Regularization of spherical cap harmonics

M. Korte and R. Holme*

GeoForschungsZentrum Potsdam, Telegrafenberg, D-14473 *Potsdam, Germany. E-mail: monika@gfz-potsdam.de*

Accepted 2002 October 31. Received 2002 September 19; in original form 2002 May 16

SUMMARY

Spherical cap harmonic analysis has become a well-known technique for regional modelling of fields that can be expressed as the gradient of a scalar potential, such as for example the geomagnetic field and its secular variation. Up to now, the method has been regularized by a purely statistical technique: coefficients that are considered to be statistically insignificantly small are simply set to zero. This method lacks physical justification and ignores resolution; individual coefficients may be small but well-resolved, while coefficients with a large value may be poorly resolved. We implement the more physical regularization technique of minimizing a certain feature of the field, for example the field strength over the cap surface, analogous to the regularization techniques widely used in global spherical harmonic analysis. The mathematical difference between spherical cap harmonics (SCHA) and global spherical harmonics analyses (SHA) lies in the basis functions. While these are completely orthogonal in SHA, this is not the case in SCHA. This leads to the existence of certain linear combinations of coefficients that hardly contribute to the field, which makes the statistical rejection criterion meaningless. With the physical regularization the individual coefficients become meaningful, as we show by modelling the secular variation from a data set of 30 years of European observatory measurements, repeat station and ground vector surveys.

Key words: magnetic field modelling, secular variation, spherical cap harmonic analysis.

1 INTRODUCTION

Geomagnetic secular variation, the slow change of the Earth's main magnetic field originating in the fluid core, is a topic of frequent studies because it can provide information concerning the dynamics of the core. The observable length-scales of secular variation at the Earth's surface are of several thousand km, as the field is geometrically attenuated from upward continuation through the (assumed to be insulating) mantle. Global models obtained by spherical harmonic analyses are a popular method of visualizing and interpreting secular variation, both at the Earth's surface and at the core–mantle boundary. However, the data distribution for secular variation studies is very inhomogeneous over the Earth, limiting the spatial resolution of global models. To date, the operational time spans of geomagnetic satellites with good spatial resolution have been too short to provide long-term secular variation information. Geomagnetic observatories with good time-series of data are dense in Europe, but sparse elsewhere, particularly in the southern hemisphere. Additional repeat station surveys are carried out by several countries for the very purpose of determining secular variation. The results of such surveys are modelled on regional scales for the practical purpose of updating geomagnetic charts, or for more detailed studies of secular variation. A long-standing question has been whether medium-scale to small-scale structure (length-scale <1000 km) exists in secular variation data. While it is unlikely for such features to have come from the Earth's core, they could reveal lithospheric induction anomalies on comparatively large scales. Mundt (1973, 1981) claimed secular variation anomalies in Europe visible in observatory data. Mundt & Porstendorfer (1977, 1978) suggested a large-scale electrical conductivity anomaly in the Earth's upper mantle as the cause of the observed anomaly, and Porstendorfer *et al.* (1979) concluded from model calculations that lithospheric conduction anomalies should be detectable in secular variation observations. Studying the temporal change of the magnetic components from European observatory records in several frequency bands, Alldredge (1983) also suggested temporally varying magnetic anomalies, i.e. secular variation anomalies, on the scale of a few thousand kilometres. The problem with all of these studies, however, is that although the spatial density of observatories is highest in Europe, it is still sparse with respect to the length-scale of the effects studied. Complementing observatory data with repeat station data seems to be the obvious solution to this problem, but the lower quality of repeat station data compared with observatory data causes new difficulties here. In part because of these difficulties, Korte & Haak (2000) find that the data are unable to support the earlier claims of anomalous secular variation.

The simplest common method of regional field or secular variation modelling is to model each field component separately with a spatial polynomial. As an alternative, Haines (1985a) introduced

*Now at: Department of Earth Sciences, University of Liverpool, UK. E-mail: holme@liv.ac.uk spherical cap harmonic analysis (SCHA), which, as for global spherical harmonic analysis (SHA) has the advantage of modelling the full vectorial field as the negative gradient of a scalar potential. With both of these regional modelling methods we have to make assumptions concerning the expected smoothness of the model by limiting the maximum polynomial degree or maximum degree of truncation of the spherical cap harmonics series, to avoid overfitting the data.

We need a method of hypothesis testing: can we fit the data within the tolerance of the estimated errors with a large-scale model or do the data really require small-scale structure? We will show that the statistical method used by Haines is physically not meaningful because it ignores the resolution of coefficients.

Haines applied a statistical regularization to the method to reduce the roughness of the models: in the stepwise regression to estimate the coefficients significance tests are applied and a coefficient is only included in the model if it is considered statistically significant (Draper & Smith 1966). Coefficients with an absolute value smaller than the square root of the chosen significance level times their standard deviation are set to zero. The method has been widely used (e.g. Torta *et al.* 1992; Haines & Torta 1994; De Santis *et al.* 1997; Haines & Newitt 1997; Kotzé 2001). However, there is no physical justification for setting small coefficients to zero. A coefficient can be small but well resolved, or conversely large but poorly resolved. We will see that in the case of spherical cap harmonics in particular, there are certain linear combinations of coefficients that are very poorly resolved. In global modelling a method of hypothesis testing is achieved by using a regularization according to physical properties of the field, for example minimizing the mean field strength over the spherical surface (Shure *et al.* 1982). The misfit of the model predictions to the data is traded off against the roughness of the model, so that the smoothest model within the estimated tolerance of the errors can be found. We applied this method to SCHA. Unlike in SHA, the basis functions in SCHA are not completely orthogonal, which makes the implementation more complicated—the damping matrix has non-zero off-diagonal elements.

Studying the magnetic field or secular variation, we are often interested not only in just a snapshot of one epoch, but want to know the temporal change. Simultaneous modelling of the spatial and temporal distribution gives additional constraints if we assume that the change in time will be continuous and smooth. Polynomials have been used for time-dependent modelling with SCHA (Haines 1985b). Again, with this method the smoothness of the models is mainly determined by the arbitrary choice of maximum degree of the polynomials. In contrast, the use of cubic splines (piecewise cubic polynomials) as basis functions, offers the possibility of physically more sensible regularizations.

With European regional survey secular variation data, we investigate three different damping norms and compare the modelling results with those obtained using Haines' statistical regularization method. The data set had been tested thoroughly and modelled for six individual 5 yr intervals between 1965 and 1995 with the latter method in an earlier study (Korte 1999; Korte & Haak 2000). The data are of limited quality because of the temporal and spatial inhomogeneity of the European regional magnetic surveys and often large errors arising from uncompensated external magnetic field influences. While imperfect, this data set is useful for investigating and emphasizing the improvements that can be obtained by regularization. For time-dependent modelling we complement the data set with annual observatory secular variation to fill in the gaps between the 5 yr intervals.

