MethodsX 10 (2023) 101999

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Principal component analysis as a tool to extract Sq variation from the geomagnetic field observations: Conditions of applicability

Anna Morozova [a,c,*], Rania Rebbah [b,c]

[a] Instituto de Astrofísica e Ciências do Espaço (IA-U.Coimbra), University of Coimbra, Coimbra, Portugal,
[b] CITEUC, University of Coimbra, Department of Physics, Coimbra, Portugal
[c] Department of Physics, FCTUC, University of Coimbra, Coimbra, Portugal

ARTICLE INFO

Method name:
Extraction of the solar quiet (Sq) variation using the principal component analysis (PCA)

Keywords:
Principal component analysis
Geomagnetic field
Solar quiet variation (Sq)
Coimbra magnetic observatory (COI)

ABSTRACT

We analyzed the applicability of the principal component analysis (PCA) as a tool to extract the Sq variation of the geomagnetic field (GMF) taking into account different geomagnetic field components, data measured at different levels of the solar and geomagnetic activity, data from different months.

The validation of the method was performed with geomagnetic data obtained at the Coimbra Magnetic Observatory in Portugal (40° 13′ N, 8° 25.3′ W, 99 m a.s.l., IAGA code COI).

GMF variations obtained with PCA were "classified" as Sq_{PCA} using reference series: (1) obtained from the observational data (Sq_{IQD}), (2) simulated by ionospheric field models.

While our results show that both the data-based and model-based reference series can be used, the DIFI3 model performs better as a reference series for GMF at middle latitudes.

We also recommend to estimate the similarity of the series with a metric that account for possible local stretching/compressing of the compared series, for example, the dynamic time warping (DTW) distance.

Since the validation of the method was performed on the geomagnetic series obtained at a mid-latitudinal European observatory, we recommend performing additional tests when applying this method to data obtained in other regions/latitudes.

- For the Y and Z components of the geomagnetic field PCA can be used to extract Sq variations from the observations without any additional procedures and SqPCA is equals to PC1.
- For the X component PCA can be used to extract Sq variation from the observations of the X component, but further analysis, for example, a comparison to a set of reference curves either obtained from the data analysis or generated using models, is always needed to classify PCs of the X component.
- We recommend to use data generated by DIFI-class models as reference series and the dtw metric (dynamic time warping distance) to classify Sq_{PCA}.

* Corresponding author.
E-mail address: anna.morozova@uc.pt (A. Morozova).

https://doi.org/10.1016/j.mex.2023.101999
Received 15 July 2022; Accepted 2 January 2023
Available online 2 January 2023

Specifications table

Subject area:	Earth and Planetary Sciences
More specific subject area:	Geomagnetic field variations
Name of your method:	Extraction of the solar quiet (Sq) variation using the principal component analysis (PCA)
Name and reference of original method:	Xu, W.Y. and Kamide, Y. [38]: Decomposition of daily geomagnetic variations by using method of natural orthogonal component. Journal of Geophysical Research: Space Physics, 109(A5).doi: 10.1029/2003JA010216. Chen, G.X., Xu, W.Y., Du, A.M., Wu, Y.Y., Chen, B. and Liu, X.C. [9]: Statistical characteristics of the day-to-day variability in the geomagnetic Sq field. Journal of Geophysical Research: Space Physics, 112(A6), doi:10.1029/2006JA012059. De Michelis, P., Tozzi, R. and Meloni, A. [13] On the terms of geomagnetic daily variation in Antarctica. Ann. Geophys, 27, pp.2483–2490. De Michelis, P., Tozzi, R. and Consolini, G. [12] Principal components' features of mid-latitude geomagnetic daily variation. Ann. Geophys, 28, pp.2213–2226, doi:10.5194/angeo-28-2213-2010.
Resource availability:	The COI 1 h data for all geomagnetic components can be downloaded from the World Data center for Geomagnetism using the Geomagnetism Data Portal at http://www.wdc.bgs.ac.uk/dataportal/ (station name: "Coimbra", IAGA code: "COI"). The Sq_{IQD} and PCs analyzed in the paper can be downloaded at Mendeley Data, doi: 10.17632/jcmdrm5f5x.1, http://dx.doi.org/10.17632/jcmdrm5f5x.1 . All indices of the solar and geomagnetic activity used in this work can be downloaded from the OMNI database at https://omniweb.gsfc.nasa.gov/form/dx1.html . The CM5 model is available at https://ccmc.gsfc.nasa.gov/models/modelinfo.php?model=CM5 . The DIFI3 model is available at http://geomag.colorado.edu/difi-calculator . A package for R to calculate dtw with different algorithms that were used in this study was developed by Giorgino [15] and can be downloaded from https://cran.r-project.org/web/packages/dtw/index.html .

Method details

Main method to extract Sq variation of the geomagnetic field

There are three main types of geomagnetic field variations on the time scale from hours to several days: regular variations during a calendar or solar day (so-called "daily" or "solar" variations known as S-type variations), regular variations during a lunar month (L-variation) and irregular variations often associated with storms and substorms and called "disturbances" (Dst-variations), see Chapman and Bartels [8]. The S-type variations are divided into two main classes: the "daily (solar) quiet" variation, Sq, which is observed most clearly during the geomagnetically quiet days, and the "daily (solar) disturbed" variation, SD, [8,39]. More details on the origin and the character of the Sq variation of the geomagnetic field can be found in the Supplementary Material (SM1).

The standard method to obtain Sq from the ground observations of the geomagnetic field consists of the selection of days with the lowest level of geomagnetic field perturbations (so-called "quiet days"), typically, five days per calendar month, and averaging of the daily geomagnetic field variations for a certain component over selected days. These days can be defined using the data of an individual observatory (local quiet days) or using the data from the same set of observatories that are used to calculate the Kp index (international quiet days – IQD), see Chapman and Bartels [8]. Hereafter, the Sq variation obtained using IQD is named "Sq_{IQD}".

In this work, we used IQDs routinely provided by the GFZ German Research center for Geosciences at the Helmholtz center in Potsdam, Germany and available at https://www.gfz-potsdam.de/en/kp-index/ or ftp://ftp.gfz-potsdam.de/pub/home/obs/kp-ap/quietdst/. Please note that for the COI observatory in most cases a set of IQD coincides with a set of quiet days defined using the local K-index (local quiet days, LQD). There were only few months during the studied time interval (2007–2017) when sets of five days of IQD and LQD have differences, and those differences were of the order of one day. The sets of 10 days of IQD and LQD were always the same. Thus, we decided to use the widely used IQD for calculation of Sq. For stations located in different regions/latitudinal zones this maybe not the case.

The Sq_{IQD} variation for a certain month is calculated as the mean daily variation for five IQDs of a month. Before the averaging, a baseline was removed from the raw daily series of the X, Y and Z components. In this work, the baseline was defined as a mean calculated for the night hours: 00:30 UTC, 01:30 UTC, 02:30 UTC, 03:30 UTC and 23:30 UTC of each analyzed day (for Coimbra UTC = LT). Thus, the Sq_{IQD} variation values for the night hours are close to zero, and there is no significant difference between the night values of Sq at the beginning and the end of a day.

PCA-based methods to extract Sq variation of the geomagnetic field

Another way to extract regular variations as Sq is to apply a decomposition method to the geomagnetic field data: e.g., the wavelet analysis [22], the empirical mode decomposition [28] or the principal component analysis, PCA [9,12,13,38]. Also, the shape and position of the vortex can be deduced from the observational data using the spherical harmonic analysis by calculating the equivalent current system [20,35] or it can be reconstructed as equivalent electric current vectors (horizontal component) from the observed horizontal geomagnetic field vector [33,34].

First attempts to use PCA (sometimes known as a method of the natural orthogonal component, NOC) to extract regular variations of the geomagnetic field were made in the 1970s-1990s [16–19,29] but were not actively supported by the geomagnetic community [23]. Golovkov et al. [18,19] and Golovkov and Zvereva [16,17] showed that for the H component of the geomagnetic field and for the geomagnetically quiet time intervals, the Sq variation can be associated with the first (or first and third) principal components (PC) and the second PC can be identified as SD variation. For the geomagnetically active time intervals the first PC was identified as SD, and the second and third PCs were identified as Sq. Dependence of the order of a PC that can be identified as Sq or SD on the

Table 1
PCA variance fraction (in %) of the geomagnetic field X, Y and Z components. The minimum, maximum and mean values of the variance fraction associated with the first three principal components (PC1-PC3) and the cumulative variance fraction (Σ) for the first three PCs. Bold marks PCs that are essential for Sq extraction.

	X component			Y component			Z component		
	min	mean	max	min	mean	max	min	mean	max
PC1	**28.8**	**49.5**	**78.2**	**58.1**	**82.7**	**94.0**	**62.1**	**85.0**	**94.9**
PC2	**9.5**	**21.1**	**36.9**	1.7	6.5	22.0	1.9	6.2	17.9
PC3	**4.2**	**10.9**	**20.7**	1.1	3.5	8.5	0.8	3.1	10.9
Σ(PC1 to PC3)	67.2	81.6	93.8	82.5	92.7	98.1	83.6	94.3	98.0

latitude was also shown. Both the existence of the daily variability of the Sq field and the need for studying it was also emphasized in the early works.

Later, Xu and Kamide [38] and Chen et al. [9] revived the interest of the geomagnetic community in PCA as a useful tool that allows not only to extract regular variations of the geomagnetic field, as Sq and SD, but also to analyze seasonal and geographic variations of the phase and amplitude of the Sq and SD fields and the dependence of their intensity on the level of the solar and geomagnetic activity. Works of Wu et al. [37], De Michelis et al., [12,13], Bhardwaj et al. [5,6] and others (see also review by [39]), generally confirmed the applicability of PCA to the extraction of the regular geomagnetic field variations observed at different latitudes, and for the time intervals of different length and corresponding to different geomagnetic activity levels. However, the results obtained for different regions/time intervals were somewhat different.

In particular, it was found that for the H (X) component of the geomagnetic field for the Asian sector [5,6,9,37,38] the Sq variation is filtered to the first PC and the SD variation is filtered to the second PC. On the contrary, for the European sector [12] PC1 is associated with SD and PC2 is associated with Sq. This difference can be explained both by the different geographic positions of the stations whose data were used for PCA and by the different studied time intervals. Also, for the Y (D) and Z components of the geomagnetic field for the European sector PC1 was identified as Sq and PC2 as SD.

To our knowledge, no systematic study of the applicability of PCA as a tool to extract Sq-type variations was performed yet and no possible explanation for the differences mentioned above was proposed. In this work, we present the results of such a study: we test PCA on different components of the geomagnetic field (X, Y and Z), on the data obtained in different months and under different levels of solar and geomagnetic activity. We also tested different lengths of the input data sets.

We use the geomagnetic field data obtained at a European mid-latitudinal geomagnetic observatory – Coimbra Magnetic Observatory (COI) in Portugal. The peculiarity of COI, and this can be also true for the L'Aquila observatory [12], is that it is located near the mean latitude of the focus of the Sq ionospheric current vortex. Thus, the shape of the Sq variations for the X component at COI can vary not only due to the intensity of the vortex but also due to the position of its focus: for some days COI is located to the north of the focus, for other days it is located to the south of the focus, and there are days when COI is located very near the focus latitude. These changes of the COI relative position result in different shapes of the Sq X variation (see SM1). Finally, contrary to all previous studies, we analyzed the data not on the annual or decadal time scale but on the monthly time scale as described below and in Morozova et al. [24,25].

Description of the proposed PCA-based method to extract Sq variation of the geomagnetic field

Principal component analysis (PCA) allows the extraction of main modes of variability of an analyzed series – principal components or PCs. The full descriptions of this widely used mathematical method can be found in, e.g., Björnsson and Venegas [7], Hannachi et al. [21], Shlens [32]. PCs are orthogonal and conventionally non-dimensional. The amplitudes of a PC for each of the analyzed days are given by the corresponding empirical orthogonal function (EOF). The combination of a PC and the corresponding EOF is called a "mode". The "significance" of each of the extracted modes is estimated from the corresponding eigenvalues as variance fraction (VF). VF can be between 0 and 1 and multiplied by 100% shows the percent of the total variability of the analyzed series related to a particular mode.

The PCA input matrices were constructed as follows. For the individual months and years, the input matrices have 24 rows (24 hourly values per day) and from 28 to 31 columns (a column for a day) depending on the analyzed month. All February matrices have the size 24 × 28 (the days of February 29 of the leap years were removed to simplify comparison between different years). For the individual months but for the "all years" series the input matrices have sizes 24 × 308, 24 × 330 or 24 × 341, depending on the analyzed month. The singular value decomposition (SVD) approach was used to solve the matrix equations.

In this configuration of the input matrices, the principal components (PCs) correspond to daily variations of different types that can be matched up with Sq variation calculated using the standard approach. The amplitudes of PCs for an individual day are given by corresponding EOFs.

Only three first PCs were selected for further analysis. Overall, the first three PCA modes explain together up to 67–94% of the COI X variability, and up to 83–98% of the COI Y and COI Z series variability depending on a month and a year. Table 1 shows VFs associated with the first three PCA modes of the variations of the X, Y and Z components.

