the data. Overfitting means that the model performs very well on the seen data, but generalizes poorly to unseen data. NNs can overcome this problem as the number of neurons is limited and thus, the number of used features gets limited. In addition, the batch-wise learning of the stochastic gradient descent algorithm prevents the use of single uncorrelated features. However, NNs need a large number of data points for the gradient-based approach to converge to an optimum, representing a good calibration of the magnetometer measurements. This is also referred to as training the NN as the weights representing the model are adjusted within this process. This is contradictory to the slowly deteriorating instruments onboard the satellite system for which many temporally separated models would be better capable of representing the changing behavior of the instruments. Therefore, the approach shown in Fig. 1 was developed. After collecting, filtering,

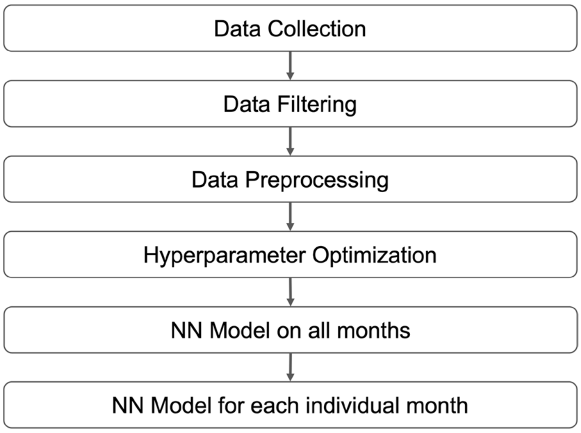


Fig. 1 Overview of the calibration process

and preprocessing the data as described in the previous section, first a Hyperparameter Optimization (HPO) on a randomly selected subset of the dataset is performed to determine the NN architecture and training parameters (see section below), then a NN model on all available data from the life span of the mission from November 2009 until September 2013 is trained using the found architecture, representing a global model of the calibration of the magnetometers. Afterward, a more fine-granularly trained model is trained for each month individually based on the global model, this is done with a much smaller learning rate when adjusting the weights of the NN. By using this approach, there is enough to train a global model which was presented with all variations of the data's statistics while still maintaining multiple models representing the differences between individual months, thus reflecting the change in the instruments throughout the duration of the mission. The fact that the GOCE mission flies in a sun-fixed orbit with stable MLTs supports the training of a global model as no changes in the behavior of the parameters are to be expected.

Before starting the training process, the STD of every feature as well as the auto-correlation among the features have been calculated. Features having an STD of 0 (depending on the computer precision) have been removed, meaning that these contain only constant values with no additional information for the calibration process. In addition, features having a correlation of 1 (depending on the computer precision) with any other feature have been removed as these represent the exact same information twice and are thus treated as duplicates of the same feature and removed. The STD-filtering removed 442 features, while the following correlation-filtering removed 780 additional features, as the telemetry dataset contains a lot of duplicated information. The reduced dataset dimension also helps to reduce the training time needed for the ML approach. In addition, the features are normalized with a min–max normalization to an interval of [− 1, 1], which is a technique by which the gradients converge to an optimal solution more smoothly when training NNs. This also ensures that each feature and the variations of the values within the feature are weighted equally during the training of the model. As mentioned earlier, there are 975 of the 2233 features taken into account during the calibration which includes the magnetometer measurements, the housekeeping data, and the telemetry data. Also, the applied filtering methods led to a reduction from about 6.4 million data points to about 4.8 million data points.

As mentioned earlier, NNs are not very well able to generalize to unseen values of features. Therefore, the training of the NN cannot be limited to the low- and mid-latitude region where the reference model is known to constitute the ground truth correctly as there are no magnetic phenomena in this region that are not modeled by the CHAOS-7 model. On the other hand, in the high-latitude regions the values measured by the magnetometers lie beyond the interval of previously seen values in the low- and mid-latitude regions, thus the NN would not be able to generalize well on these. Therefore, the high-latitude region has been taken into account for the training of the NN as well. Similar to previous work done with the ML-based calibration of the GRACE-FO satellite, a weighting was applied for the samples depending on the quasi-dipole latitude (QDLAT) they are originating from and based on the trustability of the reference model for these QDLATs. The low- and mid-latitude region samples have been weighted the highest, while the regions from 50° to 60° QDLAT have been weighted slightly lower and the high latitude regions beyond 60° QDLAT have been weighted the lowest as described in the following. The FACs appearing in the high-latitude regions were the main motivation for applying this weighting. The weights are calculated using the number of samples S_i of the three zones indexed by i in a relationship maintaining manner so that they must satisfy the following equation:

$$w_1 * S_1 + w_2 * S_2 + w_3 * S_3 = S_1 + S_2 + S_3. \tag{3}$$

Because this system of equations is undersatisfied with three weights, an additional ratio is given, set at $1 : \frac{1}{4} : \frac{1}{160}$, based on a rough estimate on the influence of FACs in the areas, so that samples from the second zone are weighted $\frac{1}{4}$ compared to the first zone, and samples from the third zone are weighted $\frac{1}{160}$ compared to the first zone. The reference model is mostly accurate for these regions as well, still, the samples containing FACs shall not be included with full weight as that would imply the model to calibrate these alongside the magnetometer calibration to meet the constraints imposed by the reference model, effectively setting the nulling of FACs as the goal of the modeling process. Thus, the final weights for the whole training dataset were set to about 1.68, 0.42, and 0.01 for the weights w_1, w_2, and w_3, respectively. When calculated for each month separately, they vary slightly from these values. The applied gradient generally is the result of the loss function that sums the equally weighted losses of batches, so only reducing the weights would lead to smaller gradients. A proportionally equally large gradient is achieved through rescaling of the weights.

Machine learning approach

Feed-forward NNs are general function approximators that are trained using the stochastic gradient descent algorithm (Hornik et al. 1989). This means that NNs can approximate any given function given enough complexity of the NN. The calibration of platform magnetometers can also be formulated as a function

$$f(x) = \hat{y}, \tag{4}$$

where x corresponds to our input including the 9 measurements of the 3 magnetometers as well as the previously processed housekeeping and telemetry data, f corresponds to the NN as the calibration function on our inputs, and $\hat{y}$ corresponds to the calibrated measurement by the platform magnetometers for the given input x. The function, or NN, is then trained using the mean squared error (MSE) which is the squared error between the predicted output $\hat{y}$ and the expected output derived from the CHAOS-7 reference model y:

$$\epsilon = (y - \hat{y})^2. \tag{5}$$

Utilizing this error ϵ, the gradients to adjust the weights of the NN are calculated and backpropagated through the different layers of the NN, thus optimizing for a low residual between the prediction of the calibration function and the expected output, reassembling a good calibration function. As there exists only one definitive position of the satellite for each measurement, there exists only one evaluation of the reference model for a given measurement. Therefore, the calibration function treats all three magnetometers as the input and outputs only one calibrated measurement. A detailed introduction to how NNs work can be found in Appendix A.

In addition, a lot of configurations for the training of the NN can be altered. For calculating the gradients during the training, a batch of data points is considered and a gradient for the whole batch is calculated for proceeding, the size of this batch is referred to as the batch size. The rate at which the calculated gradient is taken into account to adjust the weights of the NN is called the learning rate. While training the NN a step-decaying learning rate function has been chosen which reduces the learning rate after certain training progress. Also, there are a variety of activation functions to choose from as well as an arbitrary number of possible NN architectures, resulting from the number of neurons and layers used. Therefore, the architecture, as well as the configuration of the training parameters, needs to be explored which is done in the HPO.

Hyperparameter optimization

For the HPO, a variety of training and architecture configurations have been stochastically tested using the Bayesian Optimization algorithm (Snoek et al. 2012). The Bayesian Optimization algorithm is a Gaussian process that evaluates the trained NN with a certain parameterization and thus models the parametrization space connected with the target function as a Gaussian process. The search for an optimal set of parameters is then guided towards parameterizations of the NN which lead to an optimum of the target function, the smallest calibration residual, respectively. For the HPO, a variety of parameters and parameter ranges have been investigated, including the number and sizes of neuron layers within the NN architecture, the kind of activation functions used, the batch size, the number of training epochs, the learning rate, and the parametrization of the step-decaying learning rate function.

Since the model must be evaluated for each parametrization tried, a randomly chosen subset of the mission data, representing about 16% of the whole data or roughly 750000 data points, has been chosen for the HPO. This was mainly done because of performance reason as it reduces the training time of each evaluation, but still represents enough data for the NN to converge to an optimum and the data to reflect well the distribution of the entire data set. In addition, this HPO data are divided into two parts, 80% or about 600000 data points of the HPO data being used for the training while 20% of the HPO data are being used to test the trained NN. This is a common technique in ML since 80% of the HPO data is a good representation of the distribution of the data and enough data points for training the ML model while the remaining 20% of the HPO data are still a significant part of the data to test whether the model also performs well on unseen data. The error values of the trained NN will be compared between the training and test dataset to detect possible overfitting as well as the performance of the model. The score of the evaluation is then set to the MSE of the test part of the dataset.

The HPO resulted in the following parametrization of the NN: The learning rate was set to 0.01, the batch size to 1500, the number of epochs to 1200, the activation function as the Exponential Linear Unit (ELU) (Clevert et al. 2015), and the decaying function halving the learning rate every 90 epochs.

For the architecture of the NN, similar results have been suggested with a triangular-shaped architecture, consisting of a large first layer, a medium-sized second layer, and the output layer consisting of 3 neurons. Analysis has shown that with increasing layer sizes, the residual becomes smaller, but the further improvement was caused by reduced FAC amplitudes. Hence, a trade-off between a large architecture for a small residual and a small architecture that retains the FACs had to be found. Therefore, as a compromise, the architecture was set to consist of the first layer with 384 neurons, the second layer with 128, followed by the output layer with 3 neurons.

Fine-tuning and timeshift analysis

After using the found parametrization for the architecture and training of the NN, the global model was trained on all available data of the mission. Again, a split of 80% of the about 4.8 million data points for the training and 20% of the data points for testing has been used. Similar to before, this is a common step in ML to evaluate the model's quality by evaluating the model on unseen test data. In a second step, this global model was fine-tuned with the monthly data of the different months using following NN parameters: 200 epochs and a small learning rate of 0.00001, allowing only for small specific adjustments. Thus, a different calibration is found for each month which is derived from the same initial global calibration.