2 MODELLING METHOD

2.1 Spherical cap harmonics

In a source-free region, the magnetic field **B** can be given as the gradient of a scalar potential Φ, $\mathbf{B} = -\nabla\Phi$, with $\nabla^2\Phi = 0$. The general solution of Laplace's equation in the case of SCHA, written as a series up to degree $k_{\max}$, is given by Haines (1985a):

$$\Phi(r,\theta,\phi) = R_{\mathrm{E}} \sum_{k=0}^{k_{\max}} \sum_{m=0}^{k} \left(\frac{R_{\mathrm{E}}}{r}\right)^{n_k+1} \times \left[g_k^m \cos(m\phi) + h_k^m \sin(m\phi)\right] P_{n_k}^m(\cos\theta). \qquad (1)$$

The potential Φ depends on radius r, colatitude θ and longitude ϕ. R_{E} is the mean radius of the Earth and $\{g_k^m, h_k^m\}$ are the coefficients analogous to the Gauss coefficients in main field modelling. However, in SCHA $P^m_{n_k(cos\theta)}$ are not the usual Legendre polynomials with integer degree n and order m, but associated Legendre functions with a non-integer degree n_k, depending on the colatitude of the spherical cap boundary. The order m is still an integer, as the potential must be continuous in ϕ. The boundary condition on θ, however, is only to be able to give an arbitrary function at the cap boundary. To allow sufficient differentiability of the field to obtain the horizontal components, Haines (1985a) argued that two sets of basis functions were necessary, namely those with (generally non-integer) degree n_k so that either

$$\frac{\partial P_{n_k}^m}{\partial\theta} = 0 \qquad (2)$$

or

$$P_{n_k}^m = 0. \qquad (3)$$

As the degrees n_k also depend on the integer order m, they are ordered by the integer index k and $k_{\max}$ in eq. (1) is the maximum spatial index of the truncated series. The two sets of basis functions are denoted by the fact that for one set the difference $(k-m)$ is even and for the other $(k-m)$ is odd. All functions within one set are orthogonal, but the functions of one set are not completely orthogonal to those in the other. Haines (1985a) also gives the equations for products of the functions:

$$\int_0^{\theta_0} P_{n_j}^m(\cos\theta) P_{n_k}^m(\cos\theta)\sin\theta\, d\theta = -\frac{\sin\theta_0}{(n_k-n_j)(n_k+n_j+1)} P_{n_j}^m(\cos\theta_0)\frac{dP_{n_k}^m(\cos\theta_0)}{d\theta} \qquad (4)$$

for $(j-m)$ even and $(k-m)$ odd,

$$\int_0^{\theta_0} \left[P_{n_k}^m(\cos\theta)\right]^2 \sin\theta\, d\theta = -\frac{\sin\theta_0}{2n_k+1} P_{n_k}^m(\cos\theta)\frac{\partial}{\partial n}\frac{dP_{n_k}^m(\cos\theta_0)}{d\theta} \qquad (5)$$

for $(k-m)$ even and

$$\int_0^{\theta_0} \left[P_{n_k}^m(\cos\theta)\right]^2 \sin\theta\, d\theta = \frac{\sin\theta_0}{2n_k+1}\frac{dP_{n_k}^m(\cos\theta_0)}{d\theta}\frac{\partial}{\partial n} P_{n_k}^m(\cos\theta) \qquad (6)$$

for $(k-m)$ odd. Note that eqs (5) and (6) follow directly from eq. (4), using L'Hôpital's rule in the limit that n_k tends to n_j. As already noted by Lowes (1999), in Haines' paper the sign of our eq. (4) is incorrect.

The secular variation $\dot{\mathbf{B}}$, the first time derivative of the magnetic field **B**, can be modelled in the same way directly from secular variation data by introducing a secular variation potential and using the secular variation coefficients $\{\dot{g}_k^m, \dot{h}_k^m\}$ in eq. (1).

2.2 Inverse method and damping

As an alternative to the statistical regularization method of Haines (1985a), we develop regularization techniques following those used for global SHA. We use the linear inversion method based on the work of Whaler & Gubbins (1981) and Gubbins (1983). We minimize the function

$$(\gamma - \mathbf{Am})^{\mathrm{T}} \mathbf{C}_e^{-1} (\gamma - \mathbf{Am}) + \lambda \mathbf{m}^{\mathrm{T}} \Lambda \mathbf{m}, \tag{7}$$

where $(\gamma - \mathbf{Am})$ is the error vector given by the difference between data γ and the prediction of the model $\mathbf{m}$ and $\mathbf{A}$ is the operator calculated from eq. (1) relating the data vector to the model. $\mathbf{C}_e$ is the data error covariance matrix. The regularization is given by the second term: $\mathbf{m}^{\mathrm{T}}\, \Lambda \mathbf{m}$ is a quadratic norm of smoothness of the field over the spherical cap, Λ is a positive-definite damping matrix. λ is a Lagrange multiplier. The maximum-likelihood solution is

$$\hat{\mathbf{m}} = \left(\mathbf{A}^{\mathrm{T}} \mathbf{C}_e^{-1} \mathbf{A} + \lambda \Lambda\right)^{-1} \mathbf{A}^{\mathrm{T}} \mathbf{C}_e^{-1} \gamma. \tag{8}$$

The damping matrix is determined by the norm. In SHA, not only the basis functions but also the corresponding $\mathbf{B}_l^m$ are orthogonal over the sphere (Lowes 1966):

$$\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} \mathbf{B}_l^m \cdot \mathbf{B}_{l'}^{m'} \sin\theta \, d\theta d\phi = 0 \tag{9}$$

unless $l = l'$ and $m = m'$. Here, l has been used for the integer degree to avoid confusion with the non-integer n_k of spherical cap harmonics. For example, the 2-norm of the mean field strength averaged over the spherical surface is given by

$$\langle \mathbf{B} \cdot \mathbf{B} \rangle = \sum_{l=1}^{\infty} (l+1) \left(\frac{R_{\mathrm{E}}}{r}\right)^{2l+4} \sum_{m=0}^{\infty} \left[\left(g_l^m\right)^2 + \left(h_l^m\right)^2\right] \tag{10}$$

so the damping matrix based on minimizing this norm is diagonal with elements

$$f(l) = (l+1) \left(\frac{R_{\mathrm{E}}}{r}\right)^{2l+4}. \tag{11}$$

In SCHA, however, owing to the incomplete orthogonality of the basis functions, the fields are orthogonal only to the same extent as the basis functions (Lowes 1999), so the damping matrix is no longer diagonal. The functions are still ϕ-orthogonal, so integrals of products of functions with $m \neq m'$ are zero. As a result, most of the non-diagonal elements of the damping matrix are also zero, but the elements that combine odd and even harmonics of equal order m have finite values.

A straightforward, but rather weak, regularization is to minimize only the mean square radial component of $\mathbf{B}$. This norm is given by

$$\left\langle B_r^2 \right\rangle = \sum_{n_k} \sum_{n_j} \sum_{m} \left(g_{n_k}^m g_{n_j}^m + h_{n_k}^m h_{n_j}^m\right)(n_k + 1)(n_j + 1)$$
$$\times \left(\frac{R_{\mathrm{E}}}{r}\right)^{(n_k+n_j+4)} a \int_{\theta=0}^{\theta_0} P_{n_k}^m(\cos\theta) P_{n_j}^m(\cos\theta) \sin\theta \, d\theta, \tag{12}$$

where the θ-integral is given by (4)–(6), respectively. The mean $\langle\cdots\rangle$ refers here to the mean over the cap and the factor a is the result of the ϕ-integral, 2π for $m = 0$, π for $m \neq 0$, normalized for the area of the cap, $2\pi(1 - \cos\ \theta)$:

$$a = \begin{cases} 1/(1 - \cos\theta_0) & m = 0 \\ 1/[2(1 - \cos\theta_0)] & m \neq 0. \end{cases} \tag{13}$$

For $\theta_0 = \pi$ these values, respectively, become 0.5 and 0.25, which are the factors when normalizing the ϕ-integral with the area of the whole sphere in the case of global spherical harmonics. The square norm of the main field $\mathbf{B}$ is given by

$$\langle \mathbf{B} \cdot \mathbf{B} \rangle = \sum_{n_k} \sum_{n_j} \sum_{m} \left(g_{n_k}^m g_{n_j}^m + h_{n_k}^m h_{n_j}^m\right) \left(\frac{R_{\mathrm{E}}}{r}\right)^{(n_k+n_j+4)}$$
$$\times \left[\sin\theta_0 P_{n_j}^m(\cos\theta_0) \frac{dP_{n_k}^m(\cos\theta_0)}{d\theta} + (n_k + n_j + 1)(n_k + n_j + 2)\right.$$
$$\left. \times \frac{a}{2} \int_{\theta=0}^{\theta_0} P_{n_k}^m(\cos\theta) P_{n_j}^m(\cos\theta) \sin\theta \, d\theta \right]. \tag{14}$$