During further analyses, PCs were compared to reference series and those PCs that can be classified as Sq were denoted as Sq_{PCA}.

Below we provide a systematic analysis of the PCA's performance on mid-latitudinal geomagnetic data for different geomagnetic field components, different seasons and under different levels of the solar and geomagnetic activity. We also tested if only one PC is always sufficient to represent an Sq-type variation or a combination of two PCs should be considered as well.

Data and methods used for validation

Data

Geomagnetic field data

Geomagnetic measurements at the Coimbra Magnetic Observatory in Portugal (40° 13′ N, 8° 25.3′ W, 99 m a.s.l., IAGA code COI) have been started in 1866 [26,27]. The last changes of the instruments took place at COI in 2006: new sets of the absolute instruments were installed providing good quality measurements of geomagnetic field components with 1 hour time resolution [27]. Since that time to the present, there were no changes in the instruments or station location, and the data obtained between 2007 and the present time can be considered homogeneous [27]. A detailed description of the COI instruments and metadata for the series of the geomagnetic field components can be found in Morozova et al. [24,26,27]. The 1 h data for all geomagnetic components can be downloaded from the World Data center for Geomagnetism using the Geomagnetism Data Portal at http://www.wdc.bgs.ac.uk/dataportal/ (station name: "Coimbra", IAGA code: "COI"). These data were used to obtain both the Sq_{IQD} variation and the main PCA modes of the geomagnetic field variations.

The dataset consists of 1 h data on the variations of the X (northern), Y (eastern) and Z (vertical) components of the geomagnetic field measured at COI during 11 years from January 1, 2007, to December 31, 2017. This time interval covers (approximately) one solar cycle. The data for different components were tested separately. The data were used on the time scale of one calendar month. The Sq_{IQD} variation and the PCA modes were calculated for each month both for the individual years, i.e., using only the data for January 2007, for January 2008, etc., separately, and for each month but all years together, i.e., using the data for January 2007 and January 2008, etc. together, hereafter "all years" series. As a result, for each of the three analyzed components, there were obtained $11\times12 = 132$ series for individual months and years, and 12 "all years" series. This dataset is described in detail in Morozova et al. [24] and is available in Morozova et al. [25]. Standard errors (SE) for the Sq_{IQD} values were calculated for each month relative to the Sq_{IQD} "all years" series.

Geomagnetic field models

As reference series (see below in Sec. 6) for the ionospheric field of the X omponent the ionospheric fields generated by two geomagnetic field models, CM5 and DIFI3, were used. The CM5 and DIFI3 reference series were generated for the calendar day 15 of each of 12 months, from January to December. Since for both models the ionospheric field outputs for different years have the same shape but change only in amplitude, the CM5 and DIFI3 reference series (Sq_{CM5} and Sq_{DIFI3}, respectively) were used in arbitrary units (a.u.). Detailed descriptions of the models can be found in Sabaka et al. [30,31], and Chulliat et al. [10,11] and Thébault et al. [36], respectively, and a short summary can be found in the Supplementary Material (SM2).

Solar and geomagnetic indices

To estimate the decadal and seasonal variations of the level of the solar and geomagnetic activities we used the following indices. The solar activity was represented by the daily means of the sunspot number series (R) and series of the F10.7 index reflecting variations of the solar UV flux. To see variations of the geomagnetic activity level we used daily means of the Dst, Kp and ap, and AE geomagnetic indices. All the indices were obtained from the OMNI database at https://omniweb.gsfc.nasa.gov/form/dx1.html. The daily mean values of these indices were used to calculate both the monthly means and the IQD means (means calculated using only 5 IQD of a month) for each of the studied months. Corresponding plots can be found in the Supplementary Material (SM3, Figs. S3.1-S3.4).

Methods used to classify PCs

The daily variations obtained by PCA (PC1, PC2 and PC3) were compared to the Sq_{IQD}, Sq_{CM5} and Sq_{DIFI3} variations and classified, when possible, as Sq_{PCA} using two classification metrics: (1) the absolute value of the Pearson correlation coefficient (r), and (2) a metric called the dynamic time warping distance (dtw). Short descriptions of these metrics are given below.

We tested two approaches to the PCs' classification: allowing the combined classification (either one or a sum of two PCs can be classified as Sq_{PCA}) and not allowing the combined classification, i.e., single classification (only one PC per studied month is classified as Sq_{PCA}).

The need for the combined classification can be justified by the possibility of PCA to decompose an Sq-type variation into several modes for months when the solar and geomagnetic activities were very low. In such a case an Sq-type variation can be decomposed by PCA into several modes that contain different fine features of Sq.

The sums of PCs were calculated as weighted sums with weights being the monthly mean values of the corresponding EOFs.

For each set of PCs, the classification metrics were calculated between those PCs (or their sums) and the corresponding reference series (Sq_{IQD}, Sq_{CM5} or Sq_{DIFI3}). Only PC (or a sum of PCs) with metrics that are above (below) a predefined threshold for r (dtw) are used for further classification, and PC (or a sum of PCs) with highest (lowest) values of r (dtw) was classified as Sq_{PCA}.

Correlation analysis and correlation coefficient r

Here we used the standard Pearson correlation coefficient. Since in this work we used the SVD method to perform PCA, PCs and EOFs are resolved accurately to a sign. This is because both +1*PC & +1*EOF, and −1*PC & −1*EOF are solutions for an input PCA matrix. There is no general way to solve the sign ambiguity. Keeping this in mind we used the absolute values of the correlation coefficients |r|. The threshold for the classification using the correlation analysis was set as $|r| \geq 0.45$.

The significance of the correlation coefficients was estimated using the Monte Carlo approach with artificial series constructed by the "phase randomization procedure" Ebisuzaki [14]. The obtained statistical significance (p value) considers the probability of a random series to have the same or higher |r| as in the case of a tested pair of the original series.

Dynamic time warping and the dtw metric

When using the correlation coefficient as a measure of similarities between two series one must remember that its value is mostly affected by the similarity of main features existing in the compared series. It may be not sensitive to small-scale features or non-systematic shifts of the local minima or maxima (systematic shifts of the local minima and maxima or a relative shift of a whole series can be accounted for by the lagged correlation analysis). Thus, we would need a metric that is sensitive to irregular deformation of series.

The dynamic time warping (DTW) is a popular metric for comparing time series that is insensitive to local compression and stretches allowing to optimally deforms one of the two input series onto another and calculate a certain measure for the "distance" (dtw) between the studied series [15]. The smaller the "distance" the higher the similarity between the series, contrary to the correlation coefficient which is higher in the absolute value for the similar series. A description of the DTW algorithm can be found in Giorgino [15], see also reference herein.

In short, when the similarity of two-time series is studied, one of the series is taken as a "reference" and another is locally stretched or compressed to make it resemble the "reference" as much as possible. The distance (dtw value) between the two series is computed after all stretching/compressing are finished by summing the distances of individual aligned elements. Several DTW algorithms have been proposed in the 1970s in the context of speech recognition [15].

The dtw parameter, contrary to r, is not defined on a certain absolute scale. To be able to compare the r and dtw values we had to (1) use standardised (zero mean and unity standard deviation) series to perform the DTW analysis and (2) to compare r and dtw sets obtained for the same pairs of series to see if there is any correspondence between the values of r and dtw. In our tests, it was found that this correspondence can be well fit by Eq. (1)

$$dtw = \frac{A(1-r)}{B+(1-r)}, \tag{1}$$

where A and B are fitting coefficients. An example of a dtw fit on r is shown in the Supplementary Material (SM4, Fig. S4.1) and Table S4.1 presents dtw values for certain values of r when PCs series are compared to different reference series.

The mean dtw threshold that is equivalent to $|r| \geq 0.45$ is $dtw \leq 0.58$ but for individual pairs of PCs vs a reference series, it varies from 0.57 to 0.62. As is shown below in Sec. 6 the DTW analysis allows for a better estimation of the similarity of the studied series than the correlation analysis, and the number of the classified series using this dtw threshold is higher than the number of the classified series using the correlation analysis with the $|r| \geq 0.45$ threshold.

Performance of PCA as a tool to extract Sq

Performance of PCA for the Y and Z components

Fig. 1 shows examples of the first PCs together with Sq_{IQD} for the Y (top panel) and Z (bottom panel) components. The Sq_{IQD} series of the Y and Z components have very stable and specific shapes (see, for example, the width of the Sq_{IQD} ±SE bands in Fig. 1): the Sq_{IQD} Z variation is symmetric around the local noon, while Sq_{IQD} Y is anti-symmetric. These shapes agree well with the shapes of the Sq variations for the Y and Z components expected at a mid-latitudinal geomagnetic station (see SM1). The series for PC1s-PC3s and Sq_{IQD} for all months and all years can be found in Morozova et al. [25].

The comparison of PC1-PC3 obtained for the Y and Z components with corresponding Sq_{IQD} using the correlation analysis shows that all the PC1 series for both components can be reliably classified as Sq_{PCA}. Fig. 2 shows tile plots of the classification of PC1s for Y and Z with numbers showing values of the correlation coefficients (all shown $|r| \geq 0.88$, all p value < 0.01). PC2s of the Y and Z series rarely have a significant correlation with Sq_{IQD} (only 1 case out of 144 for Y and Z, respectively, $|r| = 0.48–0.55$, p value > 0.2), and no PC3 has such correlations (corresponding plots can be found in the Supplementary Material (SM5, Figs. S5.1-S5.4). The use of the combined classification does not significantly improve the classification of PCs: the addition of other PCs to PC1 increases the r values insignificantly (please compare Fig. 2 and Figs. S5.1-S5.2 with Figs. S5.3-S5.4 in SM5). Therefore, the combined classification is not needed in the case of the Y and Z components.

Thus, for the Y and Z components for all analyzed months and all 11 years from 2007 to 2017, the PC1 series are defined as Sq_{PCA}. This means that Sq is the dominant variation for these components. This also means that for the Y and Z components the probability for Sq variation to be extracted as PC1 is 100%, and, therefore, PCA can be used as a reliable method to extract Sq variations from the Y and Z series when the use of IQD is not possible or not applicable for some reason (e.g., gaps in the analyzed series of the geomagnetic measurements or the overall high geomagnetic activity level of the studied months). Also, for the Y and Z components, the PCA performance does not depend on the season or the level of the solar/geomagnetic activity.

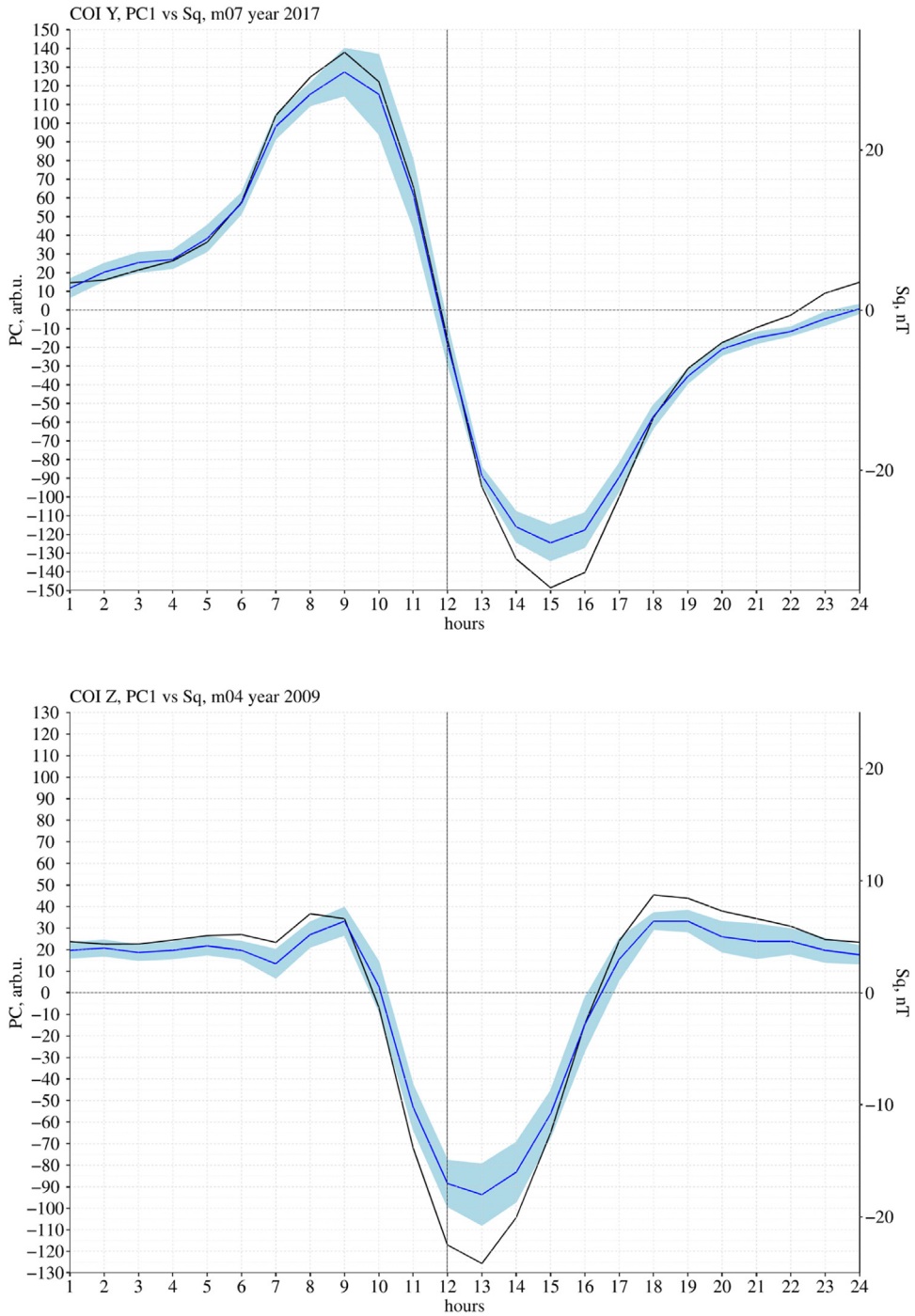

Fig. 1. Examples of Sq_{IQD} and PC1 daily variations for Y and Z: Sq_{IQD} (blue lines, in nT) and PC1 (black lines, in a.u.) daily variations for the Y (top) and Z (bottom) components. Light blue bands show Sq_{IQD} ± SE values.

