A&A 473, 673–682 (2007)
DOI: 10.1051/0004-6361:20077216

Astronomy & Astrophysics

Solar energetic particle spectra from the SOHO-EPHIN sensor by application of regularization methods

E. Böhm, A. Kharytonov, and R. F. Wimmer-Schweingruber

Institute for Experimental and Applied Physics, Extraterrestrial Physics, Christian-Albrechts-University Kiel, Leibnizstr. 11, 24098 Kiel, Germany
e-mail: [boehm;kharytonov;wimmer]@physik.uni-kiel.de

Received 1 February 2007 / Accepted 25 June 2007

ABSTRACT

Context. The Electron Proton Helium Instrument (EPHIN) on ESA's Solar and Heliospheric Observatory (SOHO) measures solar energetic electrons, protons, and alpha particles with a stack of six solid-state detectors forming a telescope. The energy deposit in these detectors must be inverted to derive the original energy of the incident particles, thus leading to the original energy spectrum of solar energetic particles. Normal inversion techniques, such as least-squares methods, rely on fitting a known functional behavior of the spectral dependence (normally a power law) to the measured data with some account taken for the instrument response. Such procedures can fail to retrieve accurate particle spectra, e.g., when count rates are low and unphysical negative counts result from the fitting procedure.
Aims. We show how regularization methods can be applied to energetic particle measurements to unambiguously derive the original particle spectrum without any assumptions about its functional behavior, while also satisfying constraints such as non-negative counts.
Methods. Such inversion techniques still require knowledge of the instrument response function, however, it is an improvement upon normal least-squares or maximum-likelihood fitting procedures because it does not require any a-priori knowledge of the underlying particle spectra. Given the instrument response function in matrix form (here derived using Monte Carlo techniques), the original Fredholm integral equations reduce to a discrete system of linear algebraic equations that can be solved by ordinary regularization methods such as singular value decomposition (SVD) or the Tikhonov method. This procedure alone may lead to unphysical negative results, requiring the further constraint of non-negative count rates. This technique avoids full deconvolution because it involves the solution of ill-conditioned or singular linear systems.
Results. We analyze data from SOHO/EPHIN by full deconvolution of the measured data with the instrument response function. We apply the SVD and Thikonov methods with and without constraints to measured data from SOHO/EPHIN.
Conclusions. The derived results agree well with those of other methods that rely on a-priori knowledge of the spectral shape of the particle distribution function, demonstrating the power of the regularization method for more general cases.

Key words. Sun: particle emission – methods: data analysis – interplanetary medium – methods: numerical

1. Introduction

The Electron Proton Helium Instrument (EPHIN), part of the Comprehensive Suprathermal and Energetic Particle Analyzer (COSTEP, Müller-Mellin et al. 1995), on ESA's Solar and Heliospheric Observatory (SOHO) measures solar energetic electrons, protons, and alpha particles in the energy range 150 keV to >5 MeV (electrons) and 4 MeV/n to >53 MeV/n (protons and helium isotopes). It addresses energetic-particle phenomena in the solar atmosphere as well as within the interplanetary medium: energetic particle sources, acceleration, transport, and propagation. All these phenomena affect the final spectrum of the particle species under investigation and may imprint a unique signature on it. However, common data-analysis procedures involve fitting well-understood, theoretically motivated functional shapes of the particle spectra to the data, accounting for unique instrumental characteristics, such as detection efficiencies. These methods suffer from at least two drawbacks: a) a pre-determined functional shape is prescribed for the particle spectrum, possibly deleting unique process-dependent signatures in the data, and b) unphysical negative fit-results often appear without appropriate penalty.

Here, we show how limitations a) and b) can be overcome by application of regularization methods which are well established in other fields, such as image restoration (e.g. Bertero & Bocacci 1998). We give the general mathematical and instrumental background in Sect. 2, a brief description of the regularization and optimization methods used here in Appendix A, as well as a discussion of the information needed for the general inversion. Regularization methods have been used e.g. by Prato et al. (2006) to reconstruct the differential emission measure from RHESSI-observed X-ray spectra from solar flares. These authors used an analytical expression for the instrument response function. Here, we use a discrete instrument-response function of EPHIN that is derived using Monte-Carlo simulations in a manner briefly described in Sect. 3. We then apply the inversion methods described in Appendix A to data from the EPHIN sensor in Sect. 4 and compare the results for the various regularization methods in Sect. 5. Throughout this paper, we restrict ourselves to analyzing electrons and protons; once the method works for these particles, the derivation of helium isotope spectra is straightforward.

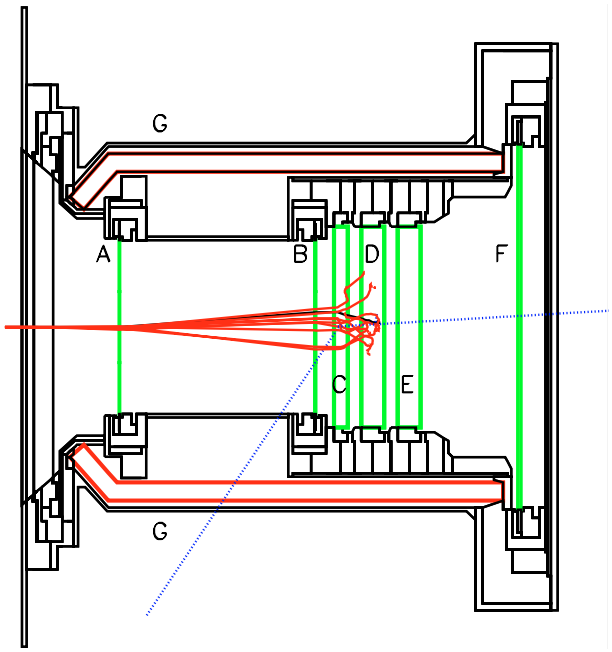

Fig. 1. Side view of the EPHIN-sensor, a telescope consisting of 6 Si-solid state detectors **A, B, C, D, E, F** (in green) surrounded by an active anticonincidence scintillator **G** (in red). 10 simulated tracks of 5 MeV electrons are shown in red as well as a pair of generated secondary gamma rays in blue.

2. Background

2.1. Instrumental background

EPHIN consists of a stack of five silicon detectors, surrounded by an anti-coincidence shield of plastic scintillator and a sixth silicon detector to distinguish between absorption and penetration (Fig. 1). Two passivated ion-implanted detectors (A and B) define the large (83°) field of view with a geometric factor of 5.1 cm^2 sr. The lithium-drifted silicon detectors C, D, and E stop electrons up to 10 MeV and hydrogen and helium nuclei up to 53 MeV/n. In this work, we will only discuss electrons and protons. The generalization to inclusion of helium ions is straightforward.

The energy measurement is pulse-height analyzed on board and sent back to Earth. On-board processing includes classification into electron (E) and proton (P) channels, depending on the energy deposited in detectors A–E. Thus, the E channels measure predominantly electrons, the P channels predominantly protons, but with cross contamination which can be appreciable, depending on the electron-to-proton ratio. The pulse-height-analyzed data contains information on the energy loss of single particles in detectors A–E in a total of 256 channels. For this work, we have rebinned this energy information into 60 logarithmically spaced energy intervals, sometimes even into 30 such bins. The sensitivity of each of these energy bins has been modeled as discussed in Sect. 3. Together, pulse-height words and energy bin sensitivities allow us to reconstruct the original particle spectra[1]. This paper addresses a new, improved method that relaxes the assumptions needed to reconstruct these original particle spectra without requiring a prescribed spectral shape.

[1] During times of high fluxes, the pulse height words need to be weighted according to another data type, the so-called histograms.