The first term in the square brackets vanishes for $k = j$, $(k - m)$ odd, or $(j - m)$ even, as either $P_n^m(\ \cos\ \theta_0)$ or $dP_n^m(\ \cos\ \theta_0)/d\theta$ is zero owing to the boundary conditions (eqs 3 and 2). The elements of the damping matrix $\mathbf{\Lambda}$ in this case are:

$$f_{kj} = \left[1 - \frac{(n_k + n_j + 2)}{(n_j - n_k)}\right] \left(\frac{R_{\mathrm{E}}}{r}\right)^{n_k+n_j+4}$$
$$\times \frac{a}{2} \sin\theta_0 P_{n_k}^m(\cos\theta_0) \frac{dP_{n_j}^m \cos\theta_0}{d\theta} \tag{15}$$

for the non-diagonal elements with $m = m'$, $(k - m)$ even and $(j - m)$ odd,

$$f_{kk} = -(n_k + 1) \left(\frac{R_{\mathrm{E}}}{r}\right)^{2n_k+4} a \sin\theta_0 P_{n_k}^m(\cos\theta_0) \frac{\partial}{\partial n} \frac{dP_{n_j}^m(\cos\theta_0)}{d\theta} \tag{16}$$

for the diagonal elements with $(k - m)$ even and

$$f_{kk} = (n_k + 1) \left(\frac{R_{\mathrm{E}}}{r}\right)^{2n_k+4} a \sin\theta_0 \frac{dP_{n_j}^m(\cos\theta_0)}{d\theta} \frac{\partial}{\partial n} P_{n_k}^m(\cos\theta_0) \tag{17}$$

for the diagonal elements with $(k - m)$ odd.

Another possibility is to minimize the 2-norm of the radial derivative of the field or its radial component, ensuring smoothness of the field with varying height. SCHA, similarly to SHA, allows upward and downward continuation, and it is unreasonable to obtain a model that is smooth only on a particular surface, but becomes very rough a short distance above the Earth's surface. Such a behaviour would surely not represent the actual geomagnetic field. Deriving the expressions for both the $(dB_r/dr)^2$ and $(d\mathbf{B}/dr)^2$ norms is straightforward from the B_r^2 and $\mathbf{B}^2$ norms: the derivative only gives an additional factor of

$$(n_k + 2)(n_j + 2) \left(\frac{R_{\mathrm{E}}}{r}\right)^2 \tag{18}$$

to (12) and (14). (Appendix A demonstrates this for the $\mathbf{B}^2$-norm.) Again all of these equations are also valid for the secular variation $\dot{\mathbf{B}}$ when substituting the time derivative of the coefficients $\{\dot{g}_k^m, \dot{h}_k^m\}$.

2.3 Temporal modelling with splines

To obtain the maximum constraint on the secular variation, we must model it simultaneously in time and space. To do this, we follow the approach of Bloxham & Jackson (1992) developed for global modelling, adopting the method of penalized least-squares splines as suggested by Constable & Parker (1988, 1991). We expand each of the SCHA Gauss coefficients in time on a basis of cubic B-splines $b_j(t)$

$$g_{n_k}^m(t) = \sum_{j=1}^{L} \alpha_{jkm} b_j(t). \tag{19}$$

α_{jkm} are temporal coefficients to be determined by an extension of the procedure outlined by (7). For example, we might minimize

$$(\gamma - \mathbf{Am})^{\mathrm{T}}\mathbf{C}_e^{-1}(\gamma - \mathbf{Am}) + \int\left[\lambda\int \mathbf{B}^2 d\Omega + \tau\int\left(\frac{\partial^2\mathbf{B}}{\partial t^2}\right)^2 d\Omega\right]dt, \tag{20}$$

where the three terms minimize the data misfit, spatial roughness and temporal roughness, respectively. λ and τ are Lagrange multipliers controlling the trade off of the misfit and roughness criteria.

Cubic B-splines are piecewise cubic polynomials, joined at points t_j which are called knots. For simple smoothing splines, the knots t_j are chosen to be the data points. However, with penalized least-squares splines, we seek to control the model fit and smoothness by damping. We thus choose a sufficiently high number of evenly spaced knots that the results are insensitive to their number or position. For modelling a data set covering 30 years, we have used 20 splines, which is more than adequate to represent the secular variation. The B-spline basis has several useful properties. First, the B-splines themselves are optimally smooth in the sense of minimizing a norm of the second derivative, and so they are a good basis to choose to represent the secular variation, which we expect (and preferentially choose) to be smooth. They have *small support*: the kth cubic B-spline is such that $b_k(t) > 0$ if $t_j < t < t_{j+4}$ and zero otherwise, and the sum of the B-splines is unity at any point. A more detailed discussion of splines and the B-spline basis is given by de Boor (1978).

3 DATA

To test the new method, we model the secular variation in Europe from a data set we had previously tested thoroughly and modelled with Haines' original method (Korte & Haak 2000). The data are taken from regional magnetic surveys of 12 European countries from 1955 to 1996. Owing to the temporally and spatially inhomogeneous distribution of the data some unifying processing was necessary. The processing step most important to keep in mind when interpreting the results is a reduction of the original data to a small number of epochs common to all points by means of a regional but large-scale 'normal' main field/secular variation model. The data set finally used consists of average annual secular variation data for 5 year intervals between 1965 and 1995. An average network of this processed data set, which consists of 300–400 points for every time interval, is shown in Fig. 1. A detailed description of the data set and data handling is given by Korte & Haak (2000).

A severe problem with these data is that they may contain errors up to the order of magnitude of the secular variation itself. As the raw data are momentary values measured at the survey stations on different days, they are generally reduced to 'annual means' using recordings of a nearby observatory. Under the assumption that all geomagnetic variations are the same at the stations and observatories, this procedure yields values of the internal magnetic field that are no more influenced by external variations than observatory annual means. Spatial differences in secular variation and the resulting errors are indeed negligible for reductions over short time spans. However, external and particularly induced variations can differ significantly even over short distances, producing errors of 10 nT and more in the desired internal results. These errors can only be reduced when the variations are recorded directly at the survey station, which in older surveys was rare. Moreover, our additional reduction of data to common epochs, inevitable in our first study, could have distorted the amplitude of any small-scale secular variation anomalies. We estimate these additional errors to be small enough not to prevent the detection of secular variation anomalies; nevertheless, they complicate the interpretation as to whether small-scale structure in the model is caused by induced secular variation effects or just data uncertainties.

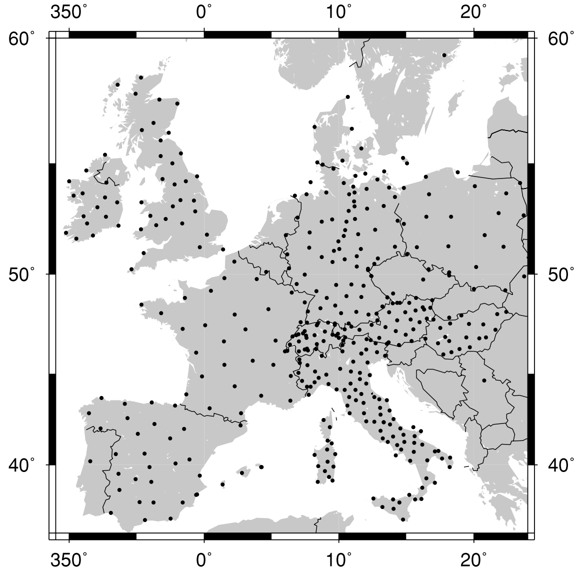

Figure 1. Network of magnetic repeat station and ground vector data used in our study, Mercator projection.