Also, using PCA we can estimate a part of the variability of the original Y and Z series associated with the Sq variation. In the case of the validation dataset, as follows from Table 1, the mean variance fraction for PC1 for the Y and Z components is ~84%.

It is also possible to detect seasonal variations of VF associated with PC1: in the presented case it is higher during the summer months and lower during winter. These seasonal variations of VF are not driven by the part of the geomagnetic activity, which is described by the Kp, ap or Dst indices: these indices have semi-annual cycles (see Fig. S2.3-S2.4). On the other hand, the AE index describing the geomagnetic activity related to the high-latitudinal magnetosphere and ionosphere has an annual cycle with a

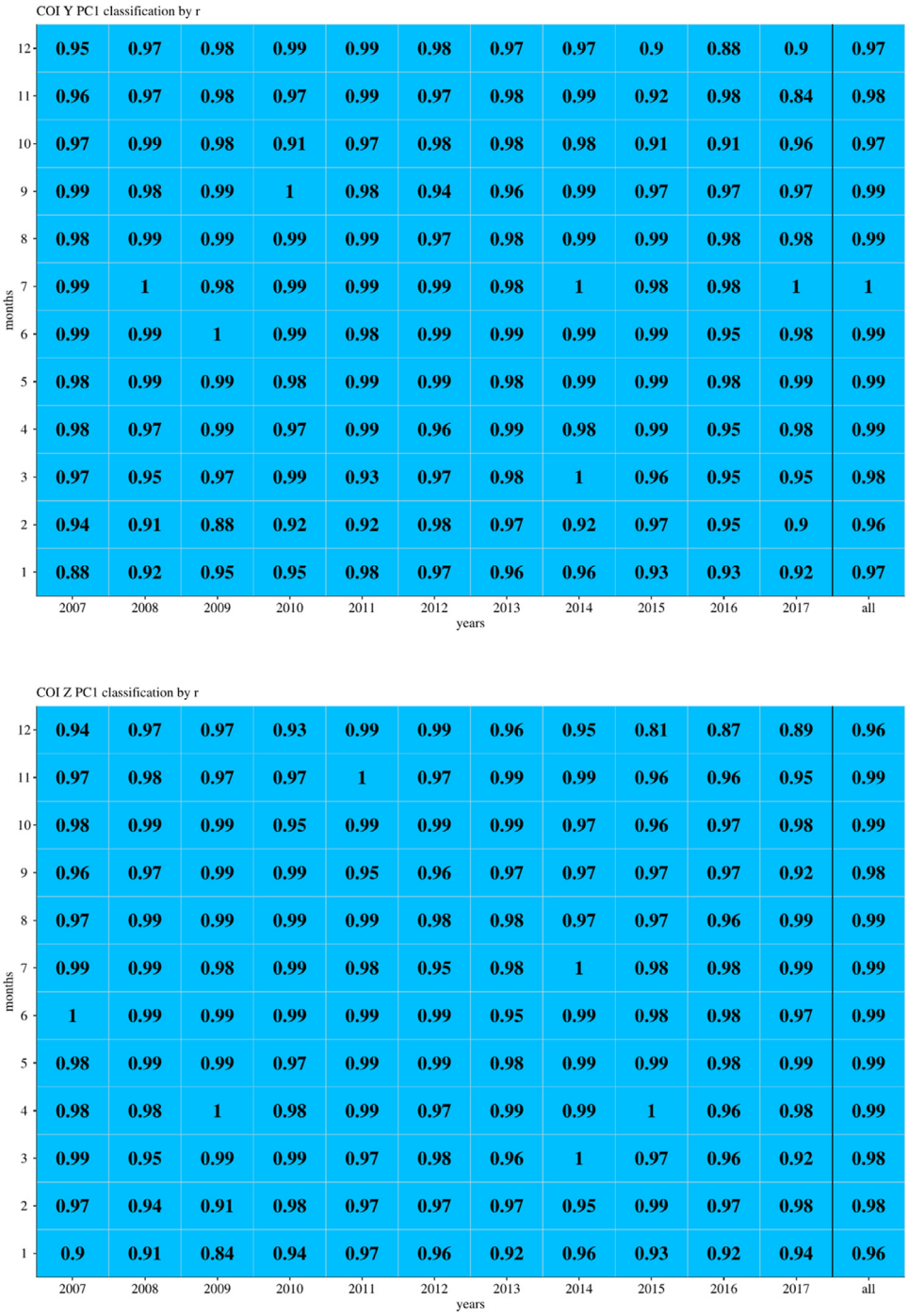

Fig. 2. Correlation coefficients between the Sq_{IQD} and PC1 series for Y (top) and Z (bottom). Numbers show correlation coefficients for different months (Y-axis) and different years (X-axis). Blue tiles mark the PCs classified as Sq (single classification using r).

maximum in summer (see Figs. S2.3-S2.4). However, to our mind, the main reason for an increase of VF for the first PCA mode during summer is the overall increase of the insolation and the intensification of the Sq current vortex during the summer months [39].

On the decadal timescale, VF of mode 1 anti-correlates with geomagnetic activity, whereas VFs for mode 2 and mode 3 correlate with geomagnetic activity level (see Table 2). This is expected since PC1s for the Y and Z components are associated with Sq, while, consequently, PC2 and PC3 contain variations related to disturbances (e.g., SD and Dst): during years with higher geomagnetic activity the contribution of the disturbance-type variations to the total variability of the Y and Z components increases resulting in higher VF values.

Table 2
Correlation between VF and the solar and geomagnetic activity. The correlation coefficients are calculated between the mean VF associated with a PC for the Y, Z and X components for a certain year, and the corresponding mean values of the solar and geomagnetic indices. Only $|r| \geq 0.3$ are shown, with *p values* in parentheses (only *p values* ≤ 0.2 are shown). Statistically significant correlation coefficients (p values ≤ 0.05) are in bold.

		Geomagnetic indices				Solar indices	
		AE	ap	Kp	Dst	R	F10.7
Y component	PC1	−0.71 (0.15)	−0.78 (0.06)	−0.74 (0.09)			
	PC2	0.57	0.66 (0.13)	0.61 (0.19)			
	PC3	0.81 (<0.01)	0.81 (0.01)	0.75 (0.03)	−0.44		
Z component	PC1	−0.50	−0.56 (0.2)	−0.50		0.43	0.44
	PC2	0.64 (0.09)	0.69 (0.05)	0.63 (0.09)			
	PC3					−0.54 (0.04)	−0.59 (0.02)
X component	PC1					0.46 (0.17)	0.54 (0.05)
	PC2	−0.41 (0.12)	−0.49 (0.04)	−0.43 (0.13)	0.31		
	PC3						

These results agree with previous findings of Golovkov et al. [18,19], Golovkov and Zvereva [16,17] and De Michelis et al. [12] for different epochs and latitudinal zones.

Performance of PCA for the X component

Fig. 3 shows examples of PC1, PC2 and PC3 together with Sq_{IQD} for the X component. All PC1, PC2 and PC3 series as well as the Sq_{IQD} series can be found at Morozova et al. [25]. There are two main types of the shape of the Sq X variations obtained from the COI data:

- Shape A: the curves with a minimum (or maximum) near the local noon and secondary (with a lower amplitude) maximum (or minimum, respectively) in the early morning or late afternoon (see Fig. 3a).

- Shape B: the curves with two minima and two maxima of comparable amplitudes (see Fig. 3b-c).

According to Amory-Mazaudier [[1,2] and [3]], and Anad et al. [4], these two types of the Sq X shape can be interpreted, e.g., as caused by an Sq current vortex with a focus located to the south (or to the north, respectively) of the COI location - shape A, or very close to the latitude of COI (40°N) - shape B.

Classification of X component PCs by correlation analysis. The comparison of PC1-PC3 obtained for the X component with corresponding Sq_{IQD} using the correlation analysis shows that, contrary to the Y and Z components, no PC is always classified as Sq. Figs. 4-6 show the classification of PCs for the X component based on the correlation analysis using single and combined classification.

Single classification (Fig. 4): for the individual years' series PC1s and PC2s were classified as Sq at about the same rate (59 and 52 series, respectively, or 40–45% each) while PC3s are classified as Sq about three times less often (18 series or 14%). Only in 3 cases (2%) none of the first three PCs was classified as Sq. On the contrary, for the "all years" series (see Fig. 4, last columns) in 6 cases (50%) PC2s were classified as Sq and in 3 cases each either PC1s or PC3s (25% each) were classified as Sq.

Thus, for the single classification, the probabilities of PC1 or PC2 to be classified as Sq_{PCA} (or the probabilities of Sq-type variation to be filtered to the 1st or 2nd mode) are approximately equal and about three times higher than the probability of PC3 to be classified as Sq.

Combined classification (Figs. 5-6): for the individual years' series PC1s and PC3s were classified as Sq at the same rate (8 and 6 series, respectively, or 5–6%) while PC2s are classified as Sq about two-three times more often (15 series or 12%). The combinations of PCs were classified as Sq in 43 cases for PC1+PC2 (33%), in 25 cases for PC1+PC3 (19%) and 32 cases for PC2+PC3 (24%). Only in 2 cases (~1.5%) none of the first three PCs or their combination was classified as Sq. For the "all years" series (see Figs. 5-6, last columns) in 7 cases (58%) PC2+PC3 were classified as Sq, in 4 cases (33%) PC1+PC3 were classified as Sq, in 1 case (8%) PC2 was classified as Sq.

Thus, for the combined classification the most probable scenarios to extract Sq-type variations are (in the declining order) the combination of PC1+PC2, PC2+PC3 and PC1+PC3.

The results of both kinds of classification for the X component are in general agreement with previous results obtained for the European region [12]: Sq_{PCA} tends to be more frequently associated with PC2 than with other components.

The advantage of the combined classification is that the higher values of the correlation coefficients r were obtained for sums of PCs comparing to r for the individual PCs. In many cases the increase of the r values is small, however in some cases the use of a sum of PCs allows to increase the r value, for example, from 0.6 to 0.68 to 0.83–0.91 (the cases of June "all years" series, June 2009, July 2017, April 2015, or December "all years" series).

As follows from Table 1, the mean variance fractions for PC1, PC2 and PC3 for the X component are ~50%, ~21% and ~11%, respectively. The mean VF varies throughout the year: for PC1 it is higher in winter, and VFs of PC2 and PC3 are higher in summer. On the decadal time scale, see Table 2, VF of PC1 (PC2) correlates (anti-correlates) with variations of the geomagnetic activity through the 11-year cycle.

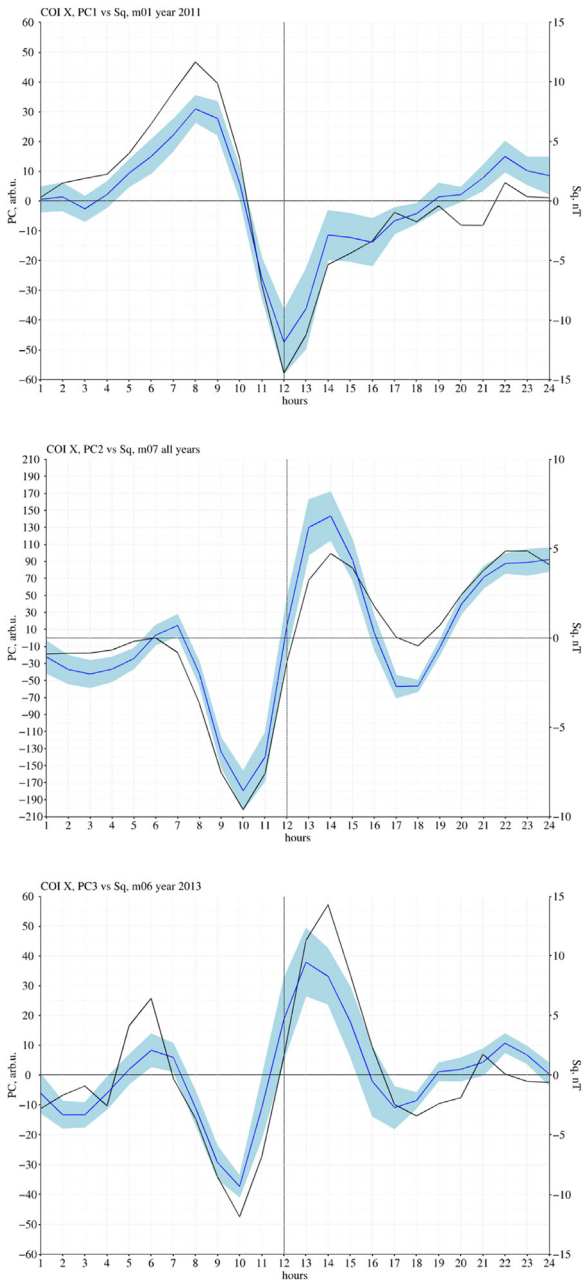

Fig. 3. Examples of Sq_{IQD} and PC1 daily variations for X. Sq_{IQD} (blue lines, in nT) and PC1 (top), PC2 (middle) and PC3 (bottom) daily variations (black lines, in a.u.) for the X component. Light blue bands show $Sq_{IQD} \pm SE$ values.