2.2. Mathematical background

Every experiment relates a set of "original" or "true" data, f, to measured data, z. The influence of the measurement apparatus is described by a function F acting on f in the most general case,

$$z = F(f). \tag{1}$$

In other words, the measuring process F acts on the "true" physical state described by variables f resulting in a measurement z. In the case of particle measurements, f are the energetic particles in interplanetary space, z are instrument count rates, and F is a description of the instrument response function. The problem of deriving z from f can be solved in two ways, directly, or indirectly. The direct problem is the calculation of z from f, whereas the inverse problem consists of trying to identify f from z. Often, Eq. (1) is solved using χ^2 or least squares methods, thus minimizing $||z - F(f)||^2$. However, these methods always rely on some underlying knowledge of the behavior of the "true" f, e.g. a power law is normally assumed for interplanetary energetic particle spectra. This is not necessarily satisfactory as particle spectra are not required to follow power laws, nor do upstream events or reconnection events. Thus, these direct methods suffer from the inclusion of sometimes unwarranted and sometimes unjustified information. On the other hand, we can consider the inverse problems, in which the functions f and z are connected by an operator equation

$$\hat{\mathbf{A}} f = z, \tag{2}$$

where the operator $\hat{\mathbf{A}}$ is the linear integral operator

$$\hat{\mathbf{A}}\colon\ \hat{\mathbf{A}} f = \int_a^b K(x, s) f(s) \mathrm{d}s, \tag{3}$$

where, again, $z(x)$ are the known measurements, $K(x, s)$ is the kernel that describes the measurement apparatus, and $f(s)$ is the sought, but unknown, function. Here, we will derive $K(x, s)$ from Monte-Carlo simulations of the EPHIN detector, thus the kernel can be considered a quantity that is known, although often with substantial uncertainty. The usual approach to solve this inverse problem is the discretization and formulation as a numerical algebra problem (Hansen 1992; Neumaier 1998; Tikhonov & Arsenin 1977). In that case, we have a system of linear algebraic equations which is often singular,

$$\mathbf{A} f = z, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix with elements a_{ij}, and $f \in \mathbb{R}^n$ is the unknown vector with components f_j, and $z \in \mathbb{R}^m$ is the known (measured) vector with components z_i. Usually, the vector z is the result of measurements contaminated by measurement errors (e.g. electronic noise), moreover, we only know the elements of the matrix $\mathbf{A}$ with limited accuracy. Thus, we need to solve the system Eq. (4), but all we know about the exact system

$$\tilde{\mathbf{A}} f = \tilde{z} \tag{5}$$

is that $||\mathbf{A} - \tilde{\mathbf{A}}|| \leq h$ and $||z - \tilde{z}|| \leq \delta$, where $h > 0$ and $\delta > 0$. However, there are infinitely many systems with such initial data $(\mathbf{A}, z)$, and, within the framework of the error level known to us, they are indistinguishable. Because we have only the approximate system, Eq. (4), instead of the exact system, Eq. (5), we can only attempt to find the optimal solution.

Unreasonable results may appear in the solution of Eq. (4), e.g., unphysical negative values in f in counting measurements. To obtain a nonnegative solution of f, the problem Eq. (4) has

to be replaced by the constrained minimization problem, which can be formulated for instance as a nonnegative least squares (NNLS) problem

$$\min \|\mathbf{A}f - z\|, \quad \text{subject to } f \geq 0 \tag{6}$$

or a similar maximum-likelihood problem with underlying Poissonian statistics. As is true with other methods, the least-squares problem always has a solution but it is non-unique if the rank of the matrix $\mathbf{A}$ is less than n, as it is in our case because the determinant of $\mathbf{A}$ is zero. NNLS problems are considered in Björck (1996); Calvetti et al. (2004); Lawson & Hanson (1974). In this work, we solve the NNLS problem Eq. (6) without any a-priori information about the unknown particle spectrum f. In this paper we emphasize the method rather than the physics behind.

3. Monte-Carlo simulations of instrument response functions

A particle entering the instrument deposits an energy E_d in the instrument. Because of incomplete charge collection and insensitive material (such as dead layers), E_d is smaller than the original energy of the particle, E_b. Because the energy deposit is also a stochastic process, the distribution of E_d will generally not be as well peaked as that of E_b. In addition, because of the different ionization cross sections of electrons and protons (and of other ions), E_d not only depends on E_b, but also on particle species. Other important factors determining E_d given E_b are angles, dead layers, detector arrangements, etc., and need to be taken into account when simulating detector response to a beam of particles with energy E_b and intensity $I(E_\mathrm{b})$. Thus, the intensity of the energy deposit, $I(E_\mathrm{d})$, is given, generally, by

$$I(E_\mathrm{d}) = \int \mathrm{d}E_\mathrm{b} \boldsymbol{K}(E_\mathrm{d}, E_\mathrm{b}) I(E_\mathrm{b}), \tag{7}$$

which is an integral equation that needs to be solved for the unknown quantity $I(E_\mathrm{b})$. Here, $\boldsymbol{K}(E_\mathrm{d}, E_\mathrm{b})$, the kernel of the integral equation, is the instrument sensitivity to measure E_d given a particle of energy E_b. It can be derived, e.g., from Monte-Carlo simulations of the detector response

$$I(E_\mathrm{d}) \cdot \mathrm{d}E_\mathrm{d} = \int (\mathrm{d}F\, \mathrm{d}\Omega\, \mathrm{d}E_\mathrm{b}\, \epsilon(x, y, \theta, \phi, E_\mathrm{b}, E_\mathrm{d}))\, \mathrm{d}E_\mathrm{d}. \tag{8}$$

The kernel, $\boldsymbol{K}$, is then given by

$$\boldsymbol{K} \doteq \int \mathrm{d}F\, \mathrm{d}\Omega\, \epsilon(x, y, \theta, \phi, E_\mathrm{b}, E_\mathrm{d}), \tag{9}$$

where ϵ is the efficiency of detecting a particle with energy in $[E_\mathrm{b}, E_\mathrm{b} + \mathrm{d}E_\mathrm{b}]$ that deposits energy in the range $[E_\mathrm{d}, E_\mathrm{d} + \mathrm{d}E_\mathrm{d}]$, hitting the detector in the area element $\mathrm{d}F$ at position (x, y) within the solid angle element $\mathrm{d}\Omega$ from the direction given by θ and ϕ. ϵ and $\boldsymbol{K}$ need to be calculated using a Monte-Carlo simulation that is described below. Here, we will assume an isotropic angular distribution of the beam and a homogeneous (uniform) distribution on the sensor area.

Equation (7) is a Fredholm integral equation which may also be interpreted as an algebraic equation if the kernel, $\boldsymbol{K}$, is known in discrete intervals, just as Eq. (3) may be interpreted as Eq. (4),

$$z = \mathbf{A}f \quad \rightarrow \quad z_i = \sum_{j=1}^{n} \boldsymbol{a}_{ij} \cdot f_j, \quad i = 1, \ldots, m, \tag{10}$$

where

$$\boldsymbol{a}_{ij} = \int_{E_\mathrm{d}}^{E_\mathrm{d}+\Delta E_\mathrm{d}} \int_{E_\mathrm{b}}^{E_\mathrm{b}+\Delta E_\mathrm{b}} \boldsymbol{K}(E_\mathrm{d}, E_\mathrm{b}) \mathrm{d}E_\mathrm{b}\, \mathrm{d}E_\mathrm{d}, \tag{11}$$

and

$$z_i = \int_{E_\mathrm{d}}^{E_\mathrm{d}+\Delta E_\mathrm{d}} z(E_\mathrm{d}) \mathrm{d}E_\mathrm{d}. \tag{12}$$

As mentioned in Sect. 2.1, we used the GEANT-4 package (The GEANT4 collaboration 2006) of the CERN program library to simulate the detector response of EPHIN to ions and electrons. The Monte-Carlo data were analyzed with the PAW package of the CERN program library, and information about energy resolution and triggering requirements of EPHIN were added at this stage. The simulated response functions, $\mathbf{A}$, of the geometric factors for the E- and P-channels are shown in Fig. 2. The four panels show graphical representations of the matrix elements $\mathbf{a}_{ij}$ in the E channels (upper panels) for electrons (left, $\mathbf{A}_\mathrm{Ee}$) and protons (right, $\mathbf{A}_\mathrm{Ep}$), and in the P channels (lower panels) for electrons (left, $\mathbf{A}_\mathrm{Pe}$) and protons (right, $\mathbf{A}_\mathrm{Pp}$). Results are shown for 30 logarithmic energy bins in E_d and E_b. Obviously, the matrices $\mathbf{A}_\mathrm{Ee}$ and $\mathbf{A}_\mathrm{Pp}$ are near diagonal and much larger than the matrices $\mathbf{A}_\mathrm{Ep}$ and $\mathbf{A}_\mathrm{Pe}$ that describe the crosstalk of protons in the E channel and vice-versa. However, during times of high particle fluxes, this crosstalk may become important when the ratio of contaminating particles to measured particles is high. Close inspection of the matrices $\mathbf{A}$ shows that they are not exactly diagonal, but that, generally, $E_\mathrm{d} \leq E_\mathrm{b}$, moreover, they have tails that extend to lower E_d. This is mainly due to insensitive material, energy escaping in the form of neutral particles, but also a background of energy deposit by penetrating particles.