In addition to the fine-tuning of the model parameters, a possible time shift in parts of the data has been searched for, similar to the time shift found in the analytical calibration of the platform magnetometer of the GRACE-FO or GOCE mission (Stolle et al. 2021a; Michaelis et al. 2022). Therefore, in parallel to the fine-tuning of the global model, for every month the interpolation neuron presented by Styp-Rekowski et al. (2021) has been used to search for a possible time shift. This interpolation neuron automatically finds a time shift which reduces the overall residual of the calibration. The results were averaged over all months and are presented in Table 1. The overall mean absolute error (MAE) which is defined as the average absolute residual over all samples as well as the MSE defined as the squared residual over all samples are very close when comparing the training and test scores which is a good sign of the model fitting the data distribution well. For the magnetometers, a mean shift of −0.51s, and for the magnetorquers, a mean shift of −3.61s have been found. But, the STD for both shifts found is very high, with 1.79s and 2.45s, respectively, rendering these time shifts very inconsistent across all months. As the time shifts cannot consistently be found during the fine-tuning of independent monthly data, they are considered to represent local minima, and thus they are not incorporated into the final solution, leaving the data as is. Also, the time shift of 0.4 s as found by Michaelis et al. (2022) has been evaluated with no observable improvements.

Final architecture

To sum up, the final approach uses the parametrization found in the HPO and applies no time shift. The carefully preprocessed data are thus input into the NN with 3 layers of 384 neurons, 128, and 3 neurons as shown in Fig. 2. As depicted in Fig. 1, the data are first collected, filtered, and preprocessed to then train the global model for 1200 epochs as described in the HPO chapter. Finally, the model is fine-tuned month-wise without applying a time shift with a lower learning rate of 0.0001 for 200 epochs. These trained monthly models are then used to generate the calibrated dataset of the platform magnetometers of the GOCE satellite.

Results and discussion

After producing the calibrated dataset, the resulting measurements of the Earth's magnetic field have been analyzed. First, the remaining residual of the calibrated data compared to a global magnetic field model is evaluated, followed by different magnetic phenomena like the FACs and magnetic storm behavior. Finally, the calibrated data have been used to support the modeling of the magnetic field of the Earth.

Residual to reference model

Table 2 shows the residuals of the calibrated dataset against the reference model CHAOS-7. Unless otherwise

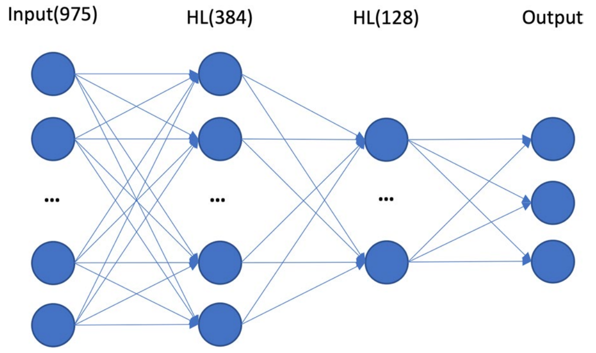


Fig. 2 Neural network architecture, with the inputs on the left side, two hidden layers (HL) and the 3-dimensional output standing for the 3-axis calibrated measurement

Table 1 Summarized residual and offset data over all available months for low- and mid-latitudes. Errors and standard deviation given in nT, while the offsets for the magnetometers (MAG) and magnetorquers (MTQ) are given in seconds

	MAE		MSE		StdDev		MAG	MTQ
	Train	Test	Train	Test	Train	Test		
Mean	13.42	13.55	854.95	882.26	28.32	28.57	− 0.51	− 3.61
Standard Dev.	3.57	3.71	464.58	552.58	7.34	8.20	1.79	2.45

Table 2 Error values in nT after fine-tuning training of the neural network for every month, on the left for the whole latitude region, on the right for low- and mid-latitudes

Month	**90-QDLAT**		**50-QDLAT**	
	MAE	**STD**	**MAE**	**STD**
	Train | Test	**Train | Test**		
200911	8.98 | 8.89	20.08 | 19.40	4.63	5.81
200912	8.88 | 8.85	19.31 | 19.07	4.61	5.78
201001	9.96 | 9.94	22.29 | 22.27	4.68	5.94
201002	10.48 | 10.38	23.50 | 23.72	4.67	6.08
201003	11.12 | 11.11	24.29 | 24.47	5.37	10.10
201004	12.05 | 12.21	26.51 | 26.75	5.37	6.88
201005	16.38 | 16.60	33.34 | 34.56	8.38	10.36
201006	17.14 | 16.93	35.20 | 34.08	8.54	10.50
201007	20.01 | 20.67	39.74 | 41.52	9.62	13.17
201009	9.94 | 10.53	20.53 | 22.34	5.56	9.60
201010	10.45 | 10.44	23.38 | 23.02	5.00	6.37
201011	10.52 | 10.45	23.24 | 22.84	4.86	6.15
201012	10.36 | 10.45	22.86 | 23.16	4.87	6.14
201101	11.02 | 10.97	24.16 | 23.76	5.38	9.95
201102	10.91 | 10.99	24.38 | 25.60	4.95	6.24
201103	10.96 | 10.99	25.20 | 25.24	4.85	6.14
201104	12.76 | 12.80	29.19 | 29.39	5.54	7.12
201105	15.58 | 15.37	34.06 | 33.09	7.04	8.90
201106	19.29 | 19.22	40.96 | 40.99	9.12	11.12
201107	18.70 | 19.09	40.11 | 41.03	8.74	10.71
201109	11.05 | 10.92	23.64 | 22.72	5.20	6.73
201110	10.82 | 11.02	23.34 | 23.78	5.15	6.63
201111	13.15 | 12.86	29.03 | 28.23	6.41	9.16
201112	11.35 | 11.41	24.51 | 24.87	5.75	7.49
201201	13.66 | 13.67	30.82 | 31.41	5.73	7.38
201202	11.66 | 11.54	25.82 | 25.59	5.09	6.46
201203	14.09 | 14.03	32.18 | 31.75	5.30	6.78
201204	14.35 | 14.45	33.20 | 33.20	5.48	7.07
201205	17.27 | 17.44	37.74 | 38.03	7.69	9.68
201206	18.38 | 18.21	39.55 | 38.63	9.24	11.57
201207	21.84 | 22.38	49.22 | 49.70	10.85	15.88
201208	15.88 | 16.09	36.26 | 36.56	6.22	8.00
201209	12.52 | 12.57	28.39 | 28.27	5.68	7.95
201210	11.29 | 11.13	24.42 | 23.38	5.48	7.09
201211	11.98 | 11.98	26.56 | 26.80	5.51	7.49
201212	11.50 | 11.57	26.13 | 26.00	5.42	6.91
201301	10.72 | 10.92	22.81 | 23.40	5.49	7.14
201302	24.03 | 23.90	53.19 | 52.30	14.99	43.09
201303	10.87 | 10.85	23.34 | 23.26	5.29	6.84
201304	12.47 | 12.43	27.60 | 27.10	5.68	7.29
201305	18.18 | 18.67	39.63 | 40.27	8.09	11.03
201306	18.69 | 18.82	39.99 | 39.49	9.20	11.33
201307	18.74 | 18.90	40.56 | 40.99	8.73	10.88
201308	15.16 | 15.67	35.63 | 37.21	6.01	8.71
201309	11.97 | 11.78	26.19 | 25.13	5.70	7.51

noted the residuals to the CHAOS-7 predictions of core, lithosphere and large-scale magnetospheric field are considered in the following sections. Each row shows the average residual for every month with the same data selection as described in the previous chapter. The left side shows the training results for the whole latitude range of the Earth, while the right side shows the low- and mid-latitudes between the 50° QDLAT. The residual is given in mean absolute error (MAE), mean squared error (MSE), and STD. As mentioned earlier, for the training the data have been randomly split into 80% of the data representing the training data, and the remaining 20% representing the test data. For the training results on the left side, the residual is always given for the training and test data.

It can be seen that the generalization error, meaning the gap between the training and test dataset is very small, indicating that the model performs similarly on unseen data. This means that the model was able to adapt to the statistics of the data in a way that holds true for unseen data. Naturally, it can be observed that the residual for the low- and mid-latitudes is smaller than for the whole latitude range as the highly fluctuating FACs are measured by the satellite in this range while not being included in the reference model, thus resulting in a higher error. For the low- and mid-latitude, a consistent residual of below 10 nT can be observed which enables scientific studies. A trend is visible, that around June solstice, in the months of May, June, and July the residuals with a range of 8-10 nT seem to be higher than for the majority of the other months with 4-6 nT, which seems to be related to a higher STD and will be discussed later. The February of 2013 appears to be an outlier.

Residual maps

In Fig. 3, the residuals of the calibrated platform magnetometer measurements against the CHAOS-7 reference model are plotted as a function of latitude and longitude for the North, East and Center B-field components on global maps. The data shown are annual averages for 2012, which have been determined by averaging all data from 2012 in bins of 4° latitude by 4° longitude. The averages have been assigned a color according to the scale in the bottom right, plotted as a contour plot.

It can be seen that the calibration performs well, especially in low- and mid-latitudes where grey is the dominating color, meaning that the mean remaining residual is close to 0. In high latitudes, the shapes of the high residuals clearly resemble the known patterns of FACs, which are not included in the reference model, thus indicating that this signal remains in the calibrated satellite measurements as desired. In the Center component, there is an anomaly visible along the magnetic equator. This remaining residual is rather small in amplitude and is suspected to be of artificial origin, as it was also observed in the calibration of the platform magnetometers of GRACE-FO and Cryosat-2, which use the same type of instrument. For the GOCE satellite, Michaelis et al. (2022) found the same anomaly which remained visible although a multitude of features have been used with the proposed approach.

Comparison of orbit residual

In Figures 4, 5 and 6 the residual of the calibration against the reference model CHAOS-7 is plotted as a function of the QDLAT. Each figure shows exemplary either the ascending or descending orbits of one axis of the platform magnetometer measurements for the whole year of 2012, separated by months starting on top with January to December on the bottom. Each drawn line represents the calibrated values of one corresponding half-orbit within that month, while the mean measured value for every QDLAT is depicted in black. It can be seen that the calibration performs especially well in the low- and mid-latitude region while in the high-latitude regions the remaining higher residuals mainly are due to the FACs that are not modeled in the CHAOS-7 model but measured by the satellite. For the southern hemisphere, it can be observed that the FACs are being correlated with features by the NN model as they are lower in amplitude and in their mean value than expected. Currently, it is subject of further research to detect the features which were used by the NN model to correct for the FACs as this occurs only in the southern hemisphere.

The higher STD in the northern hemisphere summer months is evident in a larger spread of individual orbit residuals (colored areas in the plots) for the June solstice in the months of May, June, and July, present in all three components and both orbital directions, supporting the assumption that the measurements contain higher noise as there is no systematic pattern detectable. The origin of the higher deviation is unknown. Although the complete set of housekeeping and telemetry data was used in the calibration process, that gives a multitude of information about the satellite system, there was no feature able to correct this remaining error. Thus this is subject to further investigation. Overall, the calibration works well as the residual is generally low as well as the STD, visible as the colored areas in the figures.