As one can see from Figs. 4-6, there is no clear seasonal or decadal pattern in the classification of PC1-PC3 for the X component as the Sq variation. We compared the number of months per year with PC1, PC2 or PC3 classified as Sq (single classification) with the annual mean values of the solar and geomagnetic activity indices. For the combined classification we made a similar comparison for the number of months per year with PC1+PC2, PC1+PC3 or PC2+PC3 classified as Sq (the number of single PCs classified as Sq using the combined classification is too small for a statistically significant analysis).

The obtained correlation coefficients (Table 3) are low and statistically insignificant (all p values > 0.2); however, we may conclude that, in general, the increase of the geomagnetic activity results in a more often classification of PC2 or PC1+PC3 as Sq variation; the increase of the solar activity results in a more often classification of PC1 or PC1+PC3 as Sq. We can interpret this as follows: for geomagnetically quiet epochs the Sq variation is, in most cases, the dominant variation for the X component and has a high probability to be filtered by PCA to the mode 1, while for the geomagnetically disturbed epochs the disturbance-type variations (like

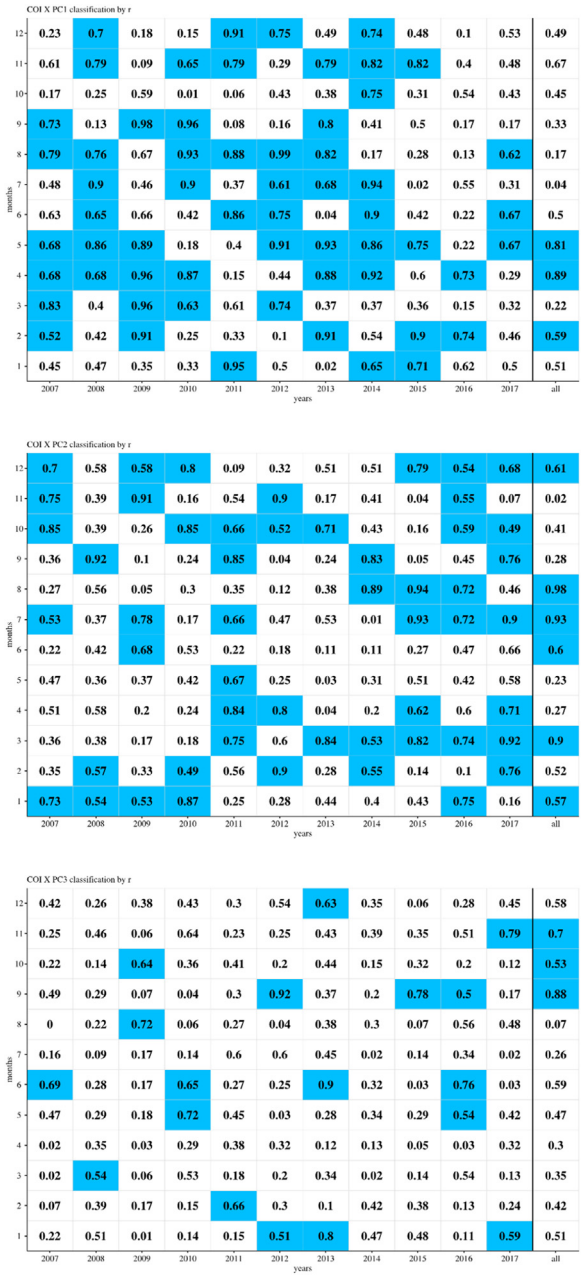

Fig. 4. Correlation coefficients between the Sq_{IQD} and PCs for X (single classification). Numbers show correlation coefficients between the Sq_{IQD} and PC1 (top), PC2 (middle) and PC3 (bottom) series for the X component for different months (Y-axis) and different years (X-axis). Blue tiles mark PCs classified as Sq (single classification using r).

Table 3
Number of months with PCs classified as Sq vs mean values of the solar/geomagnetic indices. Correlation coefficients between the number of months per year with PCs for X classified as Sq and mean annual values of the solar/geomagnetic indices. Only $|r| \geq 0.3$ are shown, with *p values* in parentheses (only *p values* ≤ 0.2 are shown).

		Geomagnetic indices				Solar indices	
		AE	ap	Kp	Dst	R	F10.7
Single classification	PC1	−0.48	−0.53 (0.18)	−0.47		0.32	0.34
	PC2	0.34	0.39	0.38			
Combined classification	PC1+PC2				0.4	−0.38	−0.33
	PC1+PC3				−0.38	0.53	0.57

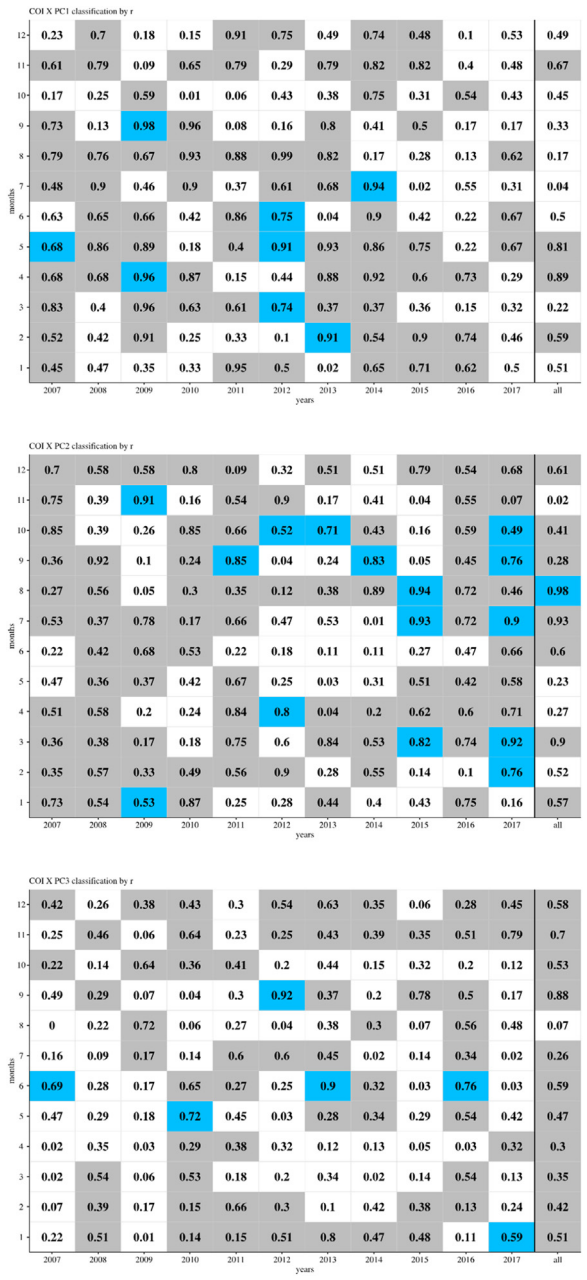

Fig. 5. Correlation coefficients between the Sq_{IQD} and PCs for X (combined classification). Same as Fig. 4 but with the combined classification allowed. gray tiles mark PCs classified as Sq in pairs with another PC (see also Fig. 6).

SD and Dst) became dominant and will be associated with PC1 while Sq will be rather filtered to the mode 2 or even to the mode 3. Similar behavior was shown by Golovkov et al. [18,19] and Golovkov and Zvereva [16,17] for data obtained at other latitudinal zones and for other decades. On the other hand, the increase of the solar activity results in a more intense flux of the solar UV radiation and, consequently, in higher ionization of the ionosphere, stronger Sq vortex and higher amplitude of the Sq geomagnetic variation. Unfortunately, the found dependence cannot be used to automatically define which PC is classified as Sq.

Thus, for the X component, PCA cannot be used as a simple method to extract Sq variations without further classification of the modes, and a comparison to a reference series is needed to identify PC that represents Sq variation.

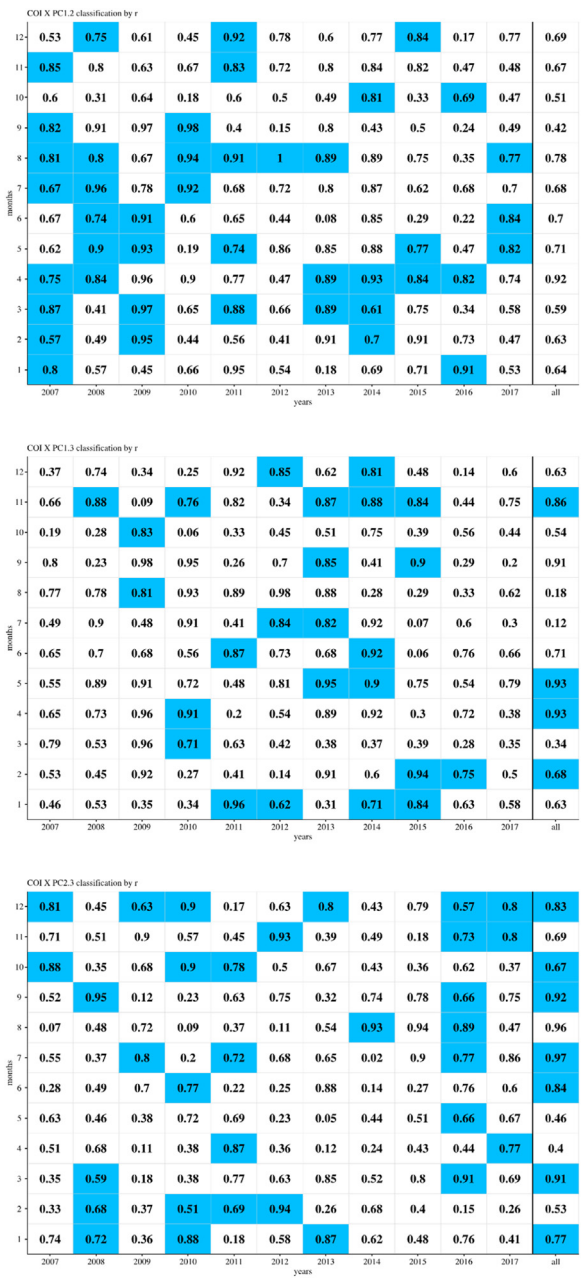

Fig. 6. Correlation coefficients between the Sq_{IQD} and sums of PCs for X (combined classification). Same as Fig. 4 but for sums of PCs: top – PC1+PC2, middle – PC1+PC3, bottom – PC1+PC3.

Adaptation of PCA for an automatic extraction of Sq for the X component

As was shown above, for the X component is it impossible to automatically extract Sq variation using just PCA. A certain reference series is needed to be compared with PCs to classify one of those PCs or a sum of PCs as Sq_{PCA}. In this work, we tested two types of reference series: (1) the mean Sq_{IQD} series obtained from geomagnetic field observations for a long time interval and (2) simulations of the ionospheric part of the geomagnetic field using geomagnetic field models.