The normally used geometric factors are projections of the matrices shown in Fig. 2 onto the E_b-axis. We show these classical EPHIN histogram geometry factors for the electron and proton histograms in the upper half of Fig. 3 and the projections onto E_d in the lower half. Figure 3 shows the projections of the matrices $\mathbf{A}_\mathrm{Ee} + \mathbf{A}_\mathrm{Ep}$ (left-hand side) and $\mathbf{A}_\mathrm{Pp} + \mathbf{A}_\mathrm{Pe}$ (right-hand side) onto E_b (upper panel) and E_d (lower panel). Two contributions are shown in each geometry factor. The main contribution to the E-channels (left-hand panels) comes from electrons (marked e) and from protons (marked p) to the proton channels (right-hand panels). The smaller contributions at higher energies in the electron channels and the lower energies in the proton channels come from the contaminating protons (for E-channels) and electrons (for P-channels) and are marked by p and e. The vertical lines show the statistical uncertainties resulting from the Monte-Carlo simulations of the geometry factors. Obviously, they are larger for the contaminating particles than for the principal ones. Substantial differences in the E_b and E_d projections can be seen at low energies for electrons. The divide between electrons and contaminating protons visible in the projection onto E_b is smeared out in the projection onto E_d and complicates the analysis.

Using the geometry factors discussed above, the intensity, I, of particles is calculated by

$$I_X(E_\mathrm{b}) = H_X(E_\mathrm{d})/g_X(E_\mathrm{b}), \quad \text{where } X = \mathbf{E}, \mathbf{P}. \tag{13}$$

Here, H is the measured number of counts with E_d per unit time, and g is the geometry factor (shown in Fig. 3). Because of the limitations just mentioned, this "classical" inversion method is not robust, especially at low and high energies.

The deficiency discussed above can be addressed by including all information available, i.e., by also including the information from the crosstalk channels. Both electrons and protons

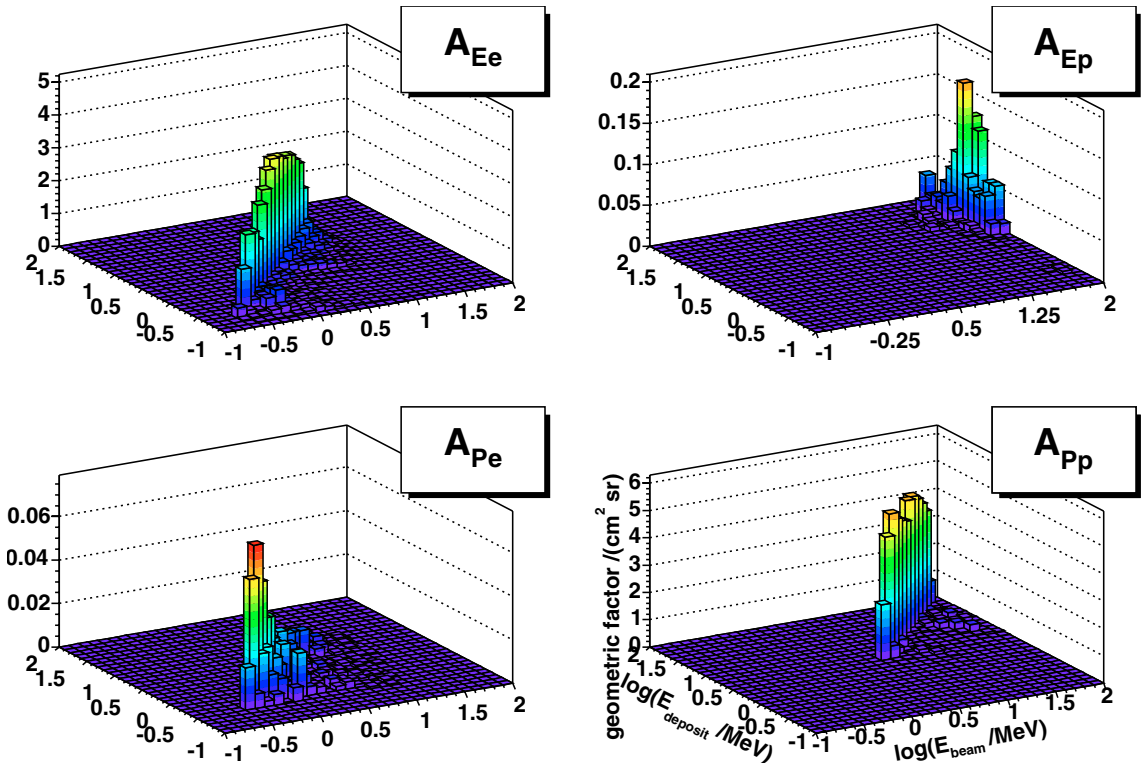

Fig. 2. Geometric factors $\mathbf{A}_{\mathrm{Ee}}$, $\mathbf{A}_{\mathrm{Ep}}$, $\mathbf{A}_{\mathrm{Pe}}$, and $\mathbf{A}_{\mathrm{Pp}}$ as derived from the Monte-Carlo simulation of the EPHIN instrument response discussed in Sect. 3. Note the different scales for the values of the geometry factor (z-axis). We have applied logarithmic binning to the geometry factors to account for the power-law particle spectra that are expected in particle acceleration processes. $\mathbf{A}_{\mathrm{Ee}}$ and $\mathbf{A}_{\mathrm{Pp}}$ are near diagonal along $E_{\mathrm{d}} = E_{\mathrm{b}}$, with some leakage towards lower energy deposits due to insensitive material and other losses. The cross-talk channels $\mathbf{A}_{\mathrm{Ep}}$ and $\mathbf{A}_{\mathrm{Pe}}$ are much smaller than $\mathbf{A}_{\mathrm{Ee}}$ and $\mathbf{A}_{\mathrm{Pp}}$ and considerably less diagonal.

contribute to the measured counts in the electron and proton channels. Formally, this can be written as the algebraic system

$$z_i^{\mathrm{E}} = \sum_{j=1}^{k} a_{ij}^{\mathrm{Ee}} f_j^{\mathrm{e}} + \sum_{j=k+1}^{2k} a_{ij}^{\mathrm{Ep}} f_j^{\mathrm{P}}, (i = 1, \ldots, k), \tag{14}$$

$$z_i^{\mathrm{P}} = \sum_{j=1}^{k} a_{ij}^{\mathrm{Pe}} f_j^{\mathrm{e}} + \sum_{j=k+1}^{2k} a_{ij}^{\mathrm{Pp}} f_j^{\mathrm{P}}, \quad (i = k+1, \ldots, 2k = n), \tag{15}$$

or, in matrix form,

$$\begin{pmatrix} z^{\mathrm{E}} \\ z^{\mathrm{P}} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{\mathrm{Ee}} & \mathbf{A}_{\mathrm{Ep}} \\ \mathbf{A}_{\mathrm{Pe}} & \mathbf{A}_{\mathrm{Pp}} \end{pmatrix} \begin{pmatrix} f^{\mathrm{e}} \\ f^{\mathrm{p}} \end{pmatrix} \tag{16}$$

where f^{e} and f^{p} are the electron and proton spectra, and z^{E} and z^{P} are the measured count rates in the E- and P-channels. This, then, is the mathematical problem that needs to be solved. Nevertheless, it is useful to consider some limiting cases. The easiest is, obviously, the case where there is no crosstalk; it results in two independent equations,

$$z^{\mathrm{E}} = \mathbf{A}_{\mathrm{Ee}} f_0^{\mathrm{e}}, \text{ and } z^{\mathrm{P}} = \mathbf{A}_{\mathrm{Pp}} f_0^{p}, \tag{17}$$

where f_0^{e} and f_0^{p} are the spectra for electrons and protons.