The remaining, artificial residual around the magnetic equator in the Center component mentioned above is found here too in both the ascending and descending orbits, and with a sinusoidal shape around the zero crossing of this magnetic component. For the East component, there is an increasing trend visible

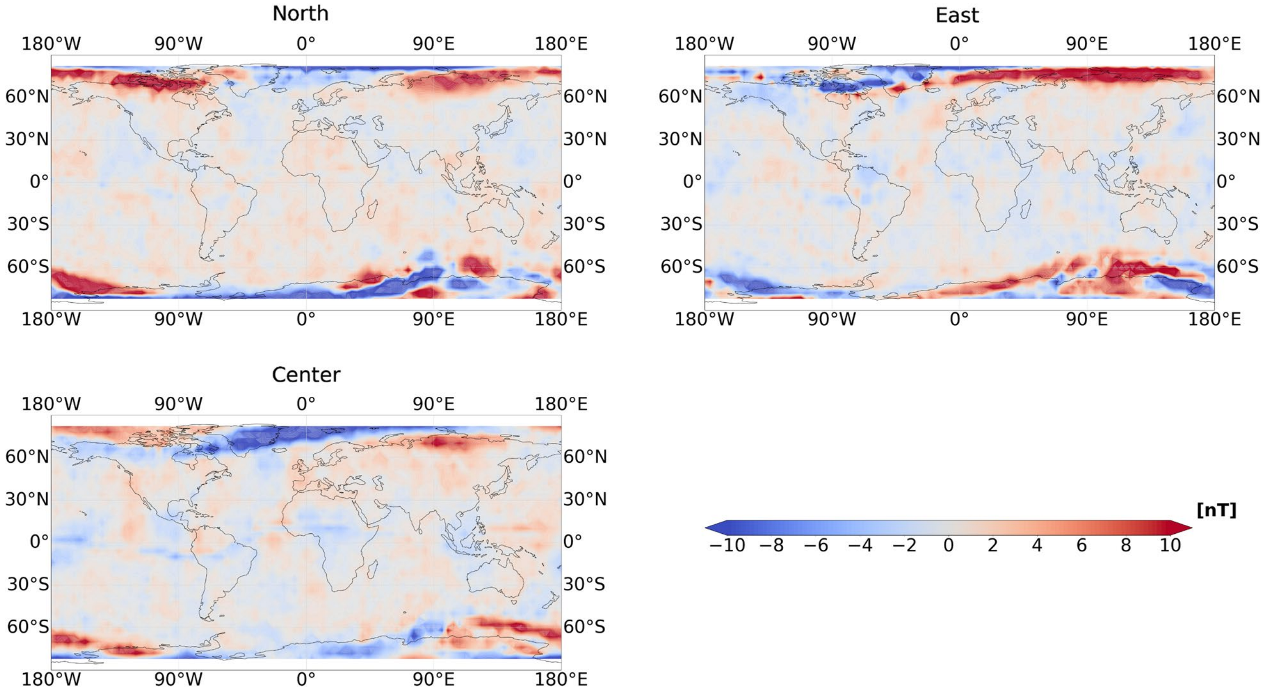


Fig. 3 Binned and averaged residual map between the calibrated platform magnetometers and the CHAOS-7 reference model for the year 2012 in the NEC Frame, North component in top left, East in top right, Center in bottom left. The color scale for all panels is given in the bottom right, and saturation of color is achieved beyond -10 or 10 nT, respectively. Residuals are low while FAC structures remain visible

around the equatorial region where the variation of the residual increases, especially for the descending orbits. This increased variation has been observed to increase throughout the duration of the mission from 2009 until 2013. It is suspected to be correlated to the solar cycle which had its minimum in December 2008 and then increased up to its maximum in April 2014.

Comparison of analytical and ML method

The calibrated dataset presented in this study is compared to the analytically calibrated dataset from Michaelis et al. (2022), in the following referred to as the Ana approach. Therefore, the Common Data Format (CDF) product files have been utilized to compare the residuals as well as differences between the calibrations. In addition, the proposed dataset was used to similarly compare against a magnetic storm event.

Residual distribution

First, the residual distribution of the ML approach with the CHAOS-7 reference model is compared to the residual distribution of the Ana approach with the CHAOS-7 reference model in Fig. 7, this was done on the December 2009 data that represent the most magnetically quiet month. The three panels show the North, East and Center component residuals of the Ana distribution in blue and the residuals of the ML distribution in orange. It can be seen that all distributions approximately follow a Gaussian distribution curve where the ML approach has a smaller standard deviation, manifesting in a steeper Gaussian distribution. This is especially the case for the North and East component where the lower overall residual of the ML approach appears to originate. The Center component looks very similar in both approaches. In addition, all distributions are centered near zero. For the East component there appears to be a small shift to the left which is suspected to be originated in the different treatment of the FACs in the southern hemisphere when comparing both approaches.

Figure 8 shows the difference between the ML and Ana calibration as a function of the calibrated measurements of the Ana approach, separated for the three North, East, and Center components, again for the most magnetically quiet December 2009, please note the different scales. The two calibrations mainly differ in the FAC regions that lie in the high-latitude regions, this means up to a value of 10000nT for the North component, the whole value range for the East component, and below about -40000nT and above about 40000nT for the Center component. We note that the difference is larger in the

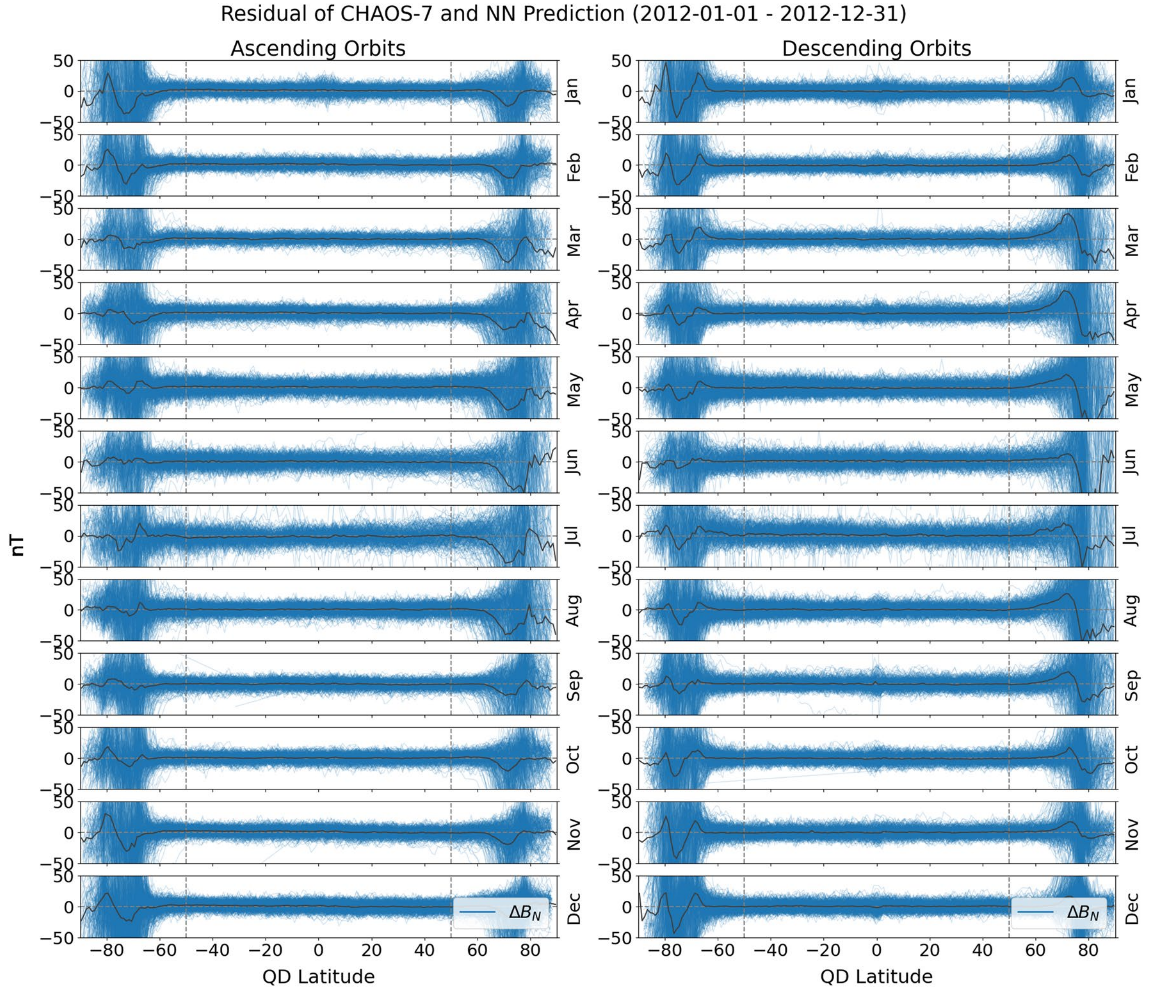


Fig. 4 Residual between the calibrated platform magnetometers and the CHAOS-7 reference model as a function of the quasi-dipole latitude (QDLAT) for every month of the year 2012 of the North component of the calibrated measurement in nT. The orbits are split into the ascending dusk-orbits on the left, and the descending dawn-orbits on the right, with the mean residual depicted in black

southern hemisphere (negative Center component), which is related to underestimation of FAC amplitudes in the ML method in this hemisphere described above that is currently under investigation. In addition, the regression curves show that there is no systematic difference in dependence of the value ranges between the two calibrations as the slope of the function is approximately zero. The shift indicates a difference between the two calibrations that lies mostly in the bit resolution of the measurement, which is about 3.05nT, being defined by the the last bit of the magnetometer measurement.