Another significant problem of the application of PCA to extract Sq-type variation from the series of the X component is that the high rate of the PCs' classification shown above was obtained for a quite low threshold ($|r| \geq 0.45$). As one can see from Figs. 4-6, higher values of the threshold would significantly decrease the number of the identified PCs. On the other hand, the visual analysis of the corresponding PCs and Sq curves shows that in some cases of the low r values the compared series show quite similar variations,

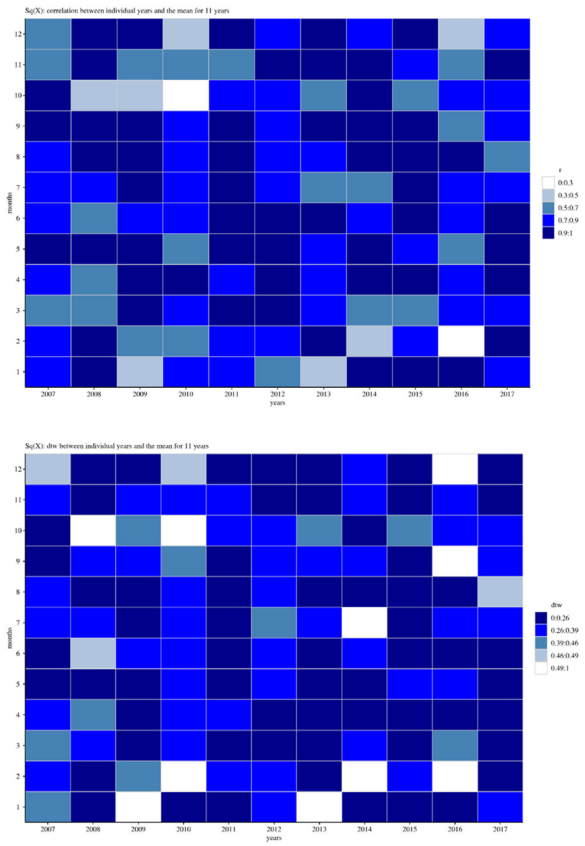

Fig. 7. Correlation coefficients r (top) and dtw values (bottom) between Sq_{IQD} (calculated for a particular month using data for individual years) and $Sq_{IQD\ allY}$ (calculated for a particular month using data for all years from 2007 to 2017). color shows the corresponding ranges of r and dtw. .

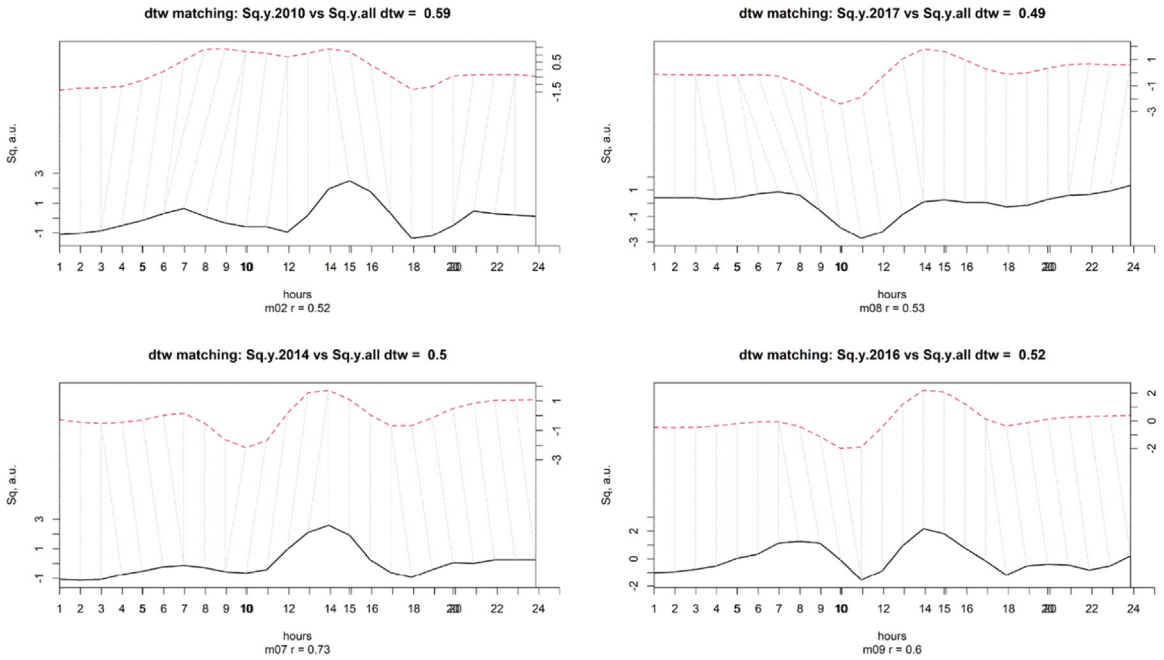

Fig. 8. Examples of the DTW matching between the Sq_{IQD} (black lines) and $Sq_{IQD\ allY}$ (red lines) when high r does not correspond to low dtw. Corresponding r and dtw values are shown below and above the plots, respectively.

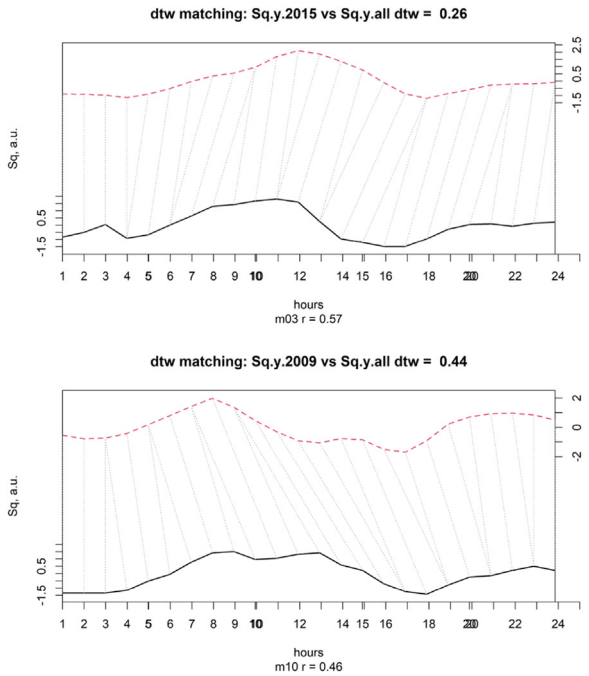

Fig. 9. Examples of the DTW matching between the Sq_{IQD} (black lines) and $Sq_{IQD\ allY}$ (red lines) when dtw values are much lower than expected for respective r. Corresponding r and dtw values are shown below and above the plots, respectively.

and the low values of r are related to local compressions and stretches of one of the series relatively to another. Thus, we need a different metric as a base for the classification. Here we tested the DTW distance as a metric of the similarity of the studied series.

Mean Sq_{IQD} as a reference series

Sq_{IQD} series obtained for the same geomagnetic station or observatory seem to be a good choice for a reference series because they automatically incorporate features of the Sq variation associated with a particular location (shape of the daily curve, characteristic seasonal variations etc.). However, as was shown above, the Sq_{IQD} series obtained for an individual month and individual year cannot be used as reliable reference series. Firstly, the automatic usage of PCA implies that a reference series already exists. Secondly, Sq_{IQD} calculated for a particular month and a particular year is strongly affected by the level of geomagnetic activity of those 5 IQDs that were used to calculate it. Thirdly, the position and the shape of the Sq current vortex in the ionosphere depends on the conditions in the upper atmosphere (wind strength, amplitude of waves and tides etc.). Thus, individual features of the vortex during the selected 5 IQDs are preserved in the Sq_{IQD} of the individual months. This can be an essential flaw for an analysis of the data obtained at observatories as COI when the position of the station to the north or the south to the Sq current vortex focus during selected days changes and affects the Sq_{IQD} variation's curve dramatically. On the other hand, averaging the Sq_{IQD} variation series obtained for a certain month but for several years may reduce the effect of individual features caused by the varying geomagnetic and atmospheric conditions. Therefore, we tested the $Sq_{IQD\ allY}$ series, which were calculated for a particular month using data for all years from 2007 to 2017, as one of the reference series for the PCs' classification.

Overall, for most of the studied series (months from January to December, years from 2007 to 2017) there is a strong correlation between Sq_{IQD} and $Sq_{IQD\ allY}$, however, for some months (mostly autumn-winter months with weak Sq current vortex) there is a large variability in the Sq_{IQD} shape resulting in a lower correlation between two types of Sq_{IQD} (individual correlation coefficients can be found in Fig. 7, top). The detailed analysis of the Sq_{IQD} and $Sq_{IQD\ allY}$ series shows that there are (1) cases of low correlation which are caused simply by shifts of maxima/minima position, and (2) cases of (relatively) high correlation that result from the similarity of the general trend but not of individual features of the compared curves. To test if the DTW analysis can perform better in these situations we calculated the dtw values for each of the corresponding pairs of the Sq_{IQD} and $Sq_{IQD\ allY}$ series (Fig. 7, bottom). In general, it seems that the DTW analysis gives a more realistic estimate of the similarity between the Sq_{IQD} and $Sq_{IQD\ allY}$ series. Some examples of the DTW matching can be found in Figs. 8 and 9: Fig. 8 gives examples of the cases when (relatively) high r values are obtained for series with similar general trends but different local features – corresponding dtw are high which means bad matching between the curves; and Fig. 9 gives examples of the cases when (relatively) low r values are obtained for series with similar features shifted locally – corresponding dtw are low which means good matching between the curves.

We used the $Sq_{IQD\ allY}$ series as reference series to classify PCs based both on the r and dtw metrics and using the combined classification option. The results (similar to Figs. 5 and 6) are shown in Figs. 10 and 11 for r classification and Figs. 12 and 13 for

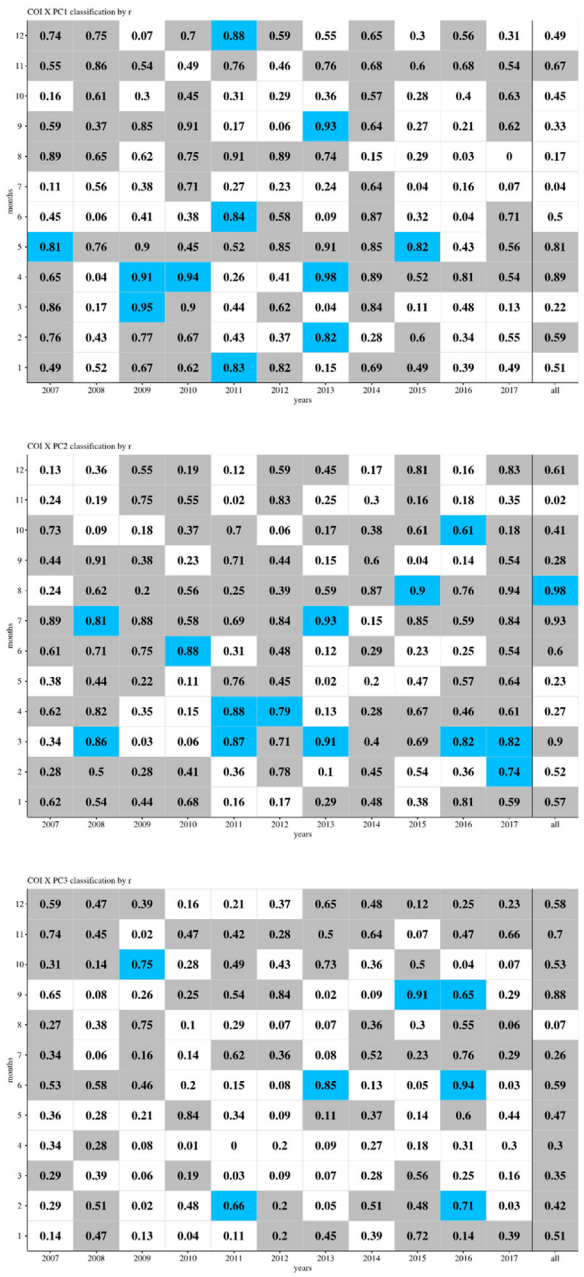

Fig. 10. Combined classification for the X component with $Sq_{IQD\ allY}$ as a reference series and r is the classification parameter: correlation coefficients (numbers) between the Sq_{IQD} and PC1 (top), PC2 (middle) and PC3 (bottom) series for different months and different years. Blue tiles mark PCs classified as Sq and gray tiles mark PCs classified as Sq in pairs with another PC (see also Fig. 11).

dtw classification. Columns 1 and 4 of Table 4 show how many different PCs or their sums were classified as Sq_{PCA} using r and dtw, respectively.