Most instruments exhibit crosstalk between P- and E-channels (protons contribute to the E-channels and electrons to the P-channels), and, hence, we will not treat this idealized case, but will solve the following cases, as well as the full Eq. (16). Formally, the electron and proton spectra may be combined into one vector describing a "combined" electron-proton spectrum, f^{ep}

$$f^{\mathrm{ep}} = f^{\mathrm{e}} + f^{\mathrm{p}}. \tag{18}$$

With this combined spectrum, Eq. (16) may be simplified,

$$\begin{pmatrix} z^{\mathrm{E}} \\ z^{\mathrm{P}} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{\mathrm{Ee}} + \mathbf{A}_{\mathrm{Ep}} \\ \mathbf{A}_{\mathrm{Pe}} + \mathbf{A}_{\mathrm{Pp}} \end{pmatrix} f^{\mathrm{ep}} = \begin{pmatrix} \mathbf{A}_{\mathrm{E}} \\ \mathbf{A}_{\mathrm{P}} \end{pmatrix} f^{\mathrm{ep}}. \tag{19}$$

On the other hand, both lines of Eq. (19) can be considered separately, leading to two separate equations,

$$\left(\mathbf{A}_{\mathrm{Ee}} + \mathbf{A}_{\mathrm{Ep}}\right) f_1^{\mathrm{ep}} = z^{\mathrm{E}} \tag{20}$$

$$\left(\mathbf{A}_{\mathrm{Pe}} + \mathbf{A}_{\mathrm{Pp}}\right) f_2^{\mathrm{ep}} = z^{\mathrm{P}}. \tag{21}$$

Because both electrons and protons contribute to the E- (Electron) and P- (Proton) channels, the solutions of Eqs. (19), (20) and (21) should agree approximately

$$f^{\mathrm{ep}} \approx f_1^{\mathrm{ep}} \approx f_2^{\mathrm{ep}} \tag{22}$$

in the sense that the accuracy differs due to the widely varying geometrical factors in the covered energy range (Fig. 3): Eq. (20) describes the E-channel and f_1^{ep} with poor information on protons, Eq. (21) describes the P-channel and f_2^{ep} with poor information on electrons. In the following, we will solve Eqs. (16) and (19), as well as Eqs. (20) and (21) and check their solutions for consistency and compare them with results from other inversion methods.

4. Application of regularization methods to EPHIN data

We chose data from three time periods around a solar particle event that occurred on days 353 and 354 in 2002. Figure 4 shows

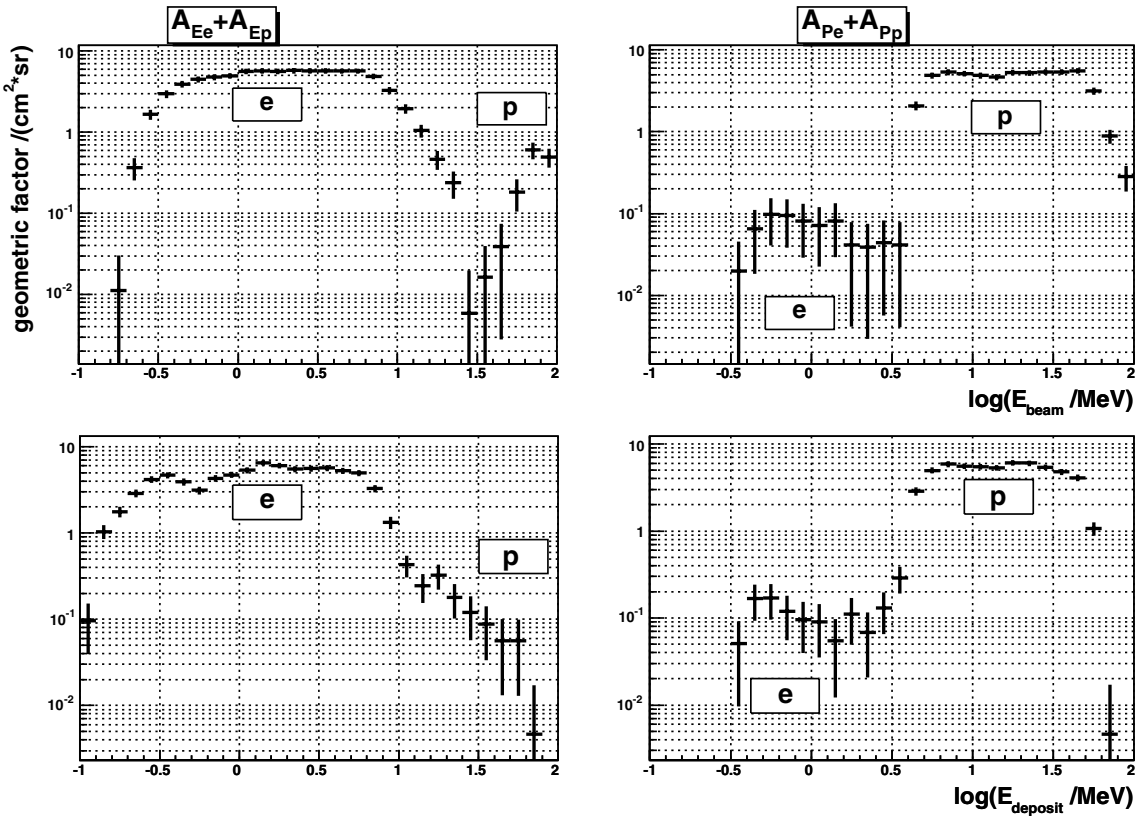

Fig. 3. Projections of the summed geometric factors shown in Fig. 2 onto the E_b (*top panels*) and E_d (*lower panels*) axes. Left-hand panels show the geometric factors $\mathbf{A}_{Ee} + \mathbf{A}_{Ep}$ for the electron (E) channels, and right-hand panels show the geometric factors $\mathbf{A}_{Pp} + \mathbf{A}_{Pe}$ for the proton (P) channels. Large differences are visible between the two projections of the electron channels, especially at low and high energies, while the proton channels appear to be shifted towards lower energies.

the count rates measured in the electron (E) and proton (P) channels. We chose the 22 h preceding the event (shaded in turquoise) to better understand contamination issues in quiet times, and the two first hours of the event (grey and red) as two separate periods because of the large variation of the electron to proton ratio.

We inverted the measured data z for electron and proton spectra f using both the singular-value-decomposition (SVD) method of Eq. (A.5) with solution f_{SVD} and the iterated Tikhonov regularization method of Eq. (A.11) with solution f_{Tikh}. This latter solution was obtained by solving the algebraic problem of Eq. (4). The two solutions f_{SVD} and f_{Tikh} were found to be completely congruent, as should be the case. However, both solutions sometimes exhibited unphysical negative values. Therefore, these solutions were subsequently used as an initial estimate for the optimization problem with constraints, Eq. (A.15). We also used the null vector as a starting value for this procedure, however, this resulted in inconsistent values, demonstrating the importance of choosing "good" initial values in optimization problems. We used MATLAB 7.0.4.352. for the calculations and the fmincon routine to compute the minimum in the optimization procedure.

Using this method, we solved Eqs. (16) and (19), as well as Eqs. (20) and (21) with $m = n = 30$ for Eqs. (20) and (21), $m = 60$; $n = 30$ for Eq. (19), and $m = n = 60$ for Eq. (16).