Comparison to Dst index

Figure 9 shows a magnetic storm in March 2013 as characterized by the Dst index that is obtained from the data of four low latitude ground magnetic observatories (Nose et al. 2015). The deviation of the horizontal component dB_H of the calibrated measurements from the CHAOS-7 core and lithospheric field model is plotted in orange and blue for the ascending and descending orbits, respectively. As the GOCE satellite flies in dusk–dawn orbits, the ascending measurements correspond to an MLT of about 18 while the descending measurements correspond to an MLT of about 6. It can be seen that the measured power follows very well the Dst index, thus showing the capability of the GOCE satellite to measure magnetic storm events. In addition, the ascending measurements are deviating stronger from the origin and the Dst index which is an expected phenomenon for the dusk side measurements, also known as the dawn–dusk asymmetry described by Anderson et al. (Anderson et al. 2005). The dusk side of the magnetic field is showing stronger

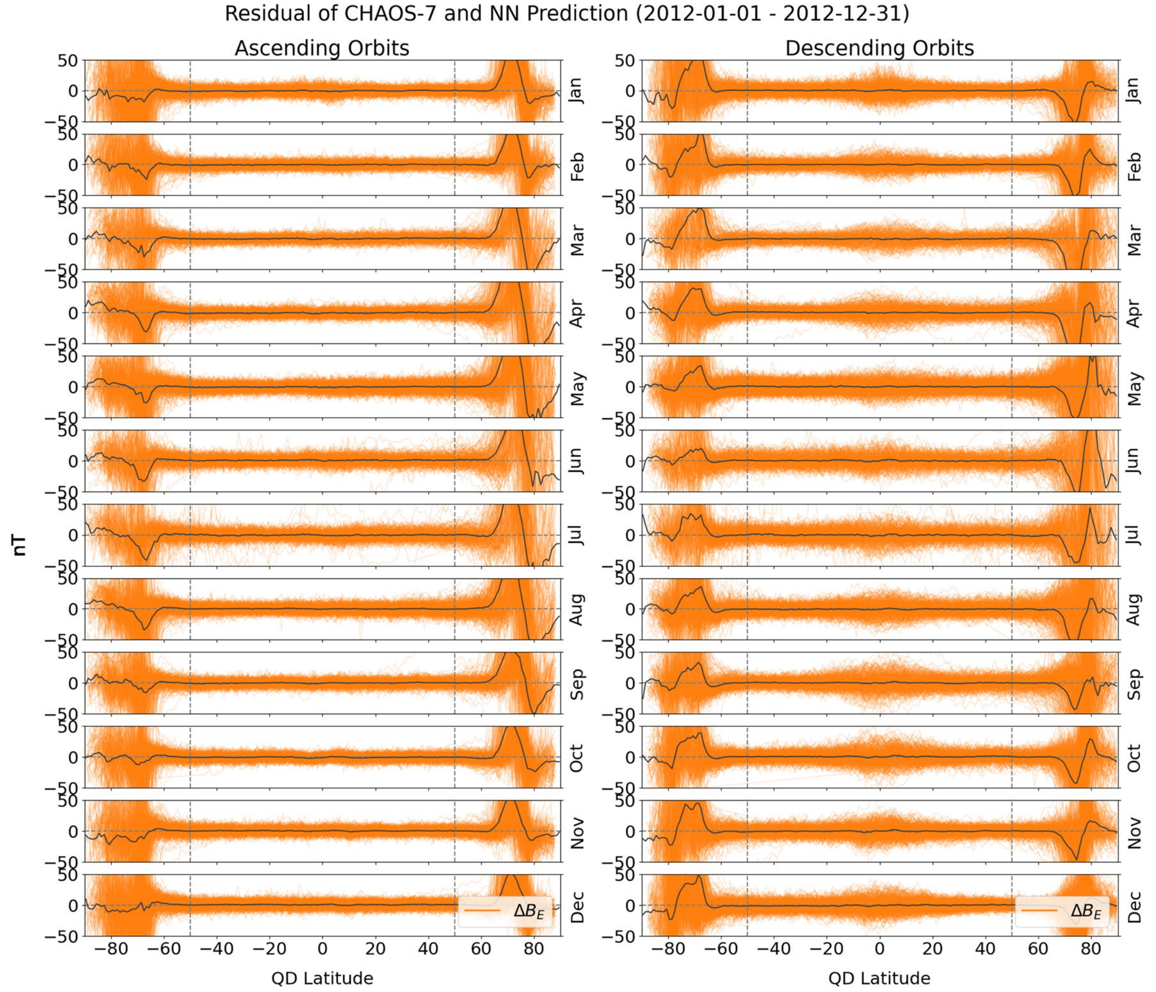


Fig. 5 Residual between the calibrated platform magnetometers and the CHAOS-7 reference model as a function of the quasi-dipole latitude (QDLAT) for every month of the year 2012 of the East component of the calibrated measurement in nT. The orbits are split into the ascending dusk-orbits on the left, and the descending dawn-orbits on the right, with the mean residual depicted in black

deviations during a magnetic storm. Similar behavior was observed for other storms during the lifetime of the GOCE mission. Similar results were obtained by Michaelis et al. (2022), confirming the quality of the calibration obtained, with even clearer separation by dusk and dawn.

Lithospheric field analysis

With an altitude of about 255km, the GOCE satellite has a rather low altitude compared to other magnetic satellite missions. Therefore, an analysis of the measured lithospheric field has been conducted.

As the data have a timely resolution of 16 s, the whole mission period has been taken into account for extracting the lithospheric field. The lithospheric field is known to not change within a time frame of about 4 years. Thébault et al (2021) recently proposed a high-resolution lithospheric field model which is depicted in the middle panel of Fig. 10. The top panel shows the lithospheric field as given by the ML-calibrated measurements of the GOCE satellite, after subtraction of the CHAOS-7 core and large-scale magnetospheric field. In general, a good agreement can be seen between the measured data of the GOCE satellite and the lithospheric field model. Anomalies like the Bangui in Africa, or the Kursk anomaly in Russia are detectable, as well as the striped ocean bottom structures in the Atlantic. The third panel shows the difference between the high-resolution model and the reconstructed lithospheric field from the GOCE measurements which, despite the small scale, shows nearly no saturation and has pale colors. This suggests that by incorporating GOCE data into lithospheric

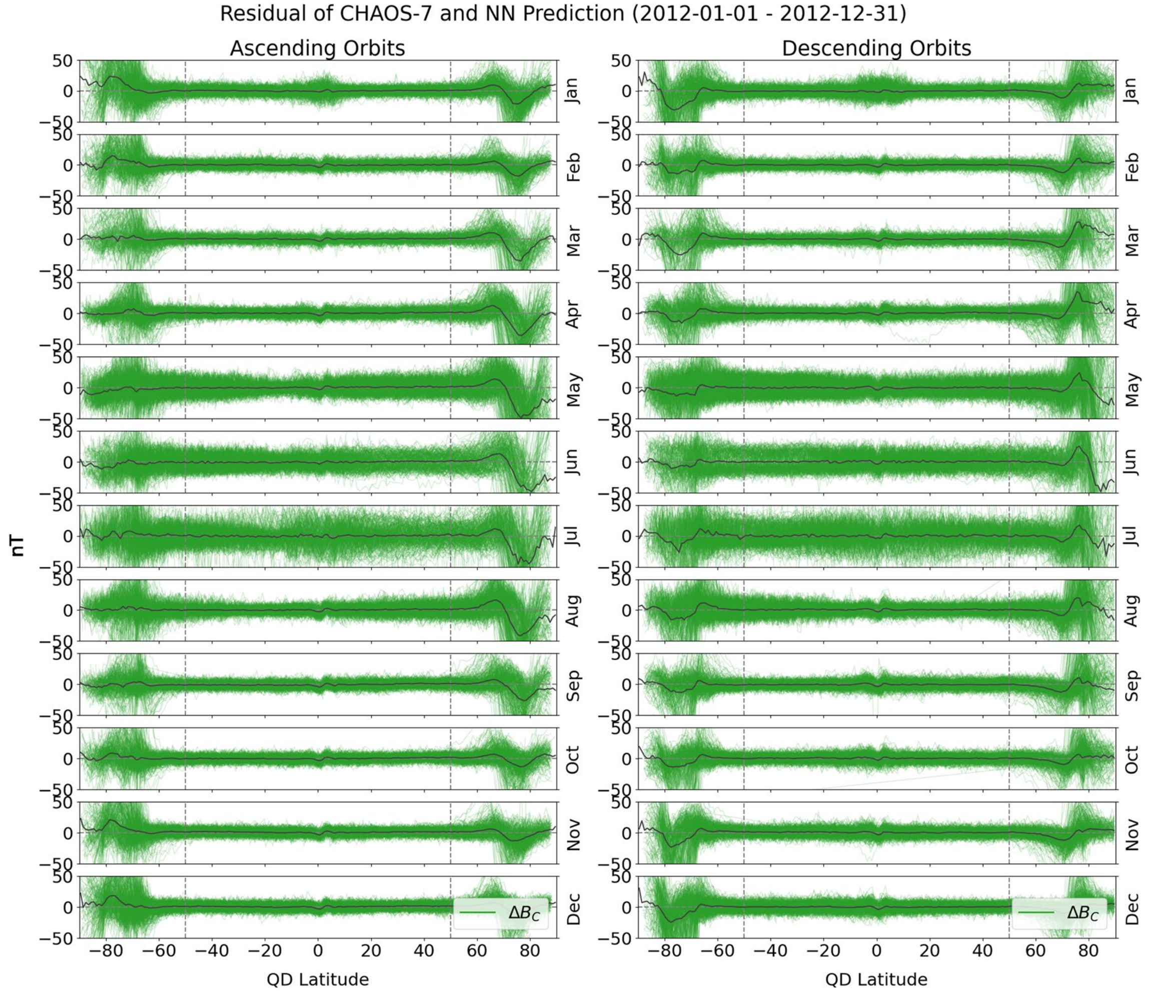


Fig. 6 Residual between the calibrated platform magnetometers and the CHAOS-7 reference model as a function of the quasi-dipole latitude (QDLAT) for every month of the year 2012 of the Center component of the calibrated measurement in nT. The orbits are split into the ascending dusk-orbits on the left, and the descending dawn-orbits on the right, with the mean residual depicted in black

field models, there is a potential improvement due to the additional amount of data points, filling the gap between CHAMP and Swarm.

Comparison to CHAMP

As the CHAMP mission ended in September 2010, there is a period of time during which both satellite missions flew simultaneously. Therefore, conjunctions have been calculated where the two missions have been close to each other to compare the retrieved measurements.

Figure 11 shows the found conjunctions between the 23rd of January 2010 and the 5th of February 2010. During this interval, the two satellite missions have been co-rotating and the magnetic environment was quiet as indicated by the first panel. The distance between the two missions was below 1000km and the relevant QDLATs are shown in the 3rd panel, the dawn and dusk orbit measurements are depicted in red and blue, respectively. The three bottom panels show the residual between the measurement of the GOCE and the CHAMP satellite for the three components of the magnetic field in the NEC frame. Most data points have an absolute difference of less than 20nT, while there is no systematic difference apparent.