Ionospheric field models as reference series

As was mentioned above (Sec. 4.1), we used two models to simulate the ionospheric part of the geomagnetic field: CM5 and DIFI3. These modeled series were used as reference series for the PCs classification using both r and dtw metrics. The classification results (similar to Figs. 5 and 6 and 10-13) can be found in Figs. 14 and 15 for r and Figs. 16 and 17 for dtw for the DIFI3 reference series and in the Supplementary Material (SM6, Figs. S6.1-S6.2 for r and Figs. S6.3-S6.4 for dtw) for the CM5 reference series. Table 4 (columns

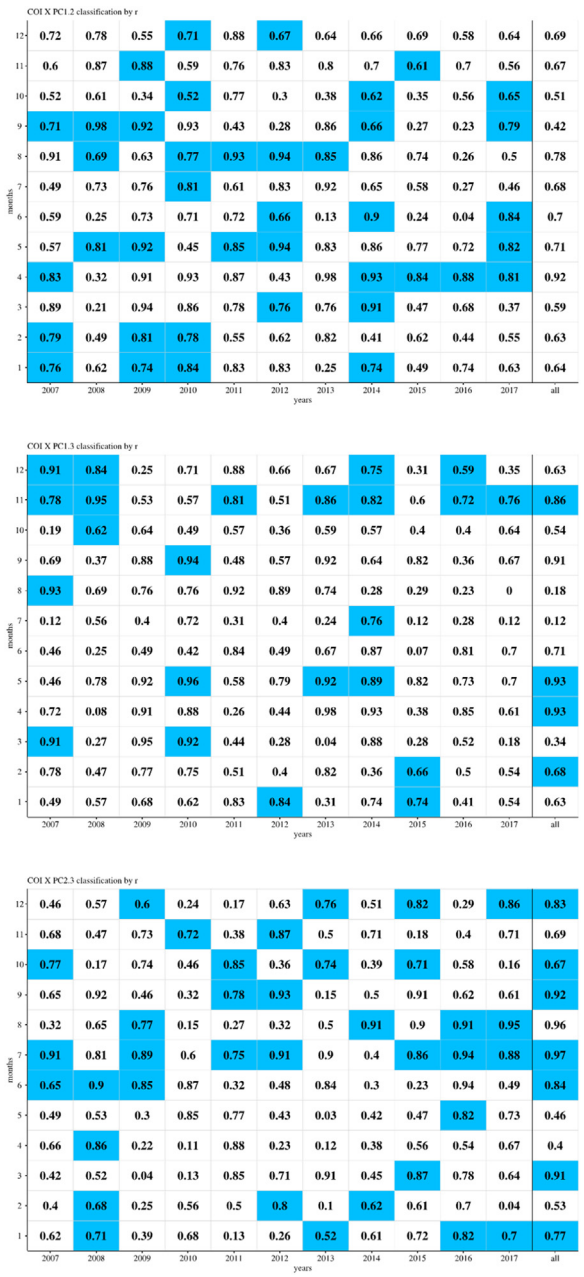

Fig. 11. Same as Fig. 10 but for a pair of PCs (top – PC1+PC2, middle – PC1+PC3, bottom – PC1+PC3).

2–3 and 5–6) shows the number of different PCs or their sums that were classified as Sq_{PCA} using different models and different metrics.

As one can see from Table 4, dtw allows the classification of more series than r. Most of the series that were identified as Sq_{PCA} are sums of PCs: the sum PC1+PC2 is most often classified as Sq_{PCA} using all studied reference series and both metrics; it is followed by the sums PC2+PC3 and PC1+PC3 which are more or less equally often classified as Sq_{PCA}.

While the differences between the performance of the analyzed reference series and metrics are not great, we recommend the DIFI3 model as a reference series and the dtw as a metric to be used to identify PCs that correspond to the Sq-type variations of the X component of the geomagnetic field.

As the final step, we compared the Sq-type variations extracted from the data using PCA and identified using combined classification with DIFI3 as a reference series and dtw as a metric to all reference series (Sq_{IQD}, $Sq_{IQD\ allY}$ and Sq_{DIFI3}). The mean and

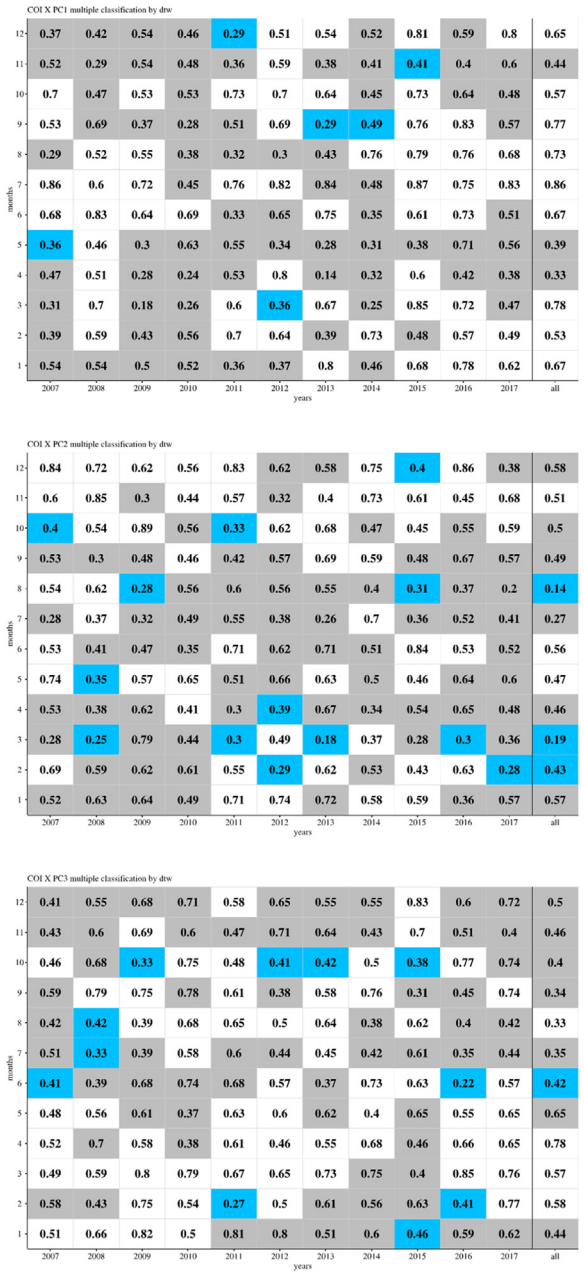

Fig. 12. Combined classification for the X component with $Sq_{IQD\ allY}$ as a reference series and dtw is the classification parameter: dtw values (numbers) between the Sq_{IQD} and PC1 (top), PC2 (middle) and PC3 (bottom) series for different months and different years. Blue tiles mark PCs classified as Sq and gray tiles mark PCs classified as Sq in pairs with another PC (see also Fig. 13).

median correlation coefficients between Sq_{PCA} and the reference series are $r_{mean} \sim 0.75$ and $r_{median} \sim 0.65$; the individual correlation coefficients can be found in the Supplementary Material (SM7).

Fig. 18 shows two examples of the comparisons of Sq_{PCA} and the reference series. These plots also allow comparing Sq_{PCA} variations obtained for a certain month using the data only for a certain year (black lines) and for the "all years" series (gray lines). It seems that for months near equinoxes and solstices (February-March, May-June, August-October, December) it is better to use only data for the studied year to obtain a Sq_{PCA}, whereas for other months it is better to use data for this month but for several years of observations (11 in our case), however, this conclusion still needs to be confirmed on longer time series or data from other locations.

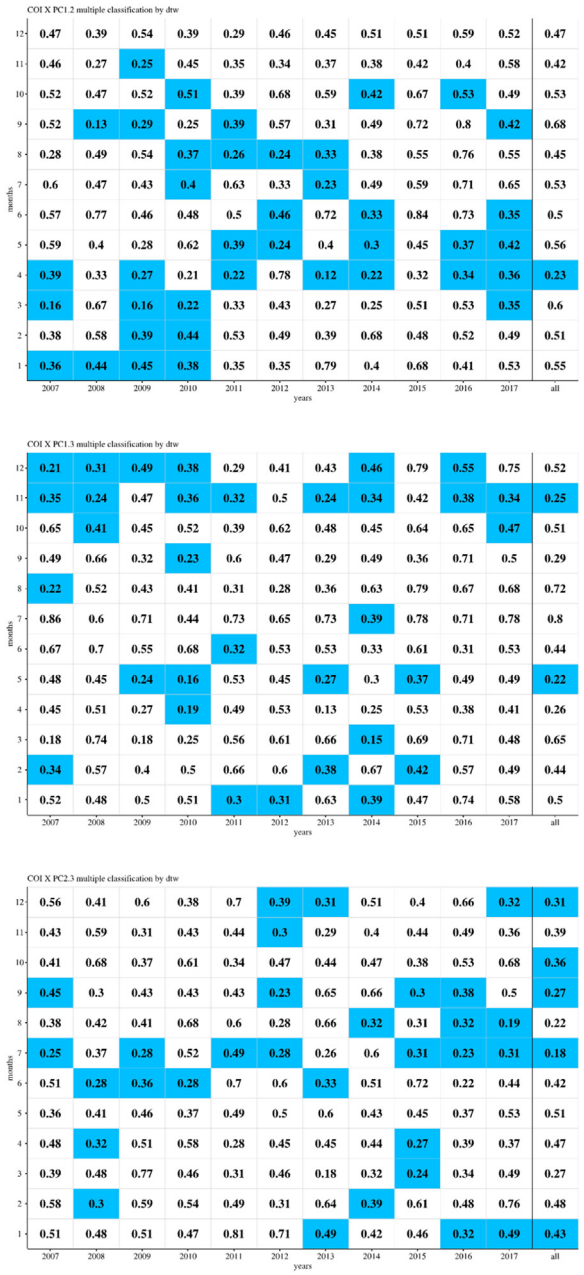

Fig. 13. Same as Fig. 12 but for a pair of PCs (top – PC1+PC2, middle – PC1+PC3, bottom – PC1+PC3).

Recommendations and comments on the usage of PCA as a method to extract Sq variation from the geomagnetic filed measurements

In this work we presented a detailed analysis of the performance of the principal component analysis (PCA) as a tool to extract Sq variation from the geomagnetic field observations (X, Y and Z components) made at a mid-latitudinal station (Coimbra Magnetic Observatory, Portugal).

The studied time interval is from January 2007 to December 2017; the time resolution is 1 h. The data were analyzed individually for all 12 months. The geomagnetic field components were analyzed separately.

The PCA modes were compared to Sq variation obtained for the same month using the standard approach based on the calculation of the mean daily variations using 5 international quiet days (IQD), Sq_{IQD}, using both the correlation and the dynamic time warping (DTW) analyses. Also, the CM5 and DIFI3 models were used to generate reference series of the ionospheric field.

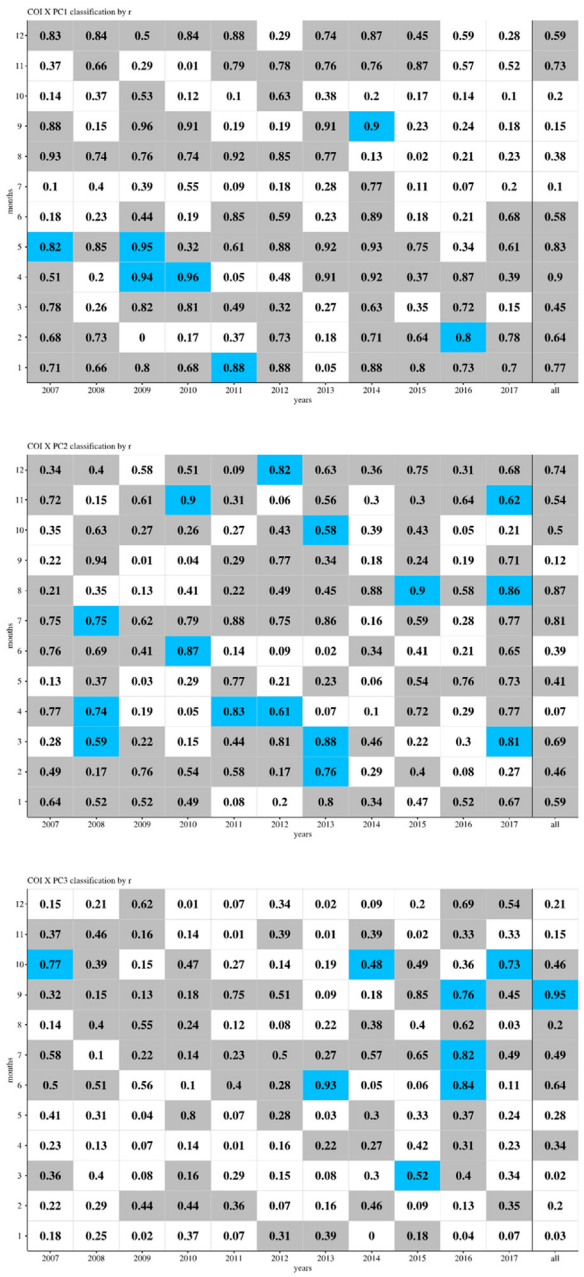

Fig. 14. Combined classification for the X component with Sq_{DIFI3} as a reference series and r is the classification parameter: correlation coefficients (numbers) between the Sq_{DIFI3} and PC1 (top), PC2 (middle) and PC3 (bottom) series for different months and different years. Blue tiles mark PCs classified as Sq and gray tiles mark PCs classified as Sq in pairs with another PC (see also Fig. 15).

Only the first three PCA modes were analyzed.

Based on the correlation or DTW analyses some PCs were classified as Sq_{PCA}. Two approaches were tested: only one PC can be classified as Sq_{PCA} (single classification) or also a weighted sum of PCs can be classified as Sq_{PCA} (combined classification).

The number of the PCs classified as Sq and their order were analyzed in relation to the component (X, Y or Z), season (mean seasonal variations through a year) and year (mean decadal variations during the 11-year solar/geomagnetic activity cycle).

Recommendations for the use of PCA to extract Sq:

1. It was found that for the Y and Z components the Sq variation is always filtered to the first PCA mode (PC1). Thus, PCA can be used to extract Sq variations from the observations of the Y and Z geomagnetic field components without any additional procedures.