In a test with synthetic data, SCHA was shown to be stable with regard to quite large, but normally distributed errors. However, the results of modelling the real data at different epochs were less convincing (Korte & Haak 2000). With the damping methods described above we hope to have better physical constraints on spatial and temporal variability to overcome some of the problems of high data errors and unsatisfactory data distributions. For the time-dependent modelling we complemented the data set with annual secular variation measurements, i.e. first differences of annual means of the geomagnetic observatories. Thus we obtained a denser time-series of data with 300–400 points every 5 years and 20–30 observatory points every year in between.

We do not have good estimates of the data uncertainties, which may differ significantly between the original data sets of the different surveys. Therefore, as in our previous study of the data, we did not apply any weights to the data. Even when we added the observatory data, which are supposed to be of significantly higher quality, we did not apply any weights, reasoning that with their higher temporal density those data automatically have more influence on the models. A major advantage of the time-dependent modelling is to make the reduction of the different survey results to common epochs unnecessary. However, for the temporally sparse ground vector surveys that had to be included in our data set, it is not possible to determine the secular variation at the different stations with a better accuracy than that given by the reduction to common epochs. To benefit from the time-dependent modelling with our data set it would be necessary to model the main field and derive the secular variation from the derivative of the coefficients. That method works well with ordinary spherical harmonics and global data. With spherical cap harmonics, however, we meet the problem that the main field cannot be modelled well directly for a comparably small spherical cap owing to the long (global dipole) wavelengths of the main field and a much shorter maximum wavelength (the size of the cap half-angle) for the first spherical cap harmonic (Torta *et al.* 1992). The widely used solution of removing a known main field model such as the IGRF from the data prior to modelling seems problematic with our method of regularization. Applying the desired damping norm, chosen for physical considerations, to the residuals after subtraction of a model field derived using a different regularization is not a consistent approach. Therefore, we do not investigate this possibility further here.

4 RESULTS

We studied modelling results both for the epoch data with different damping norms and for time-dependent modelling and compared the results with those obtained using Haines' original method. As the half-angle of the cap we adopted the value of $\theta_0 = 18^\circ$ used in our previous study. The truncation level of the spherical cap harmonics determines the smallest possible wavelength of the modelled structures. We want to investigate whether there is small-scale structure in the data, so we must not force smooth models by choosing low truncation levels. We want fine-scale model structure to be controlled by the requirements of the data, rather than by the truncation level. Thus, we chose the maximum order of the spherical harmonics $k_{\max} = 9$ and looked for models that are smooth in the radial component of the field secular variation ($\dot{B}_r^2$-norm), in the main field secular variation itself ($\dot{\mathbf{B}}^2$-norm) and in the main field secular variation and its variation with height ($\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm), respectively. In the previous case the norm of the derivative is stronger than that of the field. We multiplied it by a factor of 0.1, chosen so that the two damping norms had a similar influence: $\dot{\mathbf{B}}^2 + 0.1(d\dot{\mathbf{B}}/dr)^2$.

We met one problem with the damping method that occurred for all of the tested norms: the range of eigenvalues of the damping matrix becomes very large, the values spanning approximately seven orders of magnitude (examples from the individual epoch modelling are from 10^{-5} to 10^2 for the $\dot{\mathbf{B}}^2$-norm or from 10^{-2} to 10^5 for the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm). This broad range of eigenvalues is a symptom of the existence of certain linear combinations of the spherical cap harmonics that give almost no field, a consequence of the incomplete orthogonality of the basis functions. Such eigenvectors are also clearly poorly constrained by the data. When including high maximum spatial indices $k_{\max}$ the smallest eigenvectors for the $\dot{B}_r^2$-norm (smaller than 10^{-5}) sometimes even became negative. We confirmed that this was just a numerical problem arising from the limited accuracy in calculating the Legendre functions by comparing the eigenvalues with the numerically determined norms of the different eigenvectors. This numerical problem of negative eigenvalues can easily be overcome by slight numerical damping for stability: a constant factor times the identity matrix is added to the damping matrix. In our studies, a factor of 10^{-5} was sufficient to eliminate negative eigenvalues, yet is small enough to be insignificant with respect to the physical damping.

The smoothness and misfit of a model are determined by the Lagrange multiplier or damping factor λ from eq. (7). Lacking good estimates of data uncertainties, we choose the appropriate factor for the smoothest and at the same time best-fitting model from the knee of a trade-off curve, a plot of the norm value against misfit. Larger values of λ increase the smoothness of the models but at the cost of the fit to the data, while smaller values of λ produce a better fit at the cost of increased model complexity. Fig. 2 shows trade-off curves between the data misfit, given as an rms value, and the norm for the three damping norms and data of one epoch. In all of the curves a knee is clearly visible and well defined at the same level of misfit. Trade-off curves for other epochs show the same well-defined behaviour. Owing to the temporal inhomogeneity and the subsequent reduction procedure the data from the different epochs have mean errors of different order and thus are fitted to different levels even without damping. The range of the rms misfit without damping is from 3.2 to 8.4 nT yr^{-1}, for the example shown in the following figures it is 3.9 nT yr^{-1}. For the time-dependent modelling the trade-off curves also show well-defined knees, except for the $\dot{B}_r^2$-norm. The minimum rms misfit is 4.4 nT yr^{-1} for the data set complemented with observatory data and 4.5 nT yr^{-1} for the original 5 yr interval data set, within the range of the misfits of the single epochs as expected. The numerical problems of negative eigenvalues are clearly reduced by the additional observatory data, owing to the better data constraints.

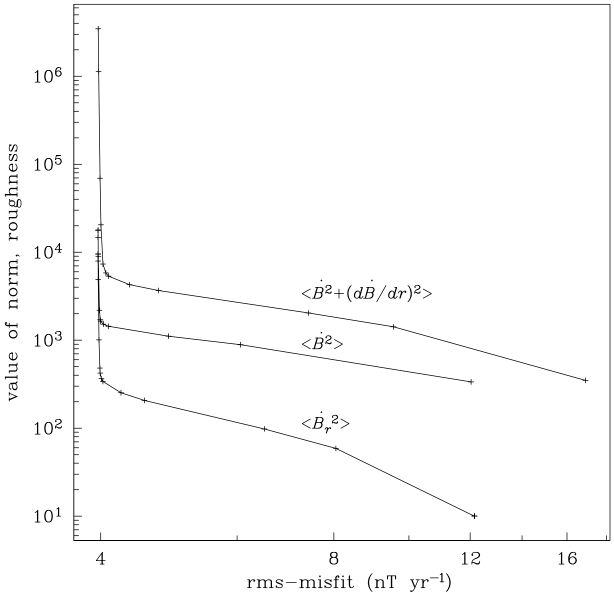

Figure 2. Trade-off curves between the data misfit and the roughness for three different norms.