Table 3 puts the results of the conjunctions in perspective. For the three components, the total count of measurements during the conjunction is shown, separately for the dawn and dusk orbit. The number of data points that lie above or below the 10nT mark is given in the fourth and fifth column. It can be seen that overall 80% of the

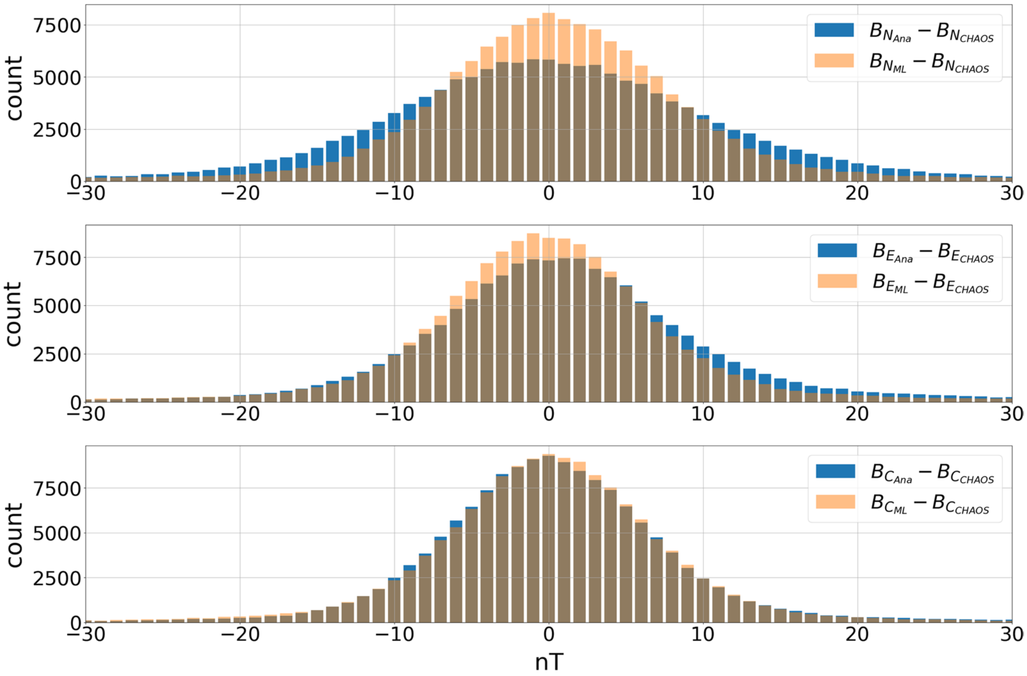


Fig. 7 Residual distribution comparison of the analytical (blue) and ML (orange) based calibrations compared to the CHAOS7 reference model for the December 2009 within a histogram plot with bin sizes of 1nT for the magnetic North (top), East (middle) and Center (bottom) component

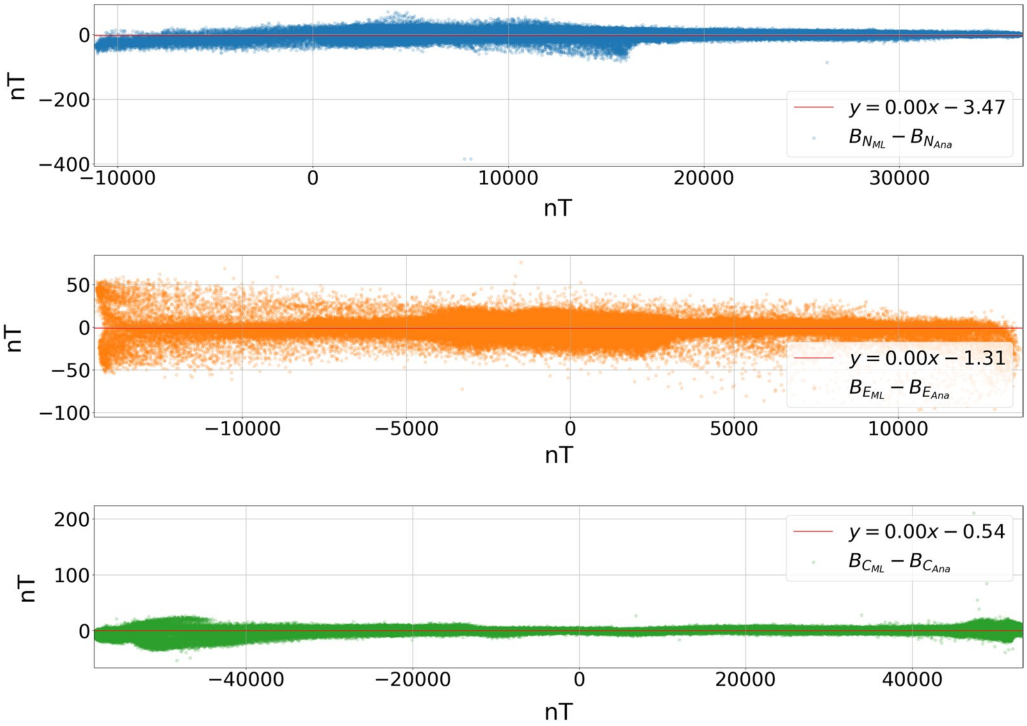


Fig. 8 Difference between ML and Ana calibrations as a function of the calibrated measured magnetic flux density by the analytical approach for the December of 2009 for the North (top), East (middle) and Center (bottom) component. The regression curves are given in red. Please note the different scales of the y-axis

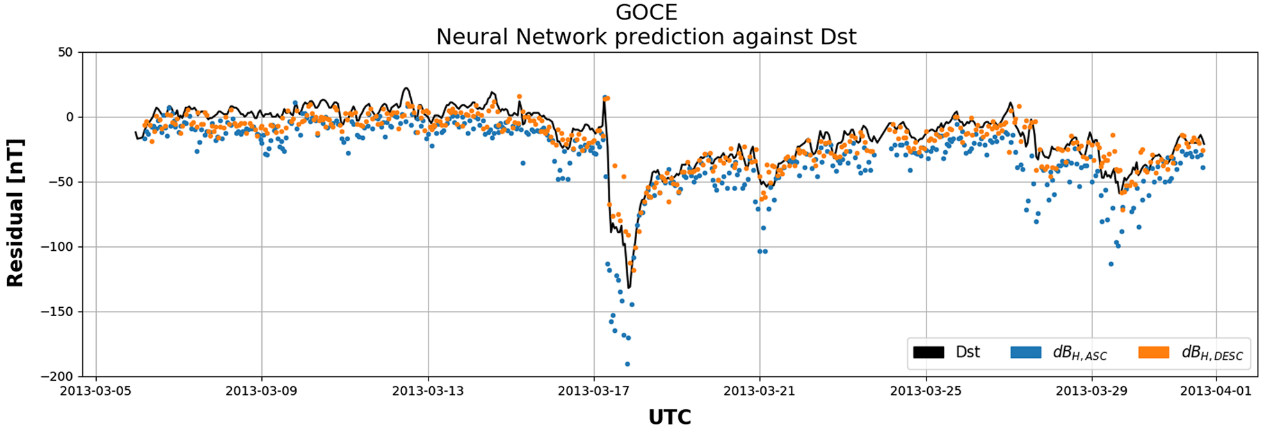


Fig. 9 Magnetic storm in March 2013 indicated by Dst index. The variations of the residual values of the calibrated GOCE platform magnetometer to the CHAOS-7 core and lithospheric field for the ascending, $dB_{H,ASC}$ (blue) and descending, $dB_{H,DESC}$ (orange) orbits follow the index line well, with some known offsets

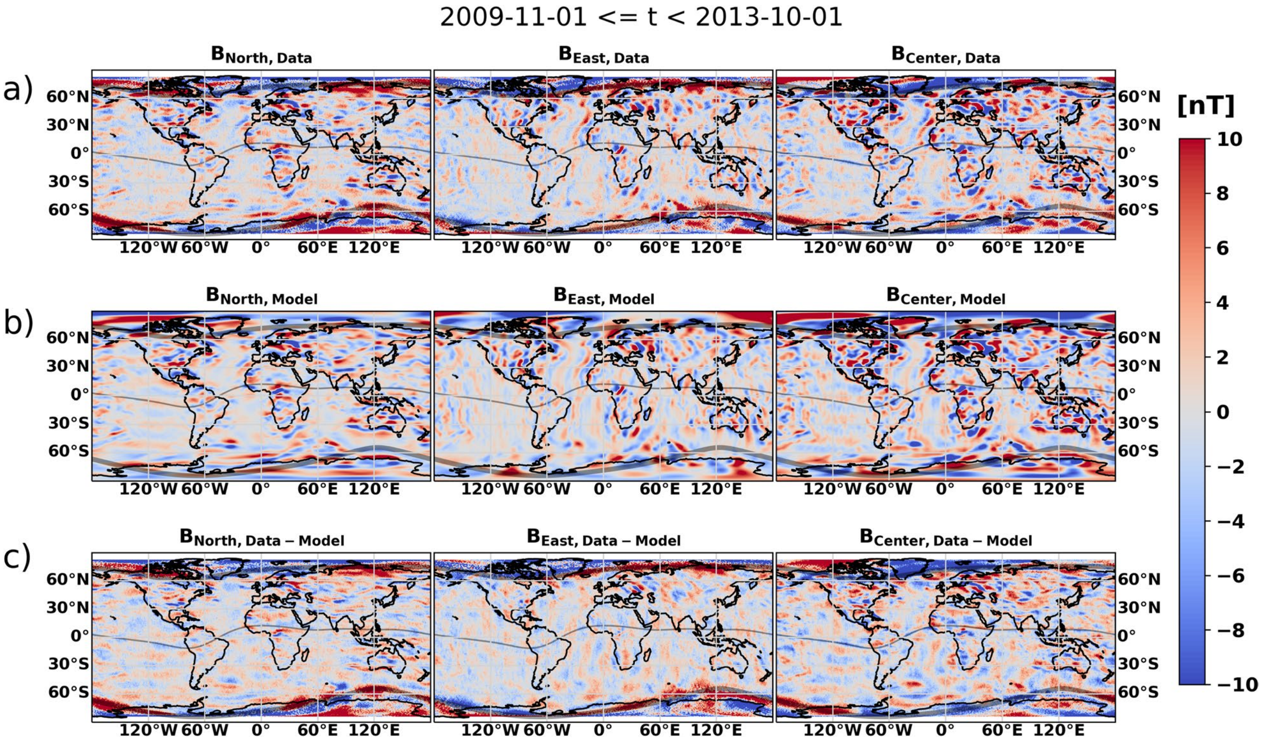


Fig. 10 Lithospheric field residual of GOCE data after subtraction of CHAOS-7 core and large-scale magnetospheric field components (top) and high-resolution lithospheric field model calculated from Thébault et al (2021) (middle). The bottom panel shows the difference between GOCE lithospheric field data and the high-resolution lithospheric model

measurements during the conjunction have a low residual compared with the CHAMP mission, although this varies depending on the component. This is an encouraging result as this result was achieved using a platform magnetometer.