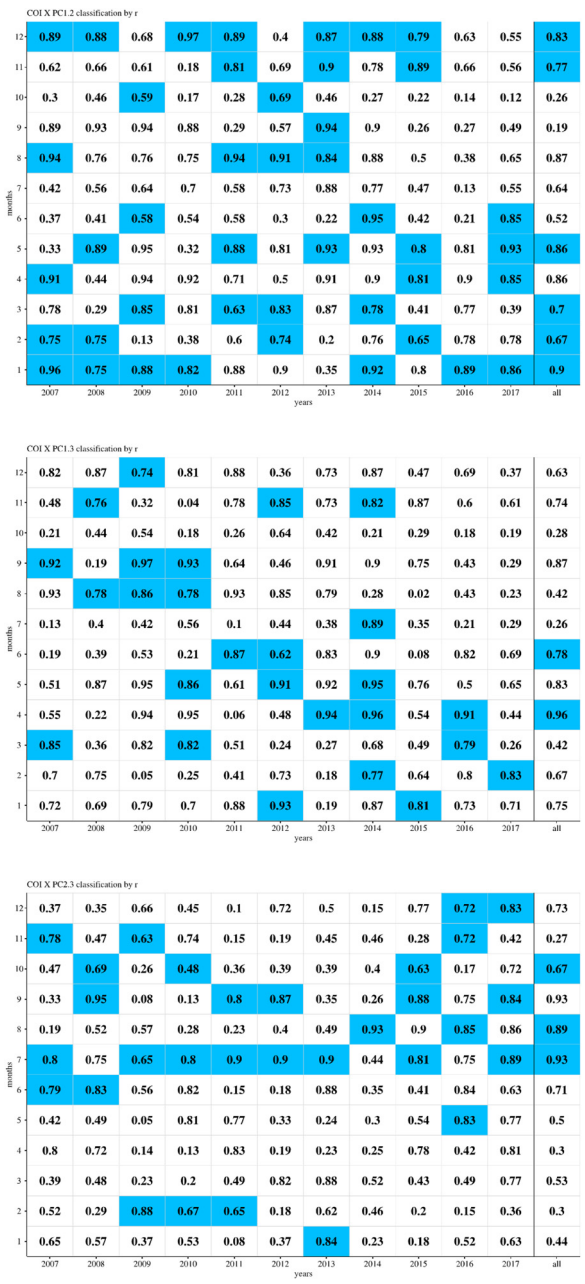

Fig. 15. Same as Fig. 14 but for a pair of PCs (top – PC1+PC2, middle – PC1+PC3, bottom – PC1+PC3).

1.1. For the studied time interval and for the mid-latitudinal geomagnetic data we found no limitations or constraints to the usage of PCA as a tool to extract Sq variations from the Y and Z components' series.
1.2. There are no significant differences in PC1s obtained for a certain month for an individual year and several years (11 years in our case), thus the input data set with the length of ~1 month (not necessary coinciding with a calendar month) will be sufficient to extract reliable Sq variation from the series of the Y and Z component.
2. The classification of PCs obtained for the X component is much more complicated, probably, due to the higher contribution of the geomagnetic disturbances in the variability of the X component at the middle latitudes:
2.1. All three first PCs can be classified as Sq.
2.2. No patterns in the classification rate of different PCs related to the season or the level of the solar/geomagnetic activity were found.

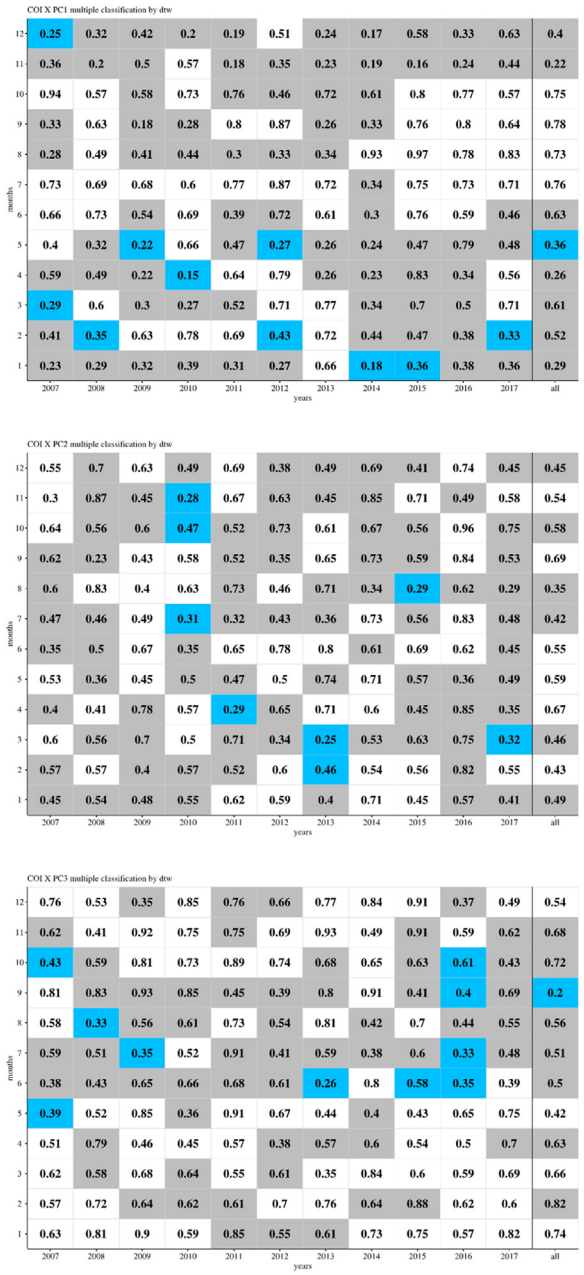

Fig. 16. Combined classification for the X component with Sq_{DIFI3} as a reference series and dtw is the classification parameter: dtw values (numbers) between the Sq_{DIFI3} and PC1 (top), PC2 (middle) and PC3 (bottom) series for different months and different years. Blue tiles mark PCs classified as Sq and gray tiles mark PCs classified as Sq in pairs with another PC (see also Fig. 17).

2.3. Thus, PCA still can be used to extract Sq variation from the observations of the X component, but further analysis, for example, a comparison to a set of reference curves either obtained from the data analysis or generated using models, is always needed to classify PCs of the X component.

2.3.1. Two types of reference series were tested: the Sq_{IQD} series obtained for each month using all years of observations (11 years), and the ionospheric magnetic field modelled using the CM5 and DIFI3 models.

2.3.2. The reference series were compared to PCs using two metrics: the correlation coefficient r and the DTW distance (dtw).

2.3.3. In general, all reference series and both metrics performed well, however only the combination of the DIFI3 model as a reference series and the dtw metric allowed us to identify Sq_{PCA} for all analyzed series.

2.3.4. We recommend to use the DIFI-class models to generate ionospheric field.

2.3.5. We recommend to use the dtw metric to classify Sq_{PCA}.

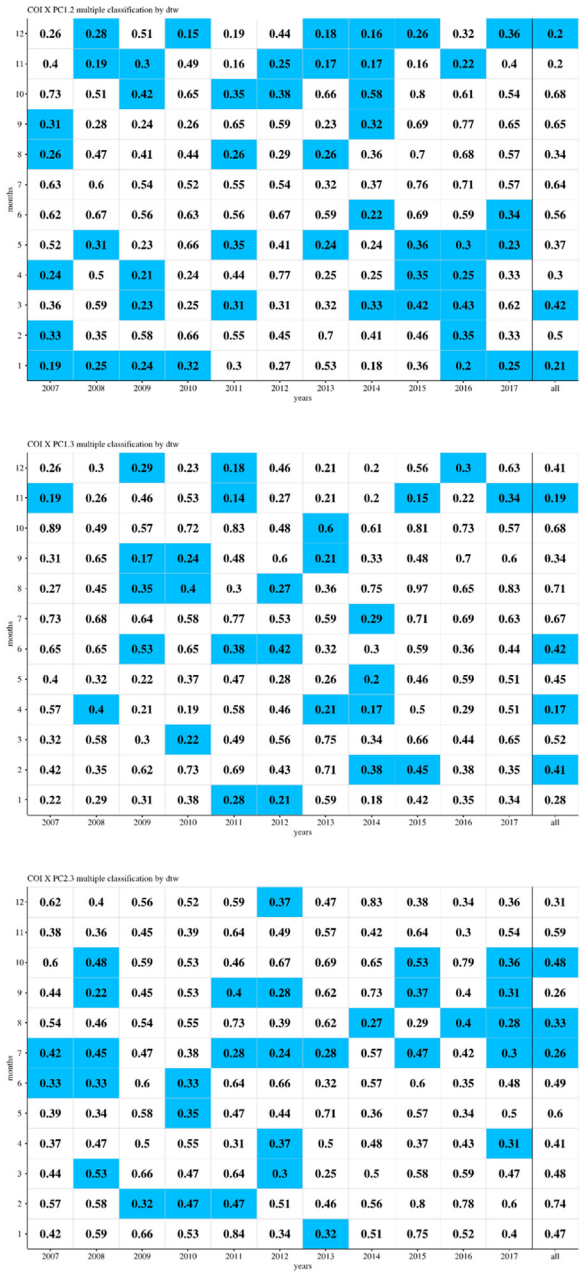

Fig. 17. Same as Fig. 16 but for a pair of PCs (top – PC1+PC2, middle – PC1+PC3, bottom – PC1+PC3).

3. The reference series may be successfully used to solve a "sign ambiguity problem" for PCs that are obtained using SVD. There is no general way to solve the sign ambiguity, and comparison to a reference series can be an easy way to do so. This may apply not only to the X but also to the Y and Z series.

Finally, we summarize the main advantages and a disadvantage/constrain of the usage of the PCA method to extract Sq-type variation from the observations of the geomagnetic field components that we found:

PCA advantages:

1. With PCA there is no need to estimate the (relative) level of geomagnetic activity of different days of an analyzed month (or another time interval of a comparable length) to find geomagnetically quiet days (e.g., international or local quiet days). All available days of an analyzed time interval can be used.
2. It is well known that Sq_{IQD} can be contaminated by the disturbance field [39] since not all IQDs of a certain month could be quiet in the absolute sense. However, since PCA is applied to the month-long time interval, this method may allow extracting the

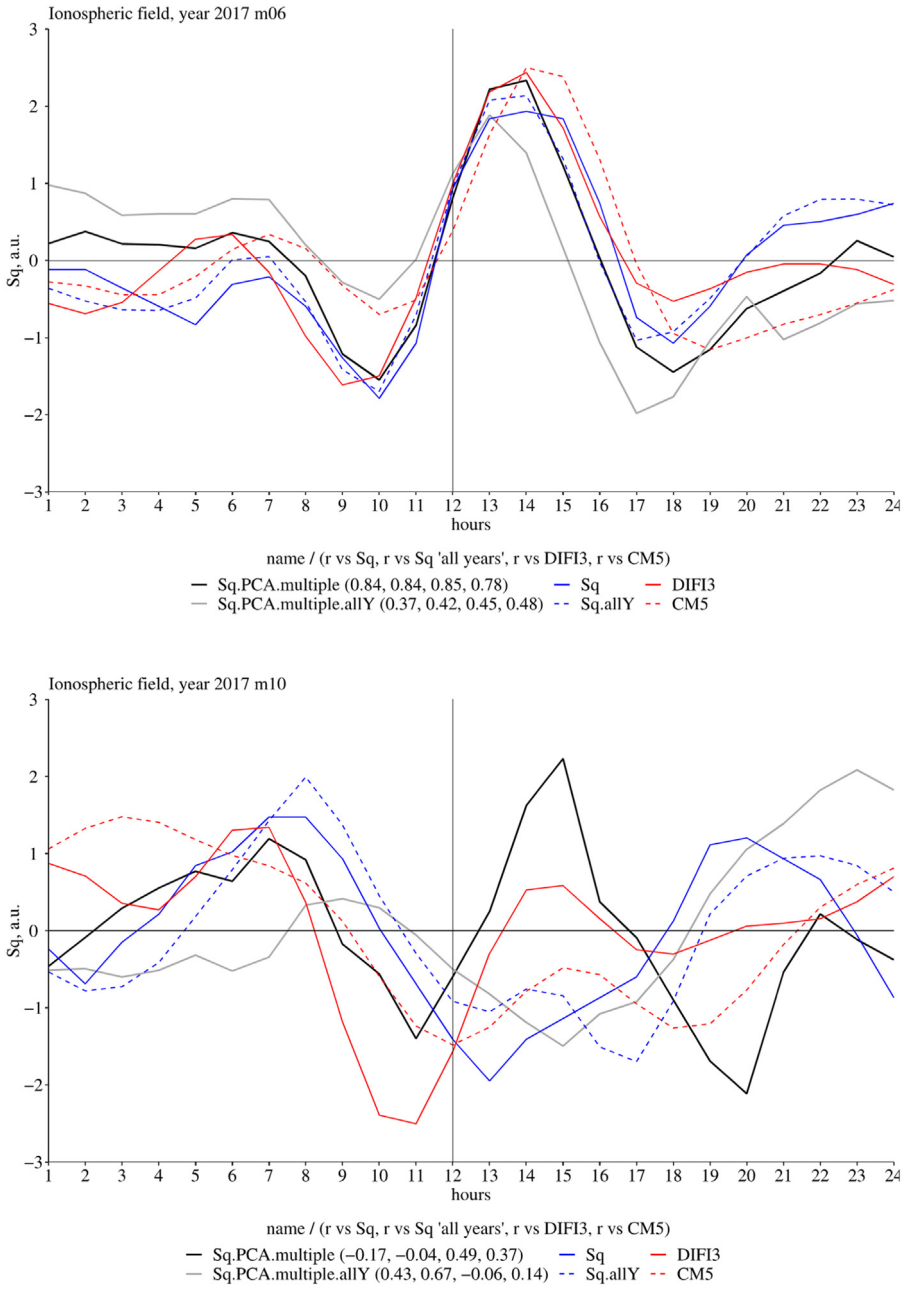

Fig. 18. Examples of different types of Sq for the X component. Sq-type variations observed or predicted for June (top) and October (bottom) of 2017: Sq_{PCA} for June or October of 2017 – black lines; Sq_{PCA} for June or October of "all years" – gray lines; Sq_{IQD} for June or October of 2017 – blue solid lines, $Sq_{IQD\ allY}$ for June or October of "all years" – blue dashed lines; Sq_{DIFI3} for June or October – red solid line; Sq_{CM5} for June or October – red dashed line. Corresponding correlation coefficients between Sq_{PCA} and reference series are shown below the plots: ($r_{vs\ Sq\_IQD:indY}$, $r_{vs\ Sq\_IQD\_allY}$, $r_{vs\ Sq\_DIFI3}$, $r_{vs\ Sq\_CM5}$).