A comparison of the resulting models with the different damping norms shows that neither the $\dot{B}_r^2$-norm nor the $\dot{\mathbf{B}}^2$-norm are adequate for our data. Models that are very strongly damped do not become smoother but even seem to show more small-scale structure while getting weaker owing to the overdamping. Fig. 3 shows this for the $\dot{\mathbf{B}}^2$-norm, for the $\dot{B}_r^2$-norm we obtain similar results for the Z-component while the X- and Y-components, as expected, are affected much less by the damping. Only the combined norm of $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$ produces the desired results of spatially smoothing the models, allowing for a sensible trade-off between the smooth model and the misfit to the data to determine how much small-scale structure is actually required by the data. With this norm the field is already much smoother for our preferred solution, see Fig. 4(a). This choice of norm is also supported by the fact that it eliminates the highly unrealistic large secular variation gradients at the southwesternmost and/or southeasternmost edge of the displayed area, which are present in several of the charts in Fig. 3. These regions still lie within the chosen spherical cap, but outside the region of the actual data (*cf.* Fig. 1). The effect mainly seems to be a problem of the data distribution, as it is also present with the statistical regularization method (see Fig. 4b). Apart from these edge-effect differences, the combined-norm models look quite similar to the models obtained

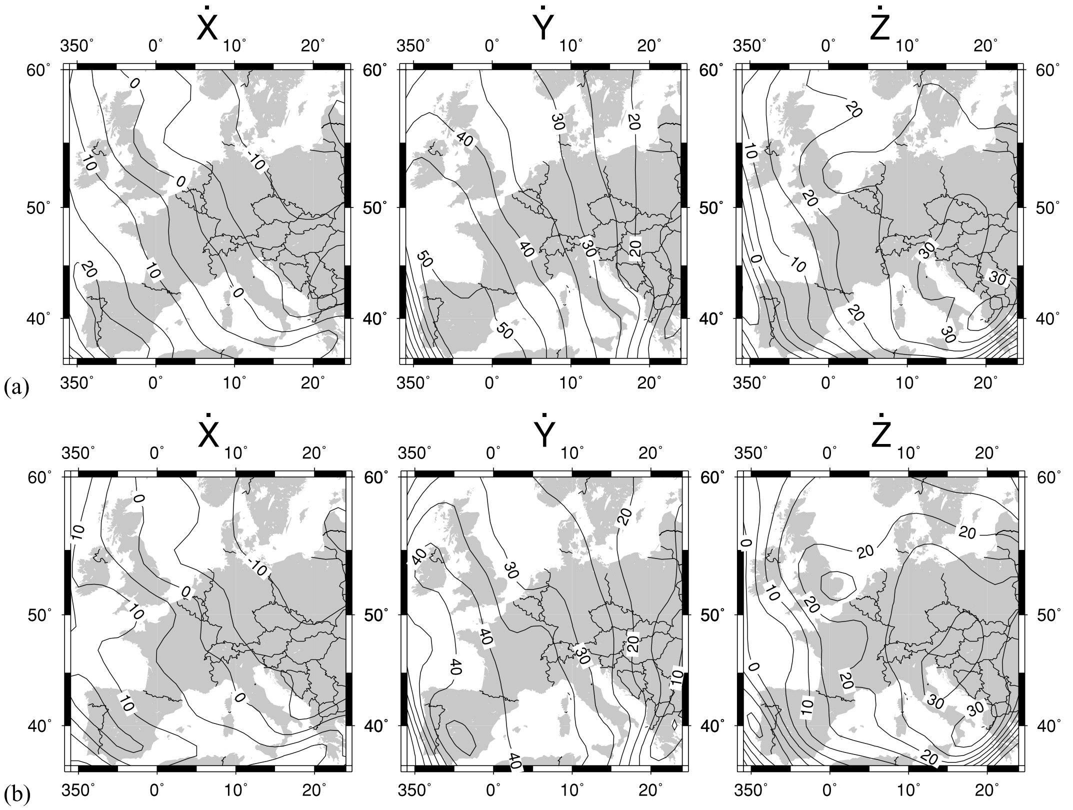

Figure 3. Model of secular variation for the time interval 1985–1990, regularized with the $\dot{\mathbf{B}}^2$-norm. (a) Damped optimally ($\lambda = 5$, residual=4.0 nT yr^{-1}), (b) damped more strongly ($\lambda = 50$, residual=5.0 nT yr^{-1}). Units are nT yr^{-1}, Mercator map projection.

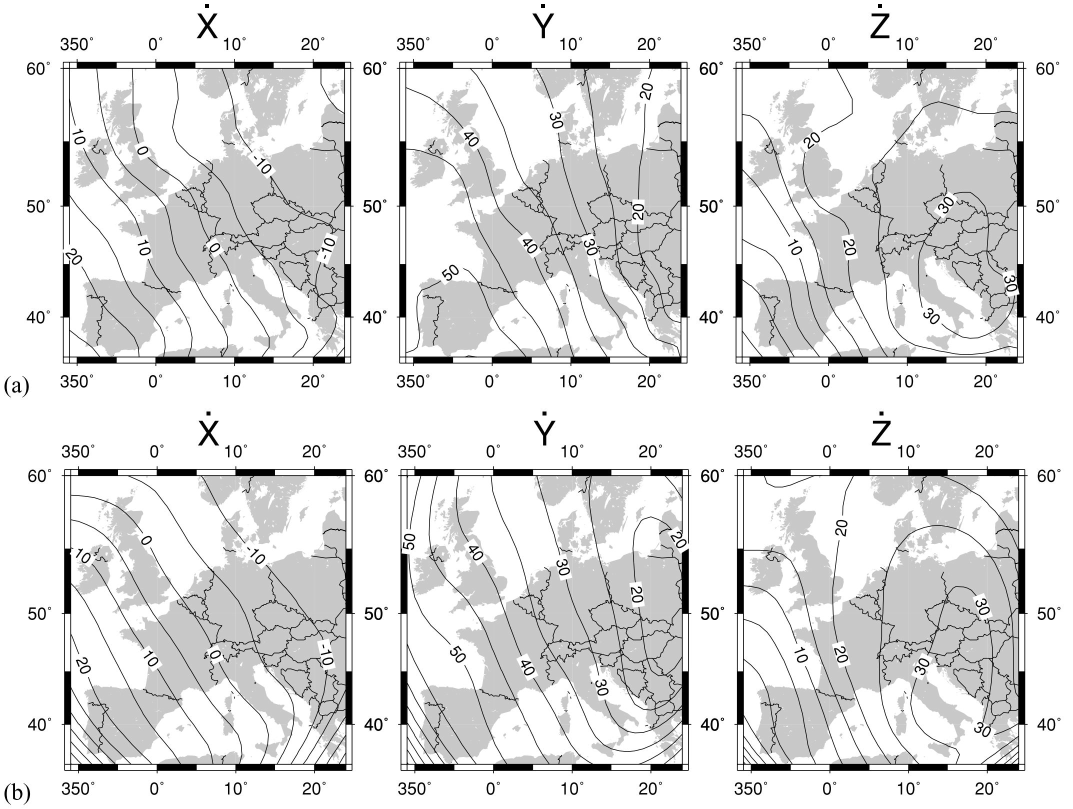

Figure 4. Model of secular variation for the time interval 1985–1990, units are nT yr^{-1}, Mercator map projection. (a) Regularized with the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm. Optimal damping ($\lambda = 1$, residual=4.0 nT yr^{-1}). (b) Model using the statistical method of rejecting coefficients (F-level of 3.5) for a maximum spatial index of $k_{\max} = 4$.