Impact analysis

An interpretation analysis of the NN and the impact of the features has been conducted. Therefore, SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) have been used. This method is able to compute the

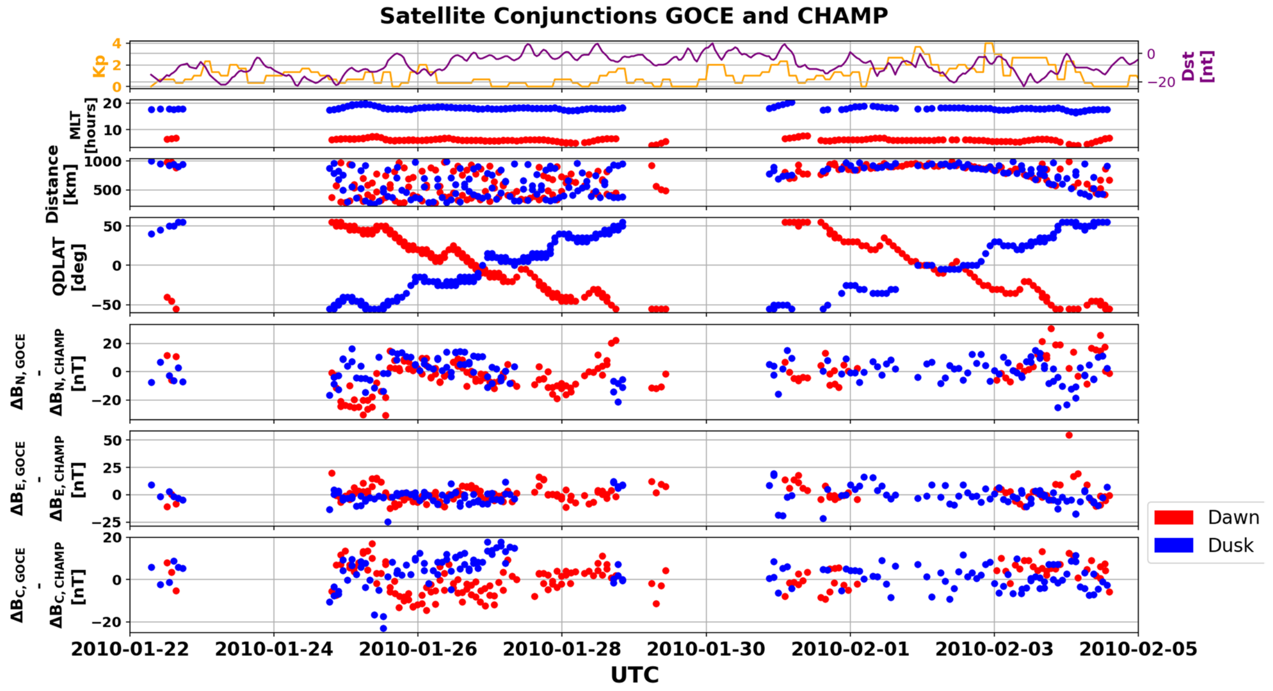


Fig. 11 Conjunctions for co-rotating GOCE and CHAMP satellites for the period of 2010-01-22 to 2010-02-05 are shown. The first panel shows the geomagnetic environment, indicators Kp and Dst show the geomagnetic quiet time. The second panel shows the MLT with descending dawn orbits in red and ascending dusk orbits in blue given by quasi-dipole latitude (QDLAT) in the 4th panel. The 3rd panel shows the distance between the GOCE and CHAMP satellites. The three bottom panels show residuals for GOCE with respect to the CHAOS7 reference model subtracted by residuals for CHAMP with respect to the CHAOS7 model for the North, East and Center components from top to bottom in nT

Table 3 Statistics of conjunction between the GOCE and CHAMP satellites corresponding to Fig. 11. The columns show different statistics for the dawn and dusk orbits for each NEC component for the residual between the satellites for distances within 1000 km, QDLAT <60° and split for below and above a residual of 10 nT

Component	Orbit	Total count	Count >= 10nT	Count < 10nT	Percentage < 10nT
North	Dawn	155	52	103	66.45
North	Dusk	148	41	107	72.30
East	Dawn	155	25	130	83.87
East	Dusk	148	20	128	86.49
Center	Dawn	155	24	131	84.52
Center	Dusk	148	23	125	84.46

contributions of each feature to the prediction of the NN. To achieve this, December 2009 is chosen as the example because it is the most magnetically quiet month. 600 data points are used to compute a mean prediction which is used as the background by the SHAP method. These data points are determined as 600 centroids of K-Means clustering, so a good representation of the monthly data is achieved. This step defines the mean predictions of the NN, which is needed so that the SHAP method can calculate the contributions of the different features to deviations from this mean expected result. In a second step, 600 randomly chosen data points and their prediction have been compared multiple times with the average result. For each data point, this step is repeated many times with masking some features absent, thus an estimate about the contribution of each feature can be given. As the calibration was done in satellite frame, this analysis has also been conducted in satellite frame.

Figure 12 shows the top 40 features used by the ML algorithm, sorted depending on their impact or contribution on the final prediction for the calibrated value. The list of features was split in two parts, depending on the scale of their contribution, please note the different scales. The most important feature can be found on the top left, while the 40th most important feature can be found in the bottom right. The order is based on the average impact on the model output, which consists of the added contributions to the X, Y, and Z components, depicted in different colors. The ESA provides an online sheet containing the feature abbreviations and their descriptions (ESA 2019). On the left, the magnetometer measurements can be found, e.g., the first three features are the Z-component of the three different magnetometers, which are accompanied by the same features found in the Telemetry data, AMT00104 to AMT00304. This confirms the assumption that the magnetometers have the largest influence on the magnetometer calibration. On the right, the most important housekeeping or telemetry features can be seen which are: temperatures (THT), Xenon Tank Heaters (XST), one Euler angle, some currents (PHD), and the magnetorquers (ATT). These are interesting findings as these generally correspond to the expectations for the calibration of platform magnetometers as used in the analytical method, but were found automatically by the ML approach. There are some differences in which features exactly are relevant when comparing both approaches, e.g., the Xenon Tank Heaters or different temperatures which were not used in the analytical approach. These have not been searched for manually in the previous analytical approach and show the ability of the ML approach to discover relevant features.

Integration and improvement in Kalmag model

To evaluate the dataset in the context of geomagnetic field modeling, it was assimilated by the Kalman filter algorithm used to build the Kalmag model (Baerenzung et al. 2020). This model is composed of 7 sources, a core field, a lithospheric field, an induced/residual ionospheric field, a remote, a close, and a fluctuating magnetospheric field, and a source associated with FACs. Each source is expanded in spherical harmonics and their dynamical evolution is controlled by scale-dependent autoregressive processes. For more detail about the Kalman filter algorithm and the spatiotemporal characterization of the different sources see Baerenzung et al. (2020).

The Kalmag model spanning the last 22 years was constructed through the assimilation of CHAMP, Swarm, and secular variation data derived from ground-based observatory measurements. It therefore is subject to the lack of observations taken by low orbiting satellites between 2010.7 and 2013.9. This new GOCE dataset could therefore have a great potential to fill this data gap. In order to evaluate this potential, three models were built, two including GOCE data and one without. The first model, referred as model C, is the one partially serving the construction of Kalmag. It spans the [2000.5; 2014.0] time period and is derived from CHAMP and observatory secular variation data. Its solution in 2009.8 was considered as a prior information for the second model (model G), which was built through the assimilation of GOCE and secular variation observations. The last model, namely model GL, is similar to model G except that the mean lithospheric field was a priori set to zero and its prior covariance initialized. Its purpose is to evaluate how the lithospheric field can be recovered from GOCE data alone. All comparisons between the different models are performed for 2014.0.

The results of this evaluation for the internal field, i.e., the sum of the core and the lithospheric field, and for the secular variation, are presented in terms of energy spectra at the Earth's surface in 2014.0 in Fig. 13. Spectra of the mean (solid lines), the standard deviation (dashed lines) and the difference with the Kalmag mean model (circles) are displayed for model G with thick lines and for model C with thin lines. For the main field (left panel), the spectra associated with mean solutions cannot be distinguished between the two models. This is not the case for the spectra of the difference with Kalmag. The mean solution of model G is globally closer to the Kalmag solution than the mean solution of model C. Since Kalmag is more accurate than both model G and C, due to the fact that it also derives from Swarm data taken before and after 2014.0, this result demonstrates that the assimilation of GOCE data helps to better resolve the main field. However, predicted uncertainties of model G, given by the spectra of the standard deviation, are slightly underestimated when compared to effective errors, as approximated by the spectra of the difference with Kalmag. This is an indication that some source contributing to the observations is not perfectly modeled. Ionospheric currents, which are still generating some magnetic signal at dawn and dusk (the orbit of the satellite), might be this source since the Kalman filter algorithm was not calibrated to account for them.

The impact of GOCE data to recover the secular variation is clearly positive as it can be observed on the right panel of Fig. 13. Not only the spatial resolution of model G is higher, but its level of error is globally lower than for model C. Furthermore, predicted and effective errors are consistent with one another. Contrary to the main field, no signal is apparently contaminating the secular variation. Note that performing the same analysis at previous epochs leads to similar results.

(See figure on next page.)