Sq variation that has less contribution from the disturbance field. Thus, PCA allows to minimize the effect of the geomagnetic activity during individual days and to obtain the Sq variation that is more typical for the studied month.

3. The shape of the Sq variation observed at a certain location depends on the position of the geomagnetic observatory relative to the focus of the Sq current vortex. Thus, under certain circumstances, Sq_{IQD} could reflect not the general conditions in the ionosphere and the upper atmosphere during a certain month but some individual features (position of the focus and the shape) of the vortex during IQDs. On the other hand, while PCA is applied to the month-long time interval, this method may allow to minimize the effect of the individual days and to obtain a more "climatological" Sq variation.

Table 4
PCs of X or their sums classified as Sq. Number (out of 144) of PCs or their sums classified as Sq using different reference series ($Sq_{IQD\ allY}$, CM5 and DIFI3) and different metrics (*r* and *dtw*).

	dtw			r		
	$Sq_{IQD\ allY}$	CM5	DIFI3	$Sq_{IQD\ allY}$	CM5	DIFI3
PC1	6	10	11	11	9	7
PC2	16	17	8	14	15	15
PC3	12	17	11	7	8	9
PC1+PC2	40	52	49	40	68	49
PC1+PC3	34	31	31	27	26	28
PC2+PC3	35	16	34	43	16	33
none	**1**	**1**	**0**	2	2	3

4. PCA allows the estimation of the variance fraction associated with a mode that is classified as Sq.
5. The EOF functions available for each of the PCs for each day of the analyzed time interval permit reconstructing the amplitudes of the Sq variation for each day individually allowing the assessment of its day-to-day variability.

PCA disadvantage:

1. The automatic classification of PCs is not always straightforward. For the Y and Z components, the Sq variations seem to be always filtered to PC1, however for the X component an additional manual or automatic classification is needed (e.g., by comparing PCs to a set of reference curves as proposed above).

The list of the abbreviations used in this paper can be found in the Supplementary Material (SM8).
The list of the datasets and software used to validate method's performance can be found in the Supplementary Material (SM9).

Funding

CITEUC is funded by the National Funds through FCT (Foundation for Science and Technology) projects UID/00611/2020 and UIDP/00611/2020.
IA is supported by FCT through the research grants UIDB/04434/2020 and UIDP/04434/2020.
This study is a contribution to the MAG-GIC project (PTDC/CTA-GEO/31744/2017), and RR is funded through this project.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Anna Morozova: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft. **Rania Rebbah:** Data curation, Investigation, Software, Visualization, Writing – review & editing.

Data Availability

The data are open access available at WDC (Edinburg). The link is provided in the text

Acknowledgments

AM is thankful to Dr. T. Giorgino for the development of the "dtw" R package (https://dynamictimewarping.github.io/, http://dtw.r-forge.r-project.org/).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2023.101999.

References

[1] C. Amory-Mazaudier, On the electric current systems in the Earth's environment some historical aspects Part I: external part/ionosphere/quiet variation, GEOACTA 21 (1994) 1–15.

[2] Amory-Mazaudier, C. (2001) On the electric current systems in the Earth's environment some historical aspects. Part II: external part/ionosphere/disturbed variation from the IAGA Assembly in Hanoï Vietnam.
[3] C. Amory-Mazaudier, Electric current systems in the earth's environment, J. Space Res. 8 (2009) 178–255 Niger.
[4] F. Anad, C. Amory-Mazaudier, M. Hamoudi, S. Bourouis, A. Abtout, E. Yizengaw, Sq solar variation at Medea Observatory (Algeria), from 2008 to 2011, Adv. Space Res. 58 (9) (2016) 1682–1695, doi:10.1016/j.asr.2016.06.029.
[5] S.K. Bhardwaj, September Rao, Longitudinal inequalities in Sq current system along 200-2100 E meridian, J. Ind. Geophys. Union 20 (5) (2016) 462–471.
[6] S.K. Bhardwaj, P.S. Rao, B. Veenadhari, Abnormal quiet day variations in Indian region along 75 E meridian, Earth Planets Space 67 (1) (2015) 1–15.
[7] H. Björnsson, S.A. Venegas, A manual for EOF and SVD analyses of climatic data, CCGCR Rep. 97 (1) (1997) 112–134.
[8] S. Chapman, J. Bartels, Geomagnetism, Oxford University Press, Oxford, 1940.
[9] G.X. Chen, W.Y. Xu, A.M. Du, Y.Y. Wu, B. Chen, X.C. Liu, Statistical characteristics of the day-to-day variability in the geomagnetic Sq field, J. Geophys. Res. 112 (A6) (2007), doi:10.1029/2006JA012059.
[10] A. Chulliat, P. Vigneron, G. Hulot, First results from the swarm dedicated ionospheric field inversion chain swarm science results after two years in Space 1. Geomagnetism, Earth Planets Space 68 (1) (2016), doi:10.1186/s40623-016-0481-6.
[11] A. Chulliat, P. Vigneron, E. Thébault, O. Sirol, G. Hulot, Swarm SCARF dedicated ionospheric field inversion chain, Earth Planets Space 65 (11) (2013) 1271–1283, doi:10.5047/eps.2013.08.006.
[12] P. De Michelis, R. Tozzi, G. Consolini, Principal components' features of mid-latitude geomagnetic daily variation, Ann. Geophys. 28 (2010) 2213–2226, doi:10.5194/angeo-28-2213-2010.
[13] P. De Michelis, R. Tozzi, A. Meloni, On the terms of geomagnetic daily variation in Antarctica, Ann. Geophys 27 (2009) 2483–2490.
[14] W. Ebisuzaki, A method to estimate the statistical significance of a correlation when the data are serially correlated, J. Clim. 10 (9) (1997) 2147–2153.
[15] T. Giorgino, Computing and visualizing dynamic time warping alignments in R: the dtw package, J. Stat. Softw. 31 (1) (2009) 1–24, doi:10.18637/jss.v031.i07.
[16] V.P. Golovkov, T.I. Zvereva, Expansion of geomagnetic variations within a year in natural orthogonal components, Geomagn. Aeron. 38 (1998) 368–372.
[17] V.P. Golovkov, T.I. Zvereva, The space-time pattern of midlatitude geomagnetic variations, Geomagn. Aeron. 40 (2000) 84–92.
[18] V.P. Golovkov, N.E. Papitashvili, Y.S. Tyupkin, E.P. Kharin, Separation of geomagnetic field variations into quiet and disturbed components by the method of natural orthogonal components, Geomagn. Aeron. 18 (1978) 342–344.
[19] V.P. Golovkov, V.O. Papitashvili, N.E. Papitashvili, Automatic calculation of K indices using the method of natural orthogonal components, Geomagn. Aeron. 29 (1989) 514–517.
[20] G.V. Haines, J.M. Torta, Determination of equivalent current sources from spherical cap harmonic models of geomagnetic field variations, Geophys. J. Int. 118 (3) (1994) 499–514.
[21] A. Hannachi, I.T. Jolliffe, D.B. Stephenson, Empirical orthogonal functions and related techniques in atmospheric science: a review, Int. J. Climatol. 27 (9) (2007) 1119–1152, doi:10.1002/joc.1499.
[22] I. Maslova, P. Kokoszka, J. Sojka, L. Zhu, Estimation of Sq variation by means of multiresolution and principal component analyses, J. Atmosph. Solar-Terrestrial Phys. 72 (7–8) (2010) 625–632, doi:10.1016/j.jastp.2010.02.005.
[23] M. Menvielle, About the scalings of K indicesfrom, IAGA News 20 (1981) 110–111.
[24] A.L. Morozova, R. Rebbah, P. Ribeiro, Datasets of the Solar Quiet (Sq) and Solar Disturbed (SD) variations of the geomagnetic field at the Coimbra Magnetic Observatory (COI) obtained by different methods, Data Brief, 37C, 2021, doi:10.1016/j.dib.2021.107174.
[25] A.L. Morozova, R. Rebbah, P. Ribeiro, Datasets of the Solar Quiet (Sq) and Solar Disturbed (SD) variations of the geomagnetic field at a midlatitudinal station in europe obtained by different methods, Mendeley Data, V1, 2021 doi: 10.17632/jcmdrm5f5x.1, doi:10.17632/jcmdrm5f5x.1.
[26] A.L. Morozova, P. Ribeiro, M.A. Pais, Correction of artificial jumps in the historical geomagnetic measurements of Coimbra Observatory, Portugal, Annal. Geophys. 32 (1) (2014) 19–40, doi:10.5194/angeo-32-19-2014.
[27] A.L. Morozova, P. Ribeiro, M.A. Pais, Homogenization of the historical series from the Coimbra Magnetic Observatory, Portugal, Earth Syst. Sci. Data 13 (2021) 809–825, doi:10.5194/essd-13-809-2021.
[28] M. Piersanti, T. Alberti, A. Bemporad, F. Berrilli, R. Bruno, V. Capparelli, V. Carbone, C. Cesaroni, G. Consolini, A. Cristaldi, A. Del Corpo, Comprehensive analysis of the geoeffective solar event of 21 June 2015: effects on the magnetosphere, plasmasphere, and ionosphere systems, Sol Phys 292 (11) (2017) 169, doi:10.1007/978-94-024-1570-4_12.
[29] Rangarajan, G.K. and Murty, A.V.S., (1980) Scaling K-indices without subjectivity From IAGA news, 19, 112–118.
[30] T.J. Sabaka, N. Olsen, R.A. Langel, A comprehensive model of the quiet-time, near-Earth magnetic field: phase 3, Geophys. J. Int. 151 (1) (2002) 32–68, doi:10.1046/j.1365-246X.2002.01774.x.
[31] T.J. Sabaka, L. Tøffner-Clausen, N. Olsen, C.C. Finlay, CM6: a comprehensive geomagnetic field model derived from both CHAMP and Swarm satellite observations, Earth Planets Space, 72 (2020) 1–24, doi:10.1186/s40623-020-01210-5.
[32] J. Shlens, in: A Tutorial on Principal Component Analysis Center For Neural Science, New York University New York City, NY, 2009, pp. 10003–10660.
[33] R. Stening, T. Reztsova, L.H. Minh, Day-to-day changes in the latitudes of the foci of the Sq current system and their relation to equatorial electrojet strength, J. Geophys. Res. 110 (A10) (2005) A10308, doi:10.1029/2005JA011219.
[34] R.J. Stening, The shape of the Sq current system, Annal. Geophys. 26 (7) (2008) 1767–1775.
[35] M. Takeda, Features of global geomagnetic Sq field from 1980 to 1990, J. Geophys. Res. 107 (A9) (2002) SIA–SI4.
[36] E. Thébault, P. Vigneron, B. Langlais, G. Hulot, A Swarm lithospheric magnetic field model to SH degree 80 Swarm Science Results after two years in Space 1. Geomagnetism, Earth Planets Space 68 (1) (2016), doi:10.1186/s40623-016-0510-5.
[37] Y.Y. Wu, W.Y. Xu, G.X. Chen, B. Chen, X.C. Liu, The evolution characteristics of geomagnetic disturbances during geomagnetic storm, Chin. J. Geophys. 50 (1) (2007) 1–11.
[38] W.Y. Xu, Y. Kamide, Decomposition of daily geomagnetic variations by using method of natural orthogonal component, J. Geophys. Res. 109 (A5) (2004), doi:10.1029/2003JA010216.
[39] Y. Yamazaki, A. Maute, Sq and EEJ—a review on the daily variation of the geomagnetic field caused by ionospheric dynamo currents, Space Sci. Rev. 206 (1–4) (2017) 299–405, doi:10.1007/s11214-016-0282-z.