Figure 5 shows the original data (solid black line) in the individual electron (E, upper panel) and proton (P, lower panel) channels for the three time periods shown in Fig. 4. The thin vertical lines show statistical uncertainties. The red stars show the results of the inversions of Eqs. (20) and (21) which have been multiplied by their corresponding matrices $\mathbf{A}_{Ee} + \mathbf{A}_{Ep}$ and $\mathbf{A}_{Pe} + \mathbf{A}_{Pp}$. The overall agreement is good, and the disagreements can be readily understood if we again consider the geometric factors in the bottom panels of Fig. 3. For electrons, g_E is small at low and large energy deposits, $E_{deposit}$, resulting in a) smaller count rates and, b), in larger relative uncertainties in the accuracy to which g_E has been determined. Both effects result in larger uncertainties at these extremes of $E_{deposit}$. The inverted data, f, are much lower than the measurements, z, at the low-$E_{deposit}$ end, indicating that effect b) is the cause for these systematic discrepancies. Furthermore, the geometric factors were simulated between 100 keV and 100 MeV, fully covering the energy range of stopping particles, the limits of the energy range usually contribute to distortion and systematic uncertainties. Improvements in the Monte-Carlo model of EPHIN will be needed, especially the extension to non-stopping, relativistic particles. The proton geometric factor, g_P is relatively constant above $E_{deposit} \approx 3.2$ MeV. The counts above this energy are mainly due to protons, while those beneath are largely due to contaminating electrons. Obviously, the counts at energies below $E_d \approx 0.3$ MeV are not expected because the geometry factor vanishes below those values. This is a clear indication that the Monte-Carlo simulation of the electron contamination is not complete at these low energies and, hence, needs improvement. Thus, comparison of inverted data with measured data can be utilized to obtain an estimate of the quality of the simulated geometry factors.

The much better agreement between f^{ep}, the inversion of Eq. (19), and f_1^{ep} in the low energy range and f_2^{ep} in the high energy range indicates that inversion of Eq. (19) is the most robust solution of the three.

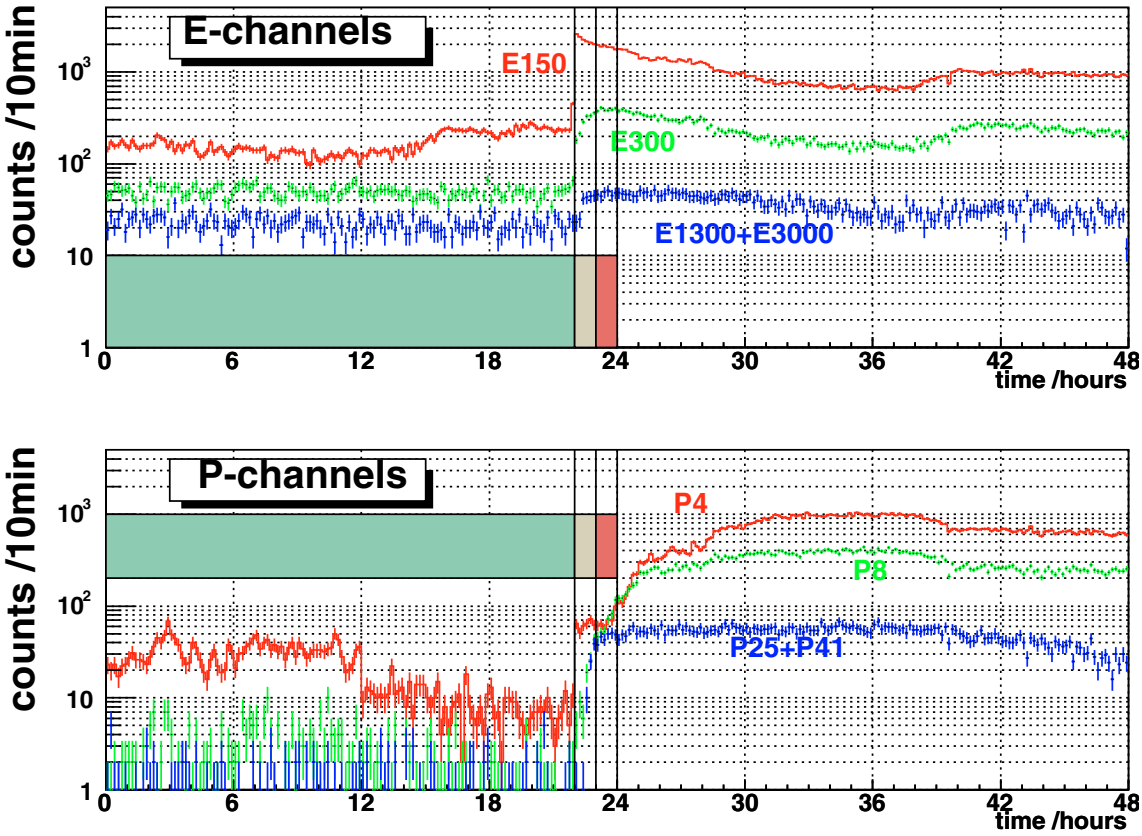

Fig. 4. Time series of predominantly electron (*upper panel*) and proton (*lower panel*) count rates around the solar particle event of days 353 and 354 in 2002. The various curves (red, green, and blue) show count rates in different energy intervals. The three regions that we analyse in more detail are shaded in color: a 22-h pre-event period in turquoise, and two consequtive hours at the beginning of the event (grey and red).

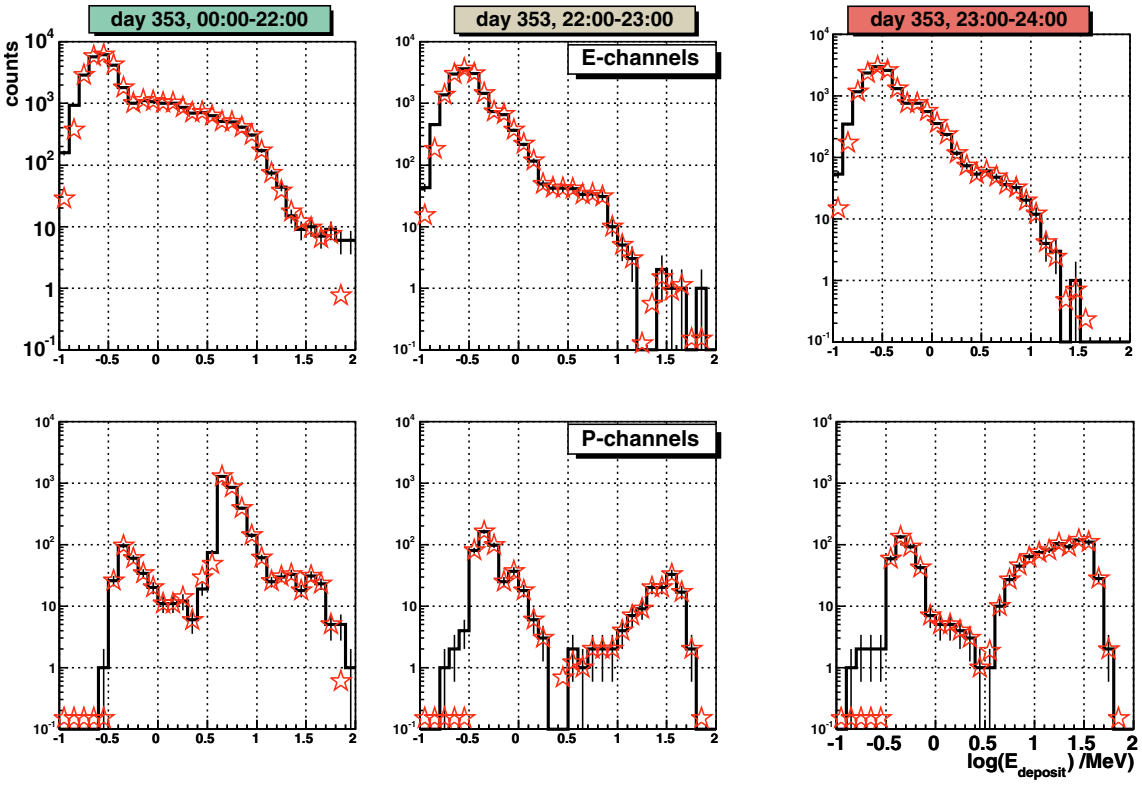

Fig. 5. EPHIN measurements (solid black curves) and inversion results (red stars) in the on-board processed electron and proton channels (*upper and lower panels*, respectively) for the three time periods of the solar particle event on days 353 and 354 of 2002. The measurements (black curves) correspond to z^{E} and z^{P} in all equations, while the red stars are the results for $(\mathbf{A}_{\mathrm{Ee}} + \mathbf{A}_{\mathrm{Ep}}) f_1^{(\mathrm{ep})}$ and $(\mathbf{A}_{\mathrm{Pe}} + \mathbf{A}_{\mathrm{Pp}}) f_2^{(\mathrm{ep})}$. Red stars along the x-axis show "underflow results" ($I \leq 10^{-1}$).