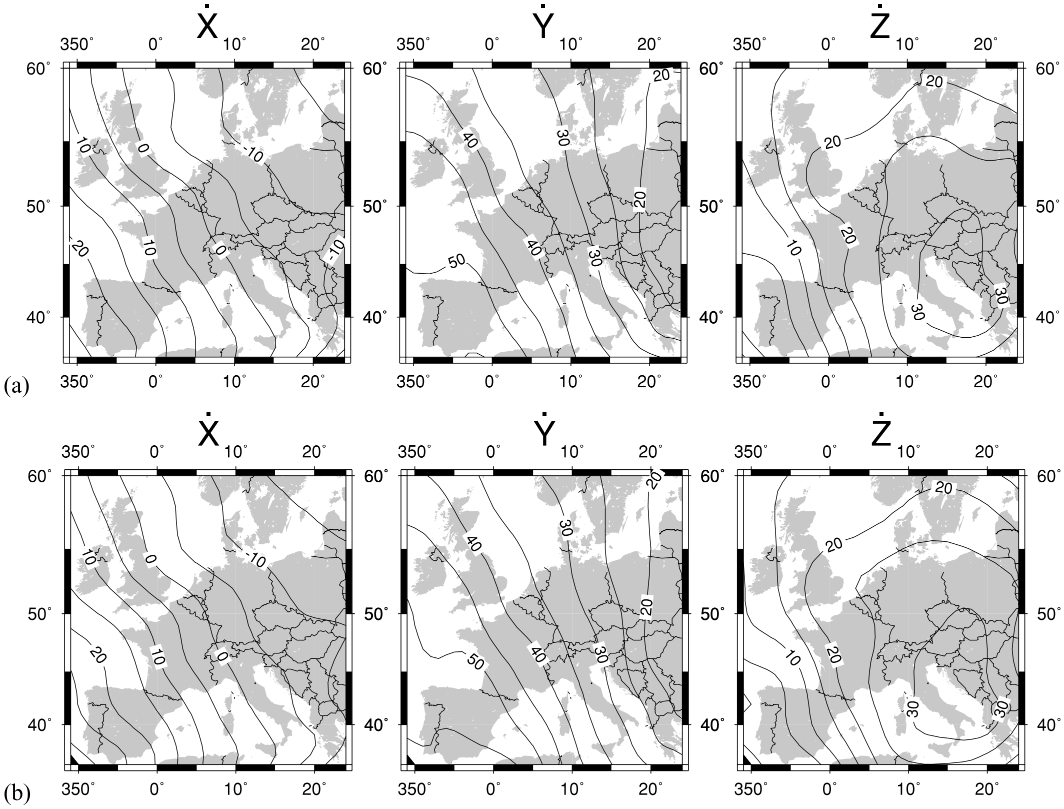

Figure 5. Results of the time-dependent models for epoch 1988.0, comparable to the individual models of the interval 1985–1990. All spatially optimally damped using the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm. Units are nT yr^{-1}, Mercator map projection (a) Original data set ($\lambda = 50$, $\tau = 1000$, residual=4.8 nT yr^{-1}). (b) Data complemented with observatory annual secular variation ($\lambda = 50$, $\tau = 5000$, residual=5.3 nT yr^{-1}).

using the statistical method with low truncation levels (Fig. 4b). For higher truncation levels the models with that method show more and more small-scale structure, which we did not believe to be reliable (*cf.* Korte & Haak 2000). If for these models we increase the F-parameter, that is we consider more small coefficients to be statistically insignificant, we also obtain an increased smoothness and misfit. However, most coefficients so rejected are of high degree, which for all of the models are small, and so the result is similar to a stronger truncation of the spherical cap harmonics series.

The time-dependent models show a similar behaviour with respect to the different norms, for both of our data sets. Figs 5(a) and (b) show time-dependent model results for the original and the complemented data set, respectively, optimally damped with the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm and for the same epoch as the previous examples. The optimum factor τ for the temporal regularization has been determined by trade-off curves of the temporal norm against the misfit in the same way as the spatial regularization factor λ had been determined previously. The addition of the observatory data does not change the models significantly, although they show slightly more small-scale structure. Generally, the appearance of the models changes only very slightly with increased temporal damping and subsequent increased misfit.

A look at the coefficients is instructive as it shows significant improvements with our method of regularization. Fig. 6 compares the temporal change of the coefficients for modelling the data for individual epochs with statistical regularization and with damping, and with time-dependent modelling. The coefficients of the 5 year intervals of the original data set are shown, joined linearly for illustration of the temporal change. In the time-dependent cases this is done for comparative reasons only, the models actually have temporally continuously changing coefficients. Only coefficients with $k = 1$–3 and $m = 0$ are shown, the behaviour of all the other coefficients is similar. It is obvious that with the statistical method (Fig. 6a) the individual coefficients are not meaningful; they vary much more in time than one would expect from looking at the field predictions of the models. Regularization by minimizing a field quantity leads to significantly less scatter of the coefficients in time (Fig. 6b), even when the epochs are still modelled individually and completely independently. With the time-dependent models (Figs 6c and d), we obtain the expected smooth change of coefficients, which is determined by the temporal damping here. Increasing or decreasing the temporal damping significantly influences the smoothness of the temporal variation of the coefficients, without influencing the field predictions of the models very much as already noted above. If we compare the actual continuous coefficient curves of the original and the complemented data set we see that the latter is less smooth owing to the additional data requiring variations shorter than 5 years.

5 CONCLUSIONS

We have developed a physical method of regularization for spherical cap harmonics analogous to that widely used for spherical harmonics. We have compared models obtained using this damping regularization to models obtained by the statistical regularization of Haines (1985a). Additionally, we have considered time-dependent models. We can show clear improvements in the performance of the damping regularization compared with the statistical method. The smoothness of the models is not arbitrarily defined by the truncation level of the series of basis functions, but can be traded off against misfit of the model to the data or, if good error estimates are

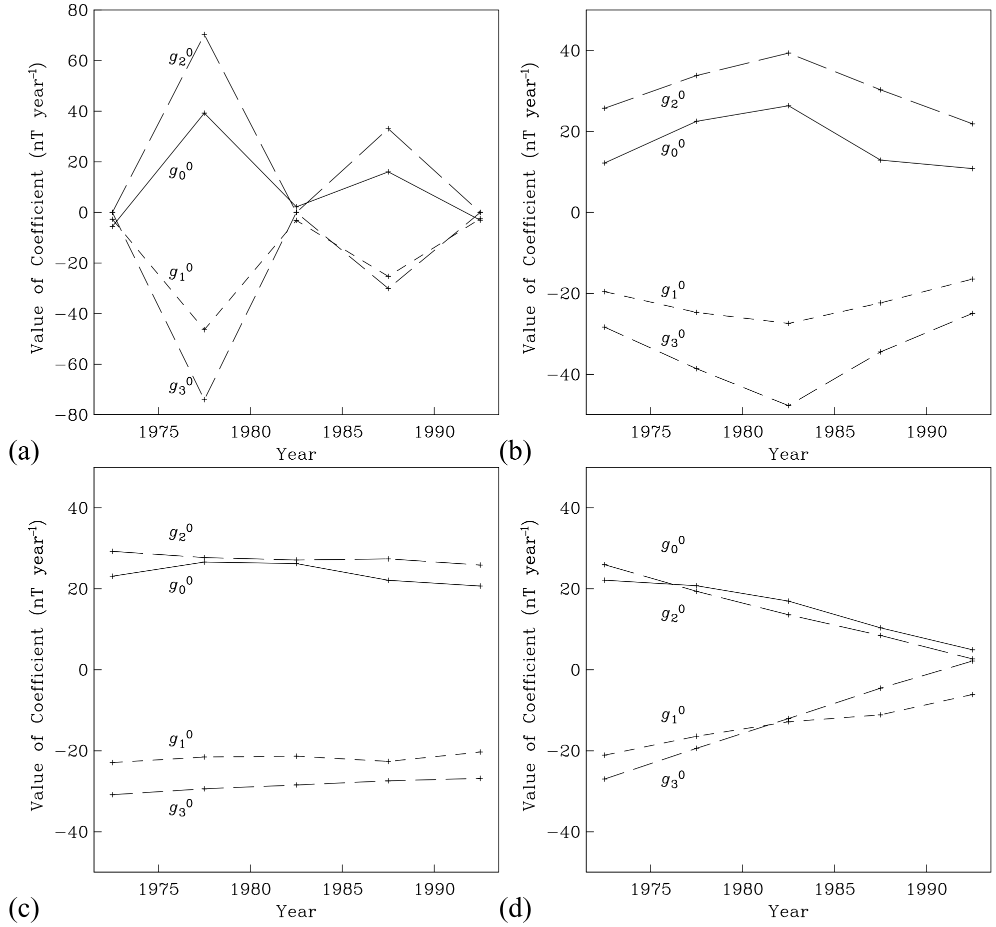

Figure 6. Change of the model coefficients in time for different methods of regularization. All coefficients plotted for the original 5 year interval epochs for comparison and joined linearly. (a) Statistical method, epochs modelled individually. (b) Damping with the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm, epochs modelled individually. (c) Time-dependent model, original data set, regularized with the $\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm, and optimal temporal regularization. (d) As (c), but the data set complemented with annual observatory data.