Fig. 12 The figure shows the top 40 features used by the ML algorithm, sorted by their average impact on the final prediction for the calibrated value, as was calculated by the SHAP method. This was evaluated for the X, Y, and Z components of the measurements in the satellite frame where each feature's contribution consists of the contribution to each component, distinguished by color. Note the different scale of the features on the left and right

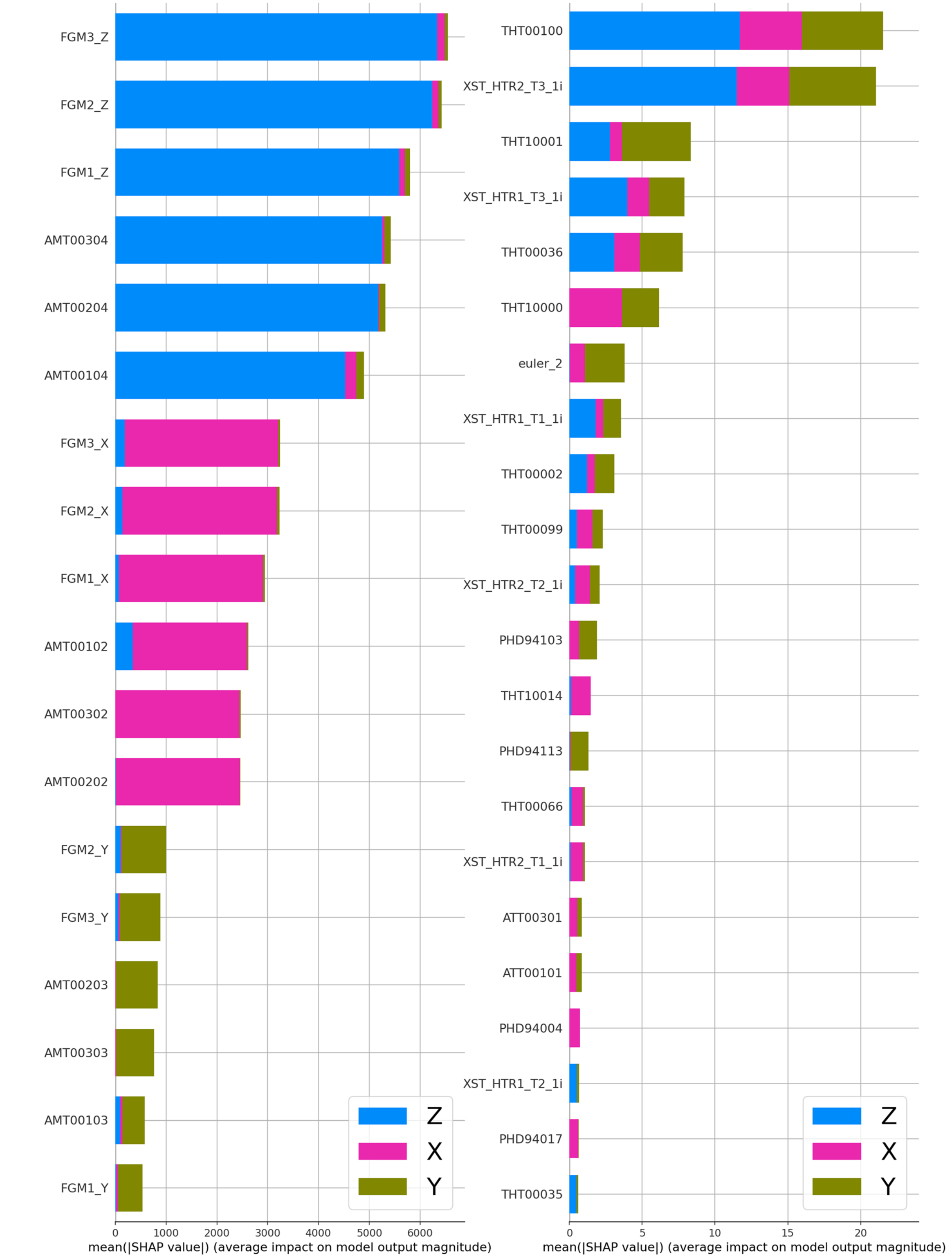


Fig. 12 (See legend on previous page.)

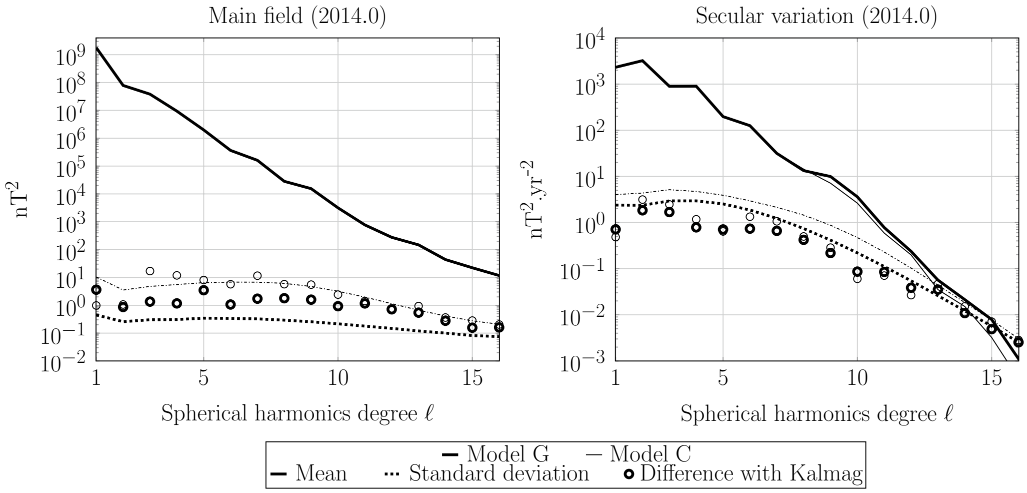


Fig. 13 Main field (left) and secular variation (right) energy spectra at the Earth's surface in 2014.0. Spectra of the mean (solid lines), standard deviation (dashed lines) and difference with the Kalmag model (circles) for model G (thick lines and symbols) and model C (thin lines and symbols)

The signal associated with the lithospheric field can also be well extracted from the dataset as illustrated in Fig. 14. Comparisons between the downward component of the field at the Earth's surface for model GL (top) and for model C (bottom) highlight their proximity. Most structures which can be recovered with CHAMP data are present within the GL solution. However, as for the core field, predicted uncertainties are also slightly underestimated (not shown).

Through this evaluation one can conclude that such a dataset is useful for the geomagnetic modeling community to cope with the lack of observations between the CHAMP and Swarm eras. Nevertheless, efforts in modeling the dayside ionospheric field are likely to be required in order to take full advantage of these new measurements.

Conclusion

To sum up, we could show that the ML-based calibration of the platform magnetometers onboard the GOCE mission yields promising results. With careful data collection and selection, as well as sophisticated data preprocessing it is possible to significantly reduce the remaining residual of the platform magnetometer measurements compared to the reference model CHAOS-7. Our evaluation has shown that on average a residual of 6.47nT for low- and mid-latitudes could be achieved which leads to a dataset that can help in studying the Earth's magnetic field. Some potential applications were shown like measurements of the lithospheric field, as well as additional information during geomagnetic storms. During a conjunction with the CHAMP satellite, it could be shown that the achieved calibration is in good agreement with other magnetic measurements. Finally, the enhancement of an existing magnetic field model could be shown for the time span of the gap of high-precision magnetic missions between 2010 and 2013. With the ML approach we are able to provide a calibration which is easily applicable to other satellite missions with some knowledge about the underlying Data Science techniques, whereas the previous analytical approach needs deep Domain knowledge to be applied. In the future, we hope that the provided dataset can support geoscientists by offering additional data to better cover the magnetic field in time, altitude, position, and MLT. The data of this study are published on the ISDC-Server of the GFZ at (Styp-Rekowski et al. 2022) ftp://isdcftp.gfz-potsdam.de/platmag/MAGNETIC_FIELD/GOCE/ML/v0204/ under the version 0204. In the future, the used features of the calibration as well as the correlation between features and the southpolar electrojet will be analyzed to investigate how to improve the calibration further.

Appendix A: Neural network introduction

A feed-forward neural network (NN) consists of multiple neurons where each neuron's output y is defined as follows:

$$y = a(w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + \cdots + w_n * x_n + b)$$

where x_1 to x_n are different features of a data point which get multiplied by their respective weight w_1 to w_n, with a bias b being added finally. The sum of these products is then fed into an activation function a which is explained in the next paragraph. NNs consist of many such neurons which will be assigned different weights. Multiple neurons in parallel, which means that they work on the same input x_1 to x_n, are called a layer of neurons. These layers

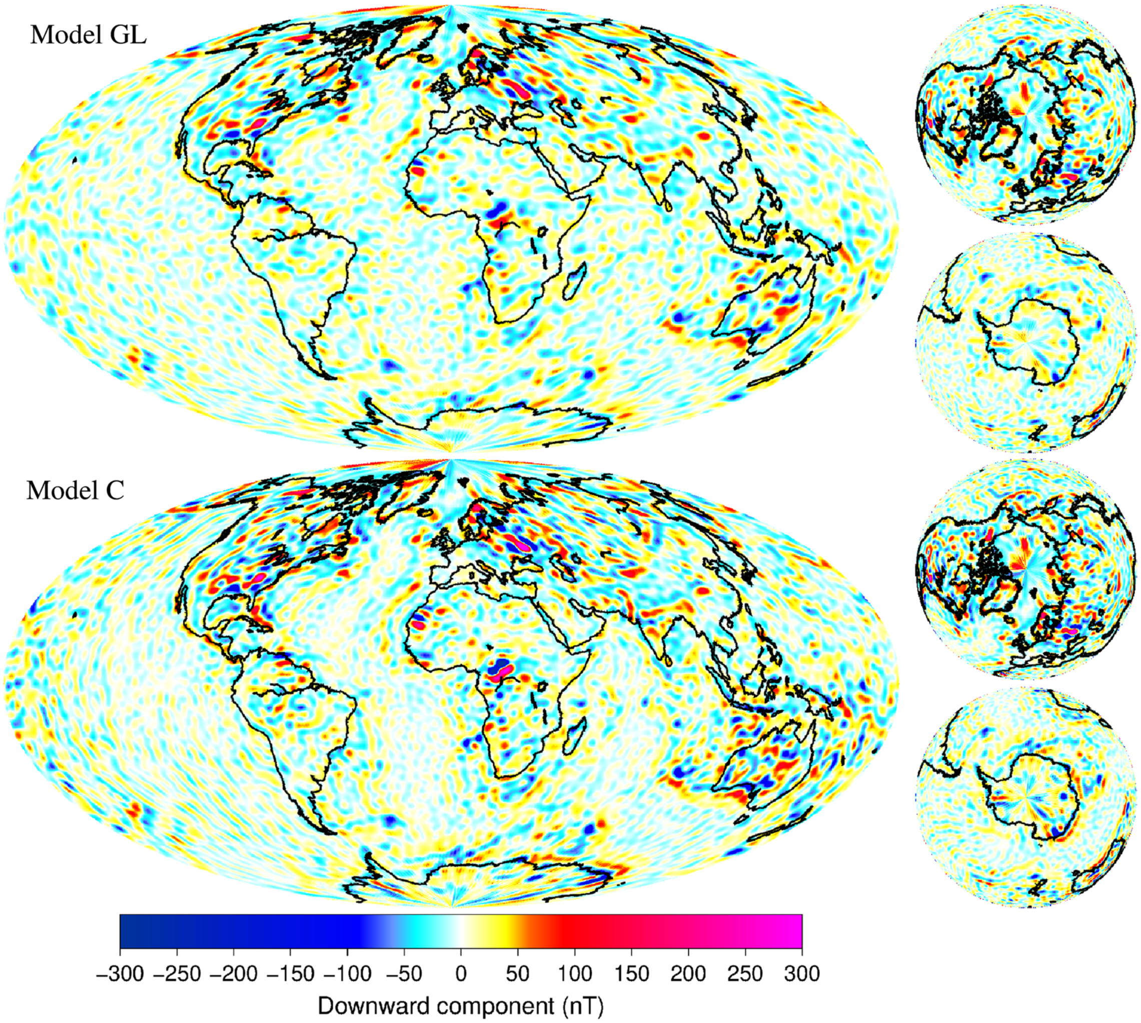


Fig. 14 Downward component of the lithospheric field at the Earth's surface expanded up to spherical harmonics degree $\ell = 100$. Top: model GL derived from GOCE data alone. Bottom: model C derived from CHAMP data

of neurons can also be stacked, meaning another layer takes the outputs y_1 to y_m of the first layer as the inputs for the neurons in the next layer. Each neuron of a layer can also be seen as a feature, as it is a new combination of the inputs that acts like a machine-crafted feature.