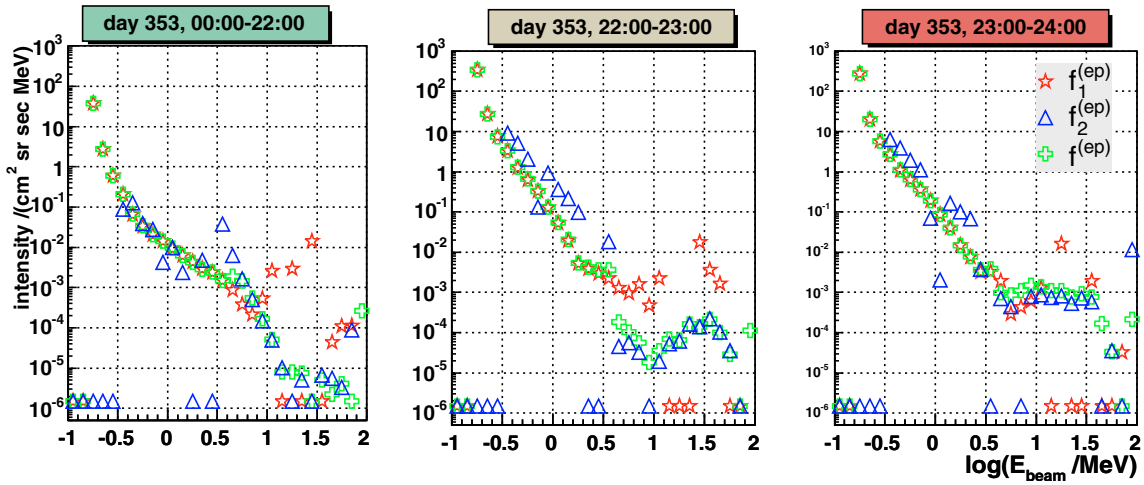

Fig. 6. Comparison of the intensities from **E**- and **P**-channels (Eq. (20) with $m = 30$, $n = 30$, $f_1^{(ep)}$), (Eq. (21) with $m = 30$, $n = 30$, $f_2^{(ep)}$), and a combined **E** and **P** analysis (Eq. (19) with $m = 60$, $n = 30$, $f^{(ep)}$). Points at intensities of 10^{-6} show underflows, ($I \leq 10^{-6}$).

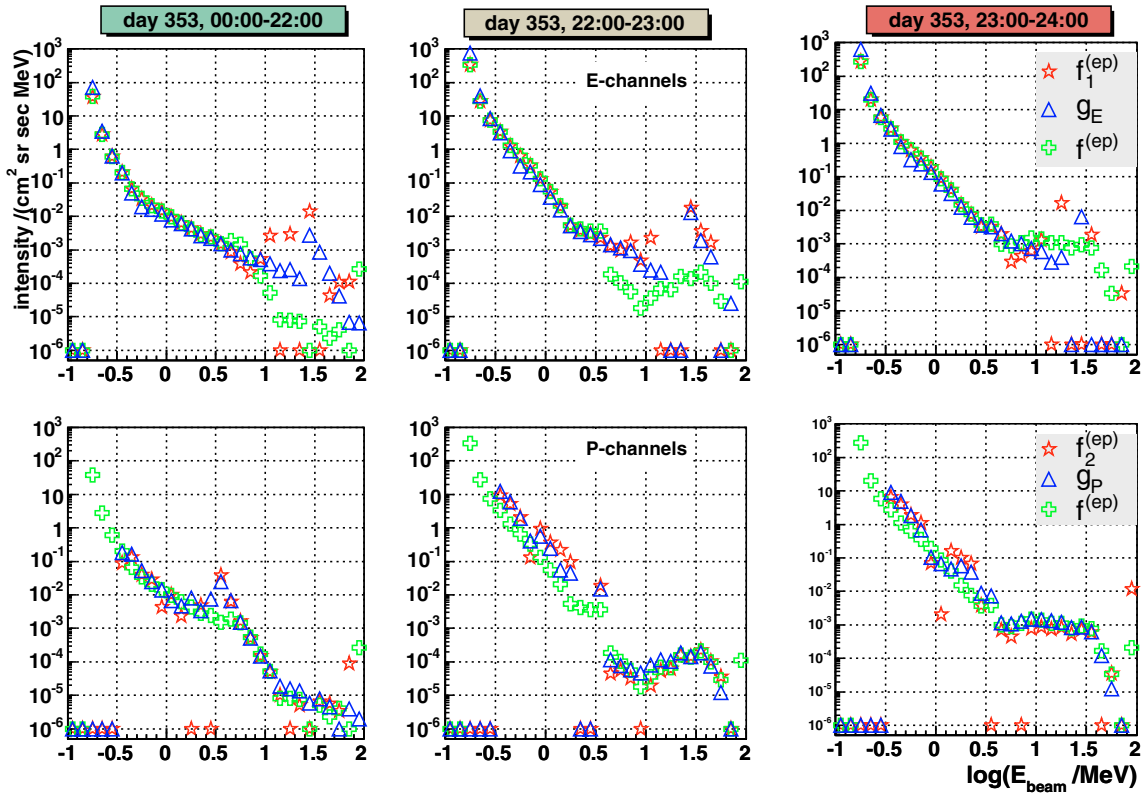

Fig. 7. Differential intensities for E and P channels (*upper and lower panels, respectively*) for the three time periods shown in Fig. 4. Results for the inversion of Eq. (19) with $m = 60$, $n = 30$, $f^{(ep)}$ are shown as green pluses in both panels, while the blue triangles show the "classical" inversion of Eq. (13). Results for the inversions of Eq. (20) ($f_1^{(ep)}$) and (21) ($f_2^{(ep)}$) are also shown as red stars in the upper and lower panels, respectively. Points at intensities of 10^{-6} show underflows, ($I \leq 10^{-6}$).

As we have already stressed, inversion of Eqs. (19), (20), and (21) should yield the same results, but with additional information about the effect of crosstalk. Therefore, we invert these equations with the method discussed in this section and the Appendix.

5. Comparison of methods

We now consider the results of our method for self consistency and compare them with those derived from different methods.

Figure 6 shows a comparison of the results for the differential intensities $f_i^{(ep)}$ in the E and P channels derived from the inversions of Eqs. (20), (21), and of the simplified full set of equations, Eq. (19). Symbols along the x-axis show "zero" measurements. As already discussed, all three results should agree, as all are derived from an equation that still contains the full information, albeit at a strongly varying level of accuracy. Thus, deviations tell us where crosstalk and/or low model accuracy limit the reliability of our results. The inversions of Eqs. (20), f_1^{ep}, and Eq. (19), f^{ep}, agree well in the low-energy range, whereas those of Eq. (21), f_2^{ep}, do not. In this low-energy range, the latter data

are derived from the proton channels and are largely due to contaminating electrons. Thus, it appears that the electron contamination is not yet correctly simulated in our instrument model. Their geometry factor is probably underestimated by a factor ∼2 which is still within the statistical uncertainties shown in Fig. 3. Nevertheless, it appears that we can derive electron spectra even from the proton-channel data. f_2^{ep} agrees with f^{ep} more closely above ∼3 MeV, where the proton geometry factor is dominated by the proton contribution, whereas f_1^{ep} agrees less well. This behavior is what is expected and, interestingly, the (contaminating) protons derived from the electron channels show at least a qualitative agreement in the shape of the spectrum in this energy range. Again, substantial improvements are needed in the knowledge of the geometry factor. Some of the points that need improvement are a more accurate knowledge of the actual experimental resolution, noise levels, and, especially, the efficiency of the anti-coincidence for near-relativistic particles which yield pulse heights near the detection threshold.