available, can be more rigorously determined by fitting the data to the estimated uncertainty. We can thus determine much better what amount of small-scale structure is actually required by the data. Without regularization we will overfit the data with its errors, including small-scale noise structure from the data errors in the model. With the statistical method this obviously also happens for higher truncation levels (Korte & Haak 2000), so we have to make a much stronger *a priori* assumption concerning the smoothness by choosing a low maximum truncation level. By damping we can choose truncation levels that we assume to be able to display structure that are much smaller than needed. The only *a priori* information is the kind of norm we choose for damping. However, here we can apply physical considerations, for example, the true secular variation field is likely to change slowly and smoothly with increasing distance from the Earth.

However, the superiority of the damping to the statistical method is most clearly shown by the coefficients. We have shown in Fig. 6(a) that the coefficients determined with the method of setting statistically insignificant coefficients to zero are not meaningful. The reason for this is the incomplete orthogonality of the basis functions. As our eigenvalue analysis of the damping matrix shows, there are linear combinations of spherical cap harmonics that contribute almost nothing to the field. Those combinations are not determined by the data, and we could add an arbitrary amount of them to the field at any epoch, changing the coefficients without changing the appearance of the model significantly. This means, however, that the choice of setting small coefficients to zero is meaningless. Depending only on the presence or absence of such poorly determined vectors, different coefficients make the criterion to be considered statistically insignificant. The spatial damping, although also unable to resolve such small eigenvalue eigenvectors, regularizes the well-resolved linear combinations in such a way that the individual coefficients become more meaningful. The change of the coefficients in time becomes much smoother, even if the data of the different epochs are modelled independently. In the time-dependent case we can apply an additional temporal regularization, which resolves some of the null space: we can smooth the temporal change of the coefficients without significantly changing the field predictions of the model. A comparison of coefficients of different epochs becomes much more meaningful if we have thus resolved the amount of temporal change in the individual coefficients required by the linear combinations contributing significantly to the field.

An additional advantage of this form of regularization is that we can vary the smoothing conditions according to the data. For our data set of limited quality in terms of errors and data distribution, the condition of just minimizing the mean field strength ($\dot{\mathbf{B}}^2$-norm) or its radial component ($\dot{B}_r^2$-norm) at the Earth's surface is not strong enough. With stronger damping, the field predictions of the model only become weaker, but no smoother (Fig. 3). The lateral structure is not significantly influenced by the damping in those cases.

This means that it does not provide the desired information concerning whether small-scale structure is actually required by the data or not. Only the combined condition of both minimum field strength and its radial derivative ($\dot{\mathbf{B}}^2 + (d\dot{\mathbf{B}}/dr)^2$-norm) gives satisfying results, the lateral smoothness of the field predictions changes with the amount of damping. Models damped optimally with that norm are much smoother than those damped with either of the other two norms (Fig. 4). The condition of a smooth horizontal derivative at the Earth's surface, even more explicitly allowing only as much lateral structure as required by the data, might also be a sensible choice for our data set but has not been implemented so far.

With regard to the data set, however, we must still conclude that it is not good enough to reveal possible lithospheric secular variation anomalies. Some of the model results still suggest that some small-scale structure might be present in the data. Given the large differences in the models with a different norm, however, we were not convinced that these are real secular variation anomalies; in particular, as we must not forget the rather strong shortcomings of the data set: the data of several countries had to be interpolated over long gaps with a smooth secular variation model. Additionally, the reduction to the 5 year intervals further distorts the amplitude of possible small-scale lithospheric secular variation anomalies. Moreover, such anomalies must be assumed to be very weak, which suggests that their reliable detection demands a much better data accuracy than most of these surveys provide. Given the encouraging results of this study, in particular of the spline-based time-dependent modelling, we aim to reprocess our repeat station data set. We will eliminate as many as possible of the additional reductions to common epochs and apply uncertainty estimates, putting more weight on the observatory data. We will also examine the effect of culling data, to investigate the appropriate density of repeat station measurements. Nevertheless, a good distribution of highly accurate time-series with no less than at least 5 year repeat intervals would be the desired data set to reliably confirm or reject the presence of lithospheric secular variation anomalies. For example, in Germany, a new network of high-quality stations has been set up in a 1999/2000 survey (Korte & Fredow 2001). This network is a subset of those used in the 1996.5 and 1992.5 surveys. A repeat interval of 2 to 4 years is planned, which combined with surveys from surrounding countries, will produce a vastly improved data set for these studies. SCHA also allows the possibility of including data at different heights, opening up the possibility of also using satellite data. With the methods described here, and this new data set, we anticipate finally being able to determine whether regional-scale variations in secular variation are observable.

ACKNOWLEDGMENTS

This paper was completed while MK was financially supported by a Feodor–Lynen fellowship from the Alexander von Humboldt foundation at Scripps Institution of Oceanography, University of California at San Diego. We appreciate the dedicated work of the many people carrying out repeat station surveys and running observatories, without whose data this study would not have been possible. Maps were produced using the GMT package (Wessel & Smith 1991, 1998).

REFERENCES

Alldredge, L.R., 1983. Varying geomagentic anomalies and secular variation, *J. geophys. Res.,* **88,** 9443–9451.

Bloxham, J. & Jackson, A., 1992. Time-dependent mapping of the magnetic field at the core–mantle boundary, *J. geophys. Res.,* **97,** 19 537–19 563.

Constable, C.G. & Parker, R.L., 1988. Smoothing, splines and smoothing splines: their application in geomagnetism, *J. Comput. Phys.,* **78,** 493–508.

Constable, C.G. & Parker, R.L., 1991. Deconvolution of long-core paleomagnetic measurements—spline therapy for the linear problem, *Geophys. J. Int.,* **104,** 453–468.

de Boor, C., 1978. *A Practical Guide to Splines,* Springer-Verlag, New York.

De Santis, A., Falcone, C. & Torta, J.M., 1997. SHA vs. SCHA for modelling secular variation in a small region such as Italy, *J. Geomag. Geoelectr.,* **49,** 359–371.

Draper, N.R. & Smith, H., 1966. *Applied Regression Analysis,* Wiley, New York.

Gubbins, D., 1983. Geomagnetic field analysis—I. Stochastic inversion, *Geophys. J. R. astr. Soc.*, **73,** 641–652.

Gubbins, D. & Roberts, P.H., 1987. Magnetohydrodynamics of the Earth's core, in *Geomagnetism,* Vol. 2, Academic, Orlando, FL.

Haines, G.V., 1985a. Spherical cap harmonic analysis, *J. geophys. Res.,* **90,** 2583–2591.

Haines, G.V., 1985b. Spherical cap harmonic analysis of geomagnetic secular variation over Canada 1960–1983, *J. geophys. Res.,* **90,** 2563–2574.

Haines, G.V. & Newitt, L.R., 1997. Canadian geomagnetic reference field, *J. Geomag. Geoelectr.,* **49,** 317–336.