The initial formula is very similar to a linear regression formula and by stacking these neurons there would be no non-linearity added to the system, which would limit its expressiveness in the kind of processes it could explain. Therefore, the sum of the products is fed into the non-linear activation function. There are many activation functions available, that act in different ways, one of these is the Rectified Linear Unit (RELU) (Nair and Hinton 2010), defined as follows:

$$f(x) = max(0, x)$$

Negative values become 0, while positive neuron calculations will get forwarded into the next layer. This enhances the expressiveness of the NN and gives the ability to model non-linearity. The Exponential Linear Unit (ELU) (Clevert et al. 2015) used in this work has a similar function shape like the RELU but overcomes some shortcomings. Normally, all intermediate layers, also called hidden layers, before the final output layer use activation functions in their output. For the final output layer, this is oftentimes not done as a linear combination of the previous layer is desired.

The construct of number of neurons and number of layers, as well as the last layer of neurons which constitutes the output is then called the NN or the model. In the beginning, all the weights within this system are initialized randomly. For the model to represent the desired function, e.g., mapping the satellite data input to the calibrated measurement, the weights need to be adjusted in a meaningful way, this is called the training of the NN. Contrary to algorithms like the least squares approach, NNs are trained with the backpropagation algorithm (Rumelhart et al. 1986). Therefore, each available data point in the training dataset is forwarded through the NN and the current result of this calculation is compared to the expected result for this data point which is also called the ground truth or the reference model. This error is then used as the gradient to adjust the weights and backpropagated through the NN to adjust the weights in the different layers in such a way that the prediction would come closer to the expected result if the calculation was repeated. To speed this process up and not rely on single data points, this error is calculated on groups of data points, also called a batch, for which one common gradient is calculated and then an adjustment to the network is made. The amount of adjustment towards the calculated gradient can be controlled with the learning rate, which normally lies in the interval of [0,1] and is multiplied with the calculated gradients before they are applied to the weights. One iteration through all batches of available data points and adjustments to the NN is called an epoch. The NN will be trained for multiple epochs and there are optimizers like the Adam optimizer (Kingma and Ba 2014) speeding up the process by modifying the gradients with historic information. After this training process, the weights of the NN do not change anymore and it is assumed that the NN represents the statistics of the data. Afterward, it can be used to perform predictions on similar data as it was trained with.

Abbreviations

CHAMP: CHAllenging Minisatellite Payload; ELU: Exponential linear unit; FAC: Field-aligned currents; GOCE: Gravity and steady-state Ocean Circulation Explorer; GRACE: Gravity Recovery And Climate Experiment; GRACE-FO: Gravity Recovery And Climate Experiment Follow-On; HPO: Hyperparameter optimization; MAE: Mean absolute error; ML: Machine learning; MLT: Magnetic Local Time; MSE: Mean squared error; NEC: North–East–Center frame; NN: Neural network; QDLAT: Quasi-dipole latitude; STD: Standard deviation.

Acknowledgements

We thank Jan Rauberg for his help in interpreting and evaluating the results. We also thank the Geomagnetic World Data Centre Kyoto for the Dst index, and the Astrophysical Observatory and Natural Resources Canada for the F10.7 index.

Author Contributions

KS preprocessed and calibrated the data, and wrote the manuscript. KS, IM, and CS designed the study. JB integrated the data. KS, IM, CS, MK, and OK evaluated and reviewed the results of the study. All authors read and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work is supported through HEIBRIDS - Helmholtz Einstein International Berlin Research School in Data Science under contract no. HIDSS-0001. This study has been partly supported by Swarm DISC activities funded by ESA under contract no. 4000109587/13/I-NB.

Availability of data and materials

The datasets generated and analyzed during the current study are available under (Styp-Rekowski et al. 2022) ftp://isdcftp.gfz-potsdam.de/platmag/MAGNETIC_FIELD/GOCE/ML/v0204/.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Distributed and Operating Systems, Technical University of Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany. [2]GFZ German Research Centre for Geosciences, Helmholtz Centre Potsdam, Telegrafenberg, 14473 Potsdam, Germany. [3]Leibniz Institute of Atmospheric Physics at the University of Rostock, Schloßstraße 6, 18225 Kühlungsborn, Germany.

Received: 6 May 2022 Accepted: 24 August 2022
Published online: 13 September 2022

References

Alken P, Olsen N, Finlay CC (2020) Co-estimation of geomagnetic field and in-orbit fluxgate magnetometer calibration parameters. Earth, Planets and Space 72:1–32. https://doi.org/10.1186/s40623-020-01163-9

Anderson B, Ohtani S-I, Korth H, Ukhorskiy A (2005) Storm time dawn-dusk asymmetry of the large-scale Birkeland currents. J Geo Res, 110(A12), A12220. https://doi.org/10.1029/2005JA011246

Baerenzung J, Holschneider M, Wicht J, Lesur V, Sanchez S (2020) The Kalmag model as a candidate for IGRF-13. Earth Planets Space 72:1–13

Billingsley Billingsley TFM100S Magnetometer. Billingsley Aerospace Defense. https://magnetometer.com/wp-content/uploads/TFM100S-Spec-Sheet-February-2008.pdf

Clevert D-A, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289

Drinkwater M, Floberghagen R, Haagmans R, Muzi D, Popescu A (2003) VII: closing session: GOCE: ESA's first earth explorer core mission. Space Sci Rev 108:419–432

European Space Agency (2019) GOCE telemetry data collection. Version 1.0. GOCE telemetry packets description. https://doi.org/10.5270/esa-7nc8pjp

Finlay CC, Kloss C, Olsen N, Hammer MD, Tøffner-Clausen L, Grayver A, Kuvshinov A (2020) The CHAOS-7 geomagnetic field model and observed changes in the South Atlantic Anomaly. Earth Planets Space 72:156. https://doi.org/10.1186/s40623-020-01252-9

Floberghagen R, Fehringer M, Lamarre D, Muzi D, Frommknecht B, Steiger C, Piñeiro J, Da Costa A (2011) Mission design, operation and exploitation of the gravity field and steady-state ocean circulation explorer mission. J Geo 85:749–758. https://doi.org/10.1007/s00190-011-0498-3

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neu Net 2:359–366

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv in Neu Inf Pro Sys 30:4765-4774

Matzka J, Bronkalla O, Tornow K, Elger K, Stolle C (2021) Geomagnetic Kp index. GFZ GRCG. https://doi.org/10.5880/Kp.0001

Michaelis I, Styp-Rekowski K, Rauberg J, Stolle C, Korte M (2022) Preprint) Geomagnetic data from the GOCE satellite mission. ESpace Science Open Archive. https://doi.org/10.1002/essoar.10511006.1
Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on international conference on machine learning. Omnipress, Haifa, Israel, pp 807-814
Nose M, Sugiura M, Kamei T, Iyemori T, Koyama Y (2015) Dst Index. WDC for Geomagnetism, Kyoto. https://doi.org/10.17593/14515-74000
Olsen N (2021) Magnetometer data from the GRACE satellite duo. Earth Planets Space 73:1–20
Olsen N, Albini G, Bouffard J, Parrinello T, Tøffner-Clausen L (2020) Magnetic observations from CryoSat-2: calibration and processing of satellite platform magnetometer data. Earth Planets Space 72:1–18
Olsen N, Friis-Christensen E, Floberghagen R, Alken P, Beggan CD, Chulliat A, Doornbos E, Da Encarnação JT, Hamilton B, Hulot G, Van Den Ijssel J, Kuvshinov A, Lesur V, Lühr H, Macmillan S, Maus S, Noja M, Olsen PEH, Park J, Plank G, Püthe C, Rauberg J, Ritter P, Rother M, Sabaka TJ, Schachtschneider R, Sirol O, Stolle C, Thébault E, Thomson AWP, Tøffner-Clausen L, Velímský J, Vigneron P, Visser PN (2013) The Swarm satellite constellation application and research facility (SCARF) and Swarm data products. Earth Planets Space 65:1189–1200. https://doi.org/10.5047/eps.2013.07.001
Olsen N, Stolle C (2012) Satellite geomagnetism. Ann Rev of Ear and Pla Sci 40:441–465
Reigber C, Lühr H, Schwintzer P (2002) CHAMP mission status. Adv in Spa Res 30:129–134
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. Adv in Neu Inf Proc Sys 25:2951-2959
Stolle C, Michaelis I, Xiong C, Rother M, Usbeck T, Yamazaki Y, Rauberg J, Styp-Rekowski K (2021) Observing Earth's magnetic environment with the GRACE-FO mission. Earth Planets Space 73:84–104. https://doi.org/10.1186/s40623-021-01364-w
Stolle C, Olsen N, Anderson B, Doornbos E, Kuvshinov A (2021) Special issue "characterization of the geomagnetic field and its dynamic environment using data from space-based magnetometers". Earth Planets Space 73:51–54. https://doi.org/10.1186/s40623-021-01409-0
Styp-Rekowski K, Stolle C, Michaelis I, Kao O (2021) Calibration of the GRACE-FO Satellite Platform Magnetometers and Co-Estimation of Intrinsic Time Shift in Data. In: 2021 IEEE International conference on Big Data (Big Data). IEEE, pp 5283-5290. https://doi.org/10.1109/BigData52589.2021.9671977
Styp-Rekowski K, Michaelis I, Stolle C, Baerenzung J, Korte M, Kao O (2022) GOCE ML-calibrated magnetic field data. V. 0204. GFZ Data Services. https://doi.org/10.5880/GFZ.2.3.2022.002
Tapping K (2013) The 10.7 cm solar radio flux (F10. 7). Space Weather 11:394–406
Thébault E, Hulot G, Langlais B, Vigneron P (2021) A spherical harmonic model of Earth's lithospheric magnetic field up to degree 1050. Geo Res Let 48:e2021GL095147. https://doi.org/10.1029/2021GL095147

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.