There appears to be one additional systematic difference between our derived electron- and proton-dominated intensities $f_1^{(\mathrm{ep})}$ and $f_2^{(\mathrm{ep})}$. Inspection of Fig. 4 shows that electrons are nearly an order of magnitude more abundant in their count rates than protons in the pre event time period, and probably even more overabundant during the two time periods following the event onset. Thus, contamination of the proton channels by electrons needs to be accounted for, whereas contamination of electron channels by protons is probably unimportant during these phases of the event. This reinforces our conclusion that inversion of Eq. (19) is the most robust solution of the three methods compared in Fig. 6. The graphics show the advantage of the inversion method, which does not make any assumption about a power law exponent. All three methods (Eq. (19), (20), Eq. (21)) agree as well as expected.

Next, we compare the results of our methods with the "classical" inversion of Eq. (13). Figure 7 shows differential intensities as derived from the proton (P) channels using Eq. (19) (green open pluses, f^{ep}) and Eq. (13) (upward-pointing open blue triangles, g_{E}). For comparison, we also show the electron- and proton-dominated inversions of Eqs. (20), f_1^{ep}), and (21), f_2^{ep}) as open red stars in the upper and lower panels of Fig. 7. No error bars have been included so far. This needs further investigation into the error propagation in inversion techniques. The lower left-hand panel shows the drawbacks of using the "classical" inversion scheme. While results are mostly satisfactory, large deviations between results and the more appropriate inversion of the full Eq. (19) occur where g_P changes abruptly and count rates are low (see Figs. 3 and/or 5). Similar discrepancies can be seen at higher energies in the upper panels. However, we also see that the two methods agree reasonably well in those regions where the geometric factors and count rates are large (see Figs. 3 and/or 5). Electron-dominated results ($f_1^{(\mathrm{ep})}$) agree over a wider range of energies than proton-dominated results. This is probably a result of the contamination of EPHIN proton measurements by electrons and much less contamination of electron measurements by protons. Note that electrons are much more abundant than protons in the event studied, as is easily seen in Fig. 4.

As already mentioned, both E- and P-channels should yield the same results but with different accuracy in the lower and higher energy ranges. The different accuracy is due to the very different responses of $\mathbf{A}_{\mathrm{E}}$ and $\mathbf{A}_{\mathrm{P}}$ at low and high energy (see Fig. 2). Indeed, this effect can be seen when comparing the upper and lower panels of Fig. 7. The wide deviations in the high-energy range in the E channels contrasts with low variability in the P channels.

We have not yet attempted to distinguish between original electrons and protons. This information is present in the full set of equations, Eq. (16). The result of their inversion is shown in Fig. 8, where it is also compared to the independent results of Gómez-Herrero (2005). Both independent methods show very satisfactory agreement, especially during the high-flux time periods. A marked disagreement can be seen at the very lowest and very highest energy bins; this is due to the small values of the geometric factors at these energy values (see Fig. 3) and the resulting large correction factors. Moreover, our simulations of the geometry factors included energies in the range 0.1 MeV $< E <$ 100 MeV and it is likely that the deconvolution of this instrument description with the measured data yields distorted results near the simulated boundaries in energy. Traditional data inversion, exemplified by the data of Gómez-Herrero (2005), have to make certain assumptions about the electron contamination in the proton channels. Such assumptions are not needed with our inversion method; the contamination by other particle species is included in a consistent manner in the relevant geometry factors.

6. Discussion and conclusions

We have compared several inversion techniques for space-based measurements of energetic particles. Such measurements suffer from the limited amount of information available to the experimenter about the exact amount of deposited energy and several coincidence conditions. We have shown that this missing information can be circumvented by applying sophisticated inversion techniques which take into account knowledge of instrument behavior.

A key advantage of the methods discussed in this paper is that no assumptions need to be made about the original spectral form, nor about spectral indices, or cross-talk of the particle spectra. Furthermore, generalization to more species, i.e., the inclusion of He ions (and even hydrogen isotopes) is straightforward.

Further work will be needed to derive quantitative estimates on the uncertainties of the results, which are essential for any interpretation of the physical results and thus are crucial to evaluate the potential of these methods. The methods presented here to solve the inverse problems associated with energetic particle data are sufficiently universal that they can be applied to other problems. Because they are relatively simple and straightforward and allow use of Monte-Carlo-simulated instrument functions (as opposed to analytic expressions), they can easily be used in other fields where the reduction of measurement data needs to be done with incomplete knowledge.

Appendix A: Regularization and optimization methods

A.1. Singular value decomposition

It is known (Björck 1996; Lawson & Hanson 1974) that the minimum norm solution of the least squares problem $\|\mathbf{A}f - z\|^2 = \min$ is given by the vector

$$f = \mathbf{A}^{+} z, \tag{A.1}$$

and is called a pseudo-solution of the system Eq. (4); the matrix $\mathbf{A}^{+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}$.

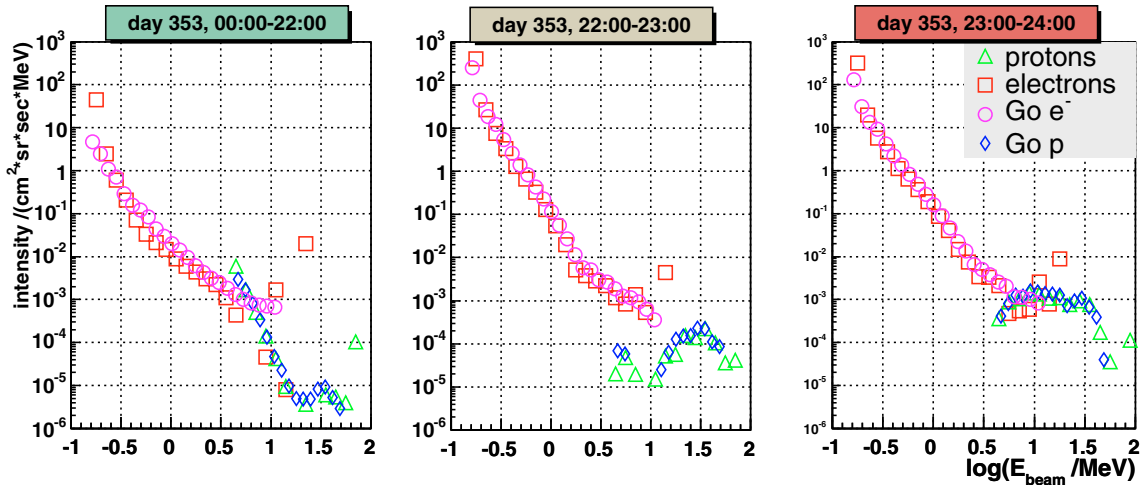

Fig. 8. Proton and electron fluxes on day 2002/353 from the present analysis (△ protons , □ electrons) together with results from Gómez-Herrero (2005) (◯ Go e$^-$, ◇ Go p).

The singular value decomposition (SVD) of the matrix $\mathbf{A}$ for the general case is

$$\mathbf{A} = \mathbf{U\Sigma V}^*, \tag{A.2}$$

where for $\mathbf{A} \in R^{m\times n}$, $\mathbf{U}$ is a square orthogonal $m \times m$-matrix, $\mathbf{V}$ is a square orthogonal $n \times n$-matrix and $\mathbf{\Sigma} = \mathrm{Diag}(\sigma_1, ..., \sigma_n)$ a rectangular diagonal $m \times n$-matrix with diagonal entries $\Sigma_{ii} = \sigma_i$. Here $\mathbf{V}^*$ denotes the adjoint of a matrix $\mathbf{V}$. Since $\mathbf{A}$ is real, the matrix $\mathbf{V}$ is also real, and, hence, the adjoint matrix $\mathbf{V}^*$ coincides with the transposed matrix $\mathbf{V}^\mathbf{T}$. If $rank(\mathbf{A}) = r < n$ for the case $m = n$, then $\sigma_{r+1} = \sigma_{r+2} = ... = \sigma_n = 0$.
The pseudo-inverse $\mathbf{A}^+$ of the matrix $\mathbf{A}$ is represented in the form

$$\mathbf{A}^+ = \mathbf{V\Sigma}^+\mathbf{U}^*, \tag{A.3}$$

where $\mathbf{\Sigma}^+$ is obtained from $\mathbf{\Sigma}$ by replacing each positive diagonal entry by its reciprocal

$$\mathbf{\Sigma}^+ = \mathrm{Diag}(\sigma_i^+), \tag{A.4}$$

$$\sigma_i^+ = \begin{cases} 1/\sigma_i & \text{if } \sigma_i \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Ill-conditioned matrices are characterized by the presence of small singular values σ_i, and it is clear that errors in z are highly magnified in the solution

$$f = \mathbf{A}^+ z = \mathbf{V\Sigma}^+\mathbf{U}^* z. \tag{A.5}$$

Therefore, the minimum norm least squares solution with SVD Eq. (A.5) is useless for problems with tiny but nonzero singular values σ_i. For such cases, one must use the truncated singular value decomposition (TSVD) as the regularization technique (Hansen 1987).