Haines, G.V. & Torta, J.M., 1994. Determination of equivalent current sources from spherical cap harmonic models of geomagnetic field variations, *Geophys. J. Int.,* **118,** 499–514.

Korte, M., 1999. Kombination regionaler magnetischer Vermessungen Europas zwischen 1955 und 1995, *PhD thesis,* Scientific Technical Report STR99/11, GeoforschungsZentrum Potsdam.

Korte, M. & Fredow, M., 2001. *Magnetic Repeat Station Survey of Germany 1999/2000,* Scientific Technical Report STR01/04, GeoforschungsZentrum Potsdam.

Korte, M. & Haak, V., 2000. Modelling European repeat station and survey data by SCHA in search of time-varying anomalies, *Phys. Earth planet. Inter.,* **122,** 205–220.

Kotzé, } P.B., 2001. Spherical cap modelling of Oersted magnetic field vectors over Southern Africa, *Earth, Planets and Space,* **53,** 357–361.

Lowes, F.J., 1966. Mean-square values on sphere of spherical harmonic vector fields, *J. geophys. Res.,* **71,** 2179.

Lowes, F.J., 1999. Orthogonality and mean squares of the vector fields given by spherical cap harmonic potentials, *Geophys. J. Int.,* **136,** 781–783.

Mundt, W., 1973. Der Character der geomagnetischen Säkularvariation in Europa im Zeitraum von 1950 bis 1970, *Veröffentl. Zentralinst. Phys. Erde,* **23,** 1–22.

Mundt, W., 1981. Regionale und lokale Anomalien der geomagnetischen Säkularvariation in Mitteleuropa, *Veröffentl. Zentralinst. Phys. Erde,* **70,** 33–43.

Mundt, W. & Porstendorfer, G., 1977. Zusammenhänge zwischen Analysen der magnetotellurik und der magnetischen Säkularvariation im Hinblick auf eine Leitfähigkeitsanomalie im oberen Erdmantel Mitteleuropas, *Gerlands Beitr. Geophys.,* **86,** 337.

Mundt, W. & Porstendorfer, G., 1978. Mögliche Zusammenhänge zwischen einer elektrischen Leitfähigkeitsanomalie im Erdmante und anomalen magnetischen Säkularvariationen in Mitteleuropa, *Geodät. Geophys. Veröff. NKGG, Reihe III,* **39,** 153–162.

Porstendorfer, G., Hassaneen, A.-R. & Otto, J., 1979. Modellierungsversuche zur Beeinflussung der magnetischen Säkularvariation durch Inhomogenitäten der elektrischen Leitfähigkeit im erdmantel, *Gerlands Beitr. Geophys.,* **88,** 467–473.

Shure, L., Parker, R.L. & Backus, G.E., 1982. Harmonic splines for geomagnetic modelling, *Phys. Earth planet. Inter.,* **28,** 215–229.

Torta, J.M., Garcá, A., Curto, J.J. & DeSantis, A., 1992. New representation of geomagnetic secular variation over restricted regions by means of spherical cap harmonic analysis: application to the case of Spain, *Phys. Earth planet. Inter.,* **74,** 209–217.

Wessel, P. & Smith, W.H.F., 1991. Free software helps map and display data, *EOS, Trans. Am. geophys. Un.,* **72,** 445–446.

Wessel, P. & Smith, W.H.F., 1998. New, improved version of the generic mapping tools releated, *EOS, Trans. Am. geophys. Un.,* **79**, 579.

Whaler, K.A. & Gubbins, D., 1981. Spherical harmonic analysis of the geomagnetic field: an example of a linear inverse problem, *Geophys. J. R. astr. Soc.,* **65,** 645–693.

APPENDIX A: DERIVATION OF THE NORMS

Obtaining the B_r^2-norm (12) is straightforward, as B_r is given by

$$B_r = -\frac{\partial \Phi}{\partial r}. \tag{A1}$$

For the $\mathbf{B}^2$-norm we have to calculate

$$\int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\left(B_\theta^2 + B_\phi^2 + B_r^2\right)\sin\theta\, d\theta d\phi. \tag{A2}$$

The integral of B_r^2 has already been calculated as the B_r^2-norm eq. (12). The remaining surface integral of $B_\theta^2 + B_\phi^2$ is determined for both those horizontal components together (e.g. Gubbins & Roberts 1987):

$$\int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\left(B_\theta^2 + B_\phi^2\right)\sin\theta\, d\theta d\phi = \int_{\phi=0}^{2\pi}\sin\theta_0\frac{\partial\Phi}{\partial\theta}\Phi\bigg|_{\theta_0} d\phi$$
$$+\int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\Phi L^2\Phi\sin\theta\, d\theta\, d\phi, \tag{A3}$$

where L^2 is the angular momentum operator of quantum mechanics:

$$L^2 = -\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]. \tag{A4}$$

The first term yields

$$\sin\theta_0\sum_k\sum_m\sum_j\sum_{m'}\left(\frac{R_\mathrm{E}}{r}\right)^{n_k+n_j+4}\left(g_{n_k}^m g_{n_j}^{m'} + h_{n_k}^m h_{n_j}^{m'}\right)\frac{a}{2}$$
$$\times\left[P_{n_j}^{m'}(\cos\theta)\frac{dP_{n_k}^m(\cos\theta)}{d\theta} + P_{n_k}^m(\cos\theta)\frac{dP_{n_j}^{m'}(\cos\theta)}{d\theta}\right] \tag{A5}$$

and the second term

$$\sum_k\sum_m\sum_j\sum_{m'}\left[n_k(n_k+1) + n_j(n_j+1)\right]\left(\frac{R_\mathrm{E}}{r}\right)^{n_k+n_j+4}$$
$$\times\left(g_{n_k}^m g_{n_j}^{m'} + h_{n_k}^m h_{n_j}^{m'}\right)\frac{a}{2}\int_0^{\theta_0}P_{n_k}^m(\cos\theta)P_{n_j}^{m'}(\cos\theta)\sin\theta\, d\theta \tag{A6}$$

where a is the result of the ϕ-integral, see eq. (13). We must define the non-diagonal elements of the damping matrix such that the final matrix is symmetric. Summing up the two terms and eq. (12) yields eq. (14). The derivative of B_r is again straightforward, but it is less obvious that the derivative of the complete vector **B** also results in the same additional prefactor. Consider the integrals of the horizontal components again:

$$\int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\left[\left(\frac{\partial B_\theta}{\partial r}\right)^2 + \left(\frac{\partial B_\phi}{\partial r}\right)^2\right]\sin\theta\, d\theta d\phi$$
$$= \int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\left\{\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial\Phi}{\partial\theta}\right)\right]^2 + \left[\frac{\partial}{\partial r}\left(\frac{1}{r\sin\theta}\frac{\partial\Phi}{\partial\phi}\right)\right]^2\right\}\sin\theta\, d\theta d\phi$$
$$= \int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\left[\left\{\frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)\right]\right\}^2\right.$$
$$\left.+\left\{\frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\left[\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)\right]\right\}^2\right]\sin\theta\, d\theta d\phi$$
$$= \int_{\phi=0}^{2\pi}\sin\theta_0\left\{\frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)\right]\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)\right\}\bigg|_{\theta_0} d\phi$$
$$+\int_{\theta=0}^{\theta_0}\int_{\phi=0}^{2\pi}\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)L^2\frac{\partial}{\partial r}\left(\frac{\Phi}{r}\right)\sin\theta\, d\theta d\phi. \tag{A7}$$

Having changed the order of differentiation, the final step is analogous to (A3) and now for all of the terms that have to be summed for the final result the differentiation yields the factor given in eq. (18).