A.2. The Tikhonov method

There are many methods for solving inverse problems, all with their advantages and disadvantages. We have found the iterated Tikhonov method (Tikhonov & Arsenin 1977) to be well suited to our ill-posed linear problem

$$\mathbf{A}f = z. \tag{A.6}$$

$\mathbf{A}$ is the linear operator or the matrix to solve the modified least square problem

$$\Phi_\alpha(f) = \|\mathbf{A}f - z\|^2 + \alpha\|f\|^2 = \min, \tag{A.7}$$

where α is the regularization parameter ($\alpha > 0$). The variational problem Eq. (A.7) is equivalent to seeking the solution of Euler's equation (see, e.g., pp. 335–336 in Bertero & Bocacci 1998; or p. 243 in Kress 1989).

$$\mathbf{A}^*\mathbf{A}f_\alpha + \alpha f_\alpha = \mathbf{A}^* z, \tag{A.8}$$

which leads to the solution

$$f_\alpha = (\mathbf{A}^*\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^* z, \tag{A.9}$$

where $\mathbf{I}$ is the identity matrix.

This unique solution f_α for every α is called the Tikhonov approximation to the generalized solution $f^+ = \mathbf{A}^+ z$. It can be shown (pp. 84–90 in Groetsch 1993) that solutions f_α converge to $\mathbf{A}^+ z$ as $\alpha \to 0$. The key issue with the Tikhonov method, or with other regularization methods, is to find the value of the regularization parameter α, that gives a good or the best solution. Methods for the choice of the regularization parameter, α, are presented in Hansen (1992).

A.3. Iterative regularization methods

As we have just seen, the SVD method breaks down in the presence of very small singular values σ_i, and the choice of the regularization parameter α in the Tikhonov method is non-trivial.

In the following, we consider iterative methods for regularization such as the Landweber method and the iterated Tikhonov method (Bertero & Bocacci 1998; Neumaier 1998).

A simple iterative method for approximating the least-squares solution of integral equations of the first kind has been introduced by Landweber,

$$f_{k+1} = f_k + \tau\mathbf{A}^*(z - \mathbf{A}f_k) \quad \text{for} \quad k = 0, 1, 2, ..., \tag{A.10}$$

where τ is the relaxation parameter. In the linear case the standard choice of initial guess is $f_0 = 0$.

On the other hand, if we use the Thikonov method (Eq. (A.9)) with iterative refinement, we obtain the following iterative scheme

$$f_{k+1} = f_k + (\mathbf{A}^*\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^* r_k, \quad \text{where} \quad r_k = z - \mathbf{A}f_k, \tag{A.11}$$

with $f_0 = 0$ and $r_0 = z - \mathbf{A}f_0 = z$. In this case, the parenthese in the denominator is a good choice of the parameter τ.

This is called the iterated Tikhonov regularization method; however, it may also be considered as a preconditioned Landweber method.

The iterated Tikhonov regularization Eq. (A.11) is not the same as iterative refinement for solving the regularized normal Eq. (A.9). The latter is obtained by replacing $\mathbf{A}^* r_k$ with the vector

$$\mathbf{A}^* z - (\mathbf{A}^*\mathbf{A} + \alpha\mathbf{I}) f_k = \mathbf{A}^* r_k - \alpha f_k \tag{A.12}$$

and has the form

$$f_{k+1} = f_k + (\mathbf{A}^*\mathbf{A} + \alpha\mathbf{I})^{-1}(\mathbf{A}^* r_k - \alpha f_k). \tag{A.13}$$

In iterative methods the number of iterations, k, plays the role of the regularization parameter as can be seen by considering the influence of errors in the data z. Suppose that the available data is a vector z^δ satisfying

$$\|z - z^\delta\| \le \delta.$$

Using the vector z^δ in the iterative Landweber method Eq. (A.10), we have

$${f^\delta}_{k+1} = {f^\delta}_k + \tau\mathbf{A}^*(z^\delta - \mathbf{A}{f^\delta}_k). \tag{A.14}$$

As with the Tikhonov regularization, we will choose the "stopping value" $k = k(\delta)$ with the property that, if the iteration is terminated at step $k = k(\delta)$, then

$${f^\delta}_{k(\delta)} \to \mathbf{A}^+ z \qquad \text{as} \quad \delta \to 0.$$

A.4. Optimization with constraints

To obtain the solution of the specific physical problem with nonnegative values of the function $f = f_{\rm opt}$, where ($f_i \ge 0$, $i = 1, ..., n$), we need to find the minimum of a constrained nonlinear multivariable function

$$y(f) = \min \sum_{i=1}^{m} \left(\frac{z_i - \sum_{j=1}^{n} a_{ij} f_j}{s_i} \right)^2 \tag{A.15}$$

subject to $f_i \ge 0$ ($i = 1, ..., n$).

This method is better than, e.g., a maximum-likelihood method with underlying Possonian statistics because it does not rely on a priori knowledge of the functional form of the particle distribution function f. Here z_i are the components of a vector $z \in R^m$, $a_{\rm ij}$ are the components of a matrix $\mathbf{A} \in R^{m\times n}$, s_i is the standard deviation of point i and $s_i > 0$ ($i = 1, ..., m$) are given numbers. In the case $s_i^2 = z_i$ we can consider Eq. (A.15) as the new variant of a minimum chi-square method, which differs from the so-called modified minimum chi-square method (Eadie et al. 1971) with constraint Eq. (A.15) and sampling the starting value of the vector f. We will use the solution of the unconstrained least squares problem Eq. (4) $f_{\rm alg} = f_{\rm tikh}$ as the starting value for the minimization problem Eq. (A.15) and obtaining the solution $f_{\rm opt}$. In all calculations we take $s_i^2 = z_i$.

Acknowledgements. We thank R. Gómez-Herrero for stimulating discussions, and A. Klassen for help in SOHO/EPHIN data analysis.

References

Bertero, M., & Bocacci, P. 1998, Introduction to Inverse Problems in Imaging (Bristol: IOP Publishing, Ltd.)
Björck, A. 1996, Numerical Methods for Least Squares problems (Philadelphia: SIAM)
Calvetti, D., Lewis, B., Reichel, L., & Sgallari, F. 2004, Electronic Transactions on Numerical Analysis, 18, 153
Eadie, W. T., Drijard, D., James, F. E., Roos, M., & Sadoulet, B. 1971, Statistical Methods in Experimental Physics (Amsterdam: North-Holland)
Gómez-Herrero, R. 2005, personal communication
Groetsch, C. W. 1993, Inverse Problems in the Mathematical Sciences (Braunschweig: Vieweg)
Hansen, P. C. 1987, BIT, 27, 534
Hansen, P. C. 1992, SIAM, 34, 561
Kress, R. 1989, Linear Integral Equations (Berlin: Springer)
Lawson, C. L., & Hanson, R. J. 1974, Solving Least Squares Problems (Eaglewood Cliffs, New Jersey: Prentice-Hall, Inc.)
Müller-Mellin, R., Kunow, H., Fleißner, V., et al. 1995, Sol. Phys., 162, 483
Neumaier, A. 1998, SIAM, 40, 636
Prato, M., Piana, M., Brown, J. C., et al. 2006, Sol. Phys., 237, 61
The GEANT4 collaboration. 2006, An Object-Oriented Toolkit for Simulation in HEP, CERN-LHCC 98-44, see also: `http://wwwinfo.cern.ch/asd/geant4/geant4.html`
Tikhonov, A. N., & Arsenin, V. Y. 1977, Solution of Ill-Posed Problems (New York: Wiley)