AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE
10.1002/2017SW001702

Key Points:
- Magnetic field model optimization improves accuracy given sufficient data
- T02 model performs best for studied intervals before optimization
- TS04 model provides best representation of magnetic field after optimization

Supporting Information:
- Supporting Information S1
- Data Set S1

Correspondence to:
T. V. Brito,
thiago.brito@zoho.com

Citation:
Brito, T. V., & Morley, S. K. (2017). Improving empirical magnetic field models by fitting to in situ data using an optimized parameter approach. *Space Weather*, *15*, 1628–1648. https://doi.org/10.1002/2017SW001702

Received 28 JUL 2017
Accepted 8 OCT 2017
Accepted article online 25 OCT 2017
Published online 19 DEC 2017

Improving Empirical Magnetic Field Models by Fitting to In Situ Data Using an Optimized Parameter Approach

Thiago V. Brito[1,2] **and Steven K. Morley**[1]

[1]Space Science and Applications (ISR-1), Los Alamos National Laboratory, Los Alamos, NM, USA, [2]Now at Department of Physics, University of Helsinki, Helsinki, Finland

Abstract A method for comparing and optimizing the accuracy of empirical magnetic field models using in situ magnetic field measurements is presented. The optimization method minimizes a cost function—τ—that explicitly includes both a magnitude and an angular term. A time span of 21 days, including periods of mild and intense geomagnetic activity, was used for this analysis. A comparison between five magnetic field models (T96, T01S, T02, TS04, and TS07) widely used by the community demonstrated that the T02 model was, on average, the most accurate when driven by the standard model input parameters. The optimization procedure, performed in all models except TS07, generally improved the results when compared to unoptimized versions of the models. Additionally, using more satellites in the optimization procedure produces more accurate results. This procedure reduces the number of large errors in the model, that is, it reduces the number of outliers in the error distribution. The TS04 model shows the most accurate results after the optimization in terms of both the magnitude and direction, when using at least six satellites in the fitting. It gave a smaller error than its unoptimized counterpart 57.3% of the time and outperformed the best unoptimized model (T02) 56.2% of the time. Its median percentage error in $|B|$ was reduced from 4.54% to 3.84%. The difference among the models analyzed, when compared in terms of the median of the error distributions, is not very large. However, the unoptimized models can have very large errors, which are much reduced after the optimization.

Plain Language Summary We present a method for comparing and optimizing the accuracy of commonly used empirical models that reproduce the Earth's magnetic field for altitudes ranging from a thousand to hundreds of thousands of kilometers. This method uses magnetic field data from satellites orbiting the planet to create a "penalty function" and uses an optimization algorithm to minimize this function and find the model input parameters that produce the best results for a given date and time. Our results show that these models can be improved by the use of satellite data. The model known as TS04 produced the best results after the optimization procedure generating a smaller error in 57.3% of the points in our data set when compared to the standard (unoptimized) inputs. The optimized TS04 also outperformed the best unoptimized model by 56.2%. The differences among all the models analyzed are usually not very large; however, the unoptimized models can have very large errors, which are much reduced by the optimization.

1. Introduction

The configuration and strength of the geomagnetic field is a major controlling factor for energetic charged particle dynamics in the magnetosphere, including the dynamics of radiation belt electrons (e.g., Roederer & Zhang, 2014), ring current ions (e.g., Zaharia et al., 2006), cosmic rays, and solar energetic particle events (e.g., Desorgher et al., 2009, Kress et al., 2010). For studies requiring magnetic conjugacy between satellites (e.g., Friedel et al., 2005; Morley et al., 2013) or between satellites and ground-based instrumentation (e.g., Hones et al., 1996; Ge et al., 2012), the global morphology and mapping of magnetic field lines is crucial (e.g., Pulkkinen & Tsyganenko, 1996). Data-based modeling of the external geomagnetic field has a long history of development in space physics (see Tsyganenko, 2013, and references therein). The most widely used semiempirical field models use a modular approach, describing the major large current systems and summing their contributions to the total magnetic field (e.g., Tsyganenko, 2002a; Tsyganenko & Sitnov, 2005). The contributions of each component current system are parametrized by upstream solar wind measurements and

geomagnetic indices (Tsyganenko, 2002b, 2013). These models have been demonstrated to specify the global structure of the inner magnetosphere well on a statistical basis (McCollough et al., 2008; Woodfield et al., 2007; Zhang et al., 2010), with the later models generally performing better, although their performance varies through individual events (Huang et al., 2008; Morley et al., 2013).

Improving models of the geomagnetic field is relevant to a broad range of studies. There are many common uses of these models including the calculation of the adiabatic invariants for radiation belt studies (e.g., Iles et al., 2006; Turner et al., 2014; Yu et al., 2014), calculation of conjunction between satellites(e.g., Friedel et al., 2005), calculation of mapping from the ionosphere to the magnetotail or vice versa (e.g., Antonova et al., 2015; Ge et al., 2012; Kubyshkina et al., 2011).

A range of approaches have been used in modifying semiempirical field models to better reproduce the magnetic field during specific events (e.g., Pulkkinen et al., 1991). For example, Sergeev et al. (1990) modified the T87 (Tsyganenko, 1987) model by introducing an inhomogenous thinning of the current sheet in the near tail and modifying the central location of the current sheet, and the free parameters of the model were fit to selected magnetic field measurements. They also compared the different parametrized levels of *Kp*, finding that only $Kp = 3$ reproduced the observations. Following this work, Pulkkinen et al. (1991) modified the *Kp*-parametrized T89 model (Tsyganenko, 1989) by including an additional current sheet where the parameters controlling the tail current and the additional sheet are determined by minimizing the root-mean-squared error of 3-min averaged magnetic field measurements from three spacecraft. Other work modifying Tsyganenko models has replaced current systems in T89 and T96 (Tsyganenko & Stern, 1996) with new formulations (e.g., Apatenkov et al., 2007; Ganushkina et al., 2002; Pulkkinen et al., 2006), although these models may not be suited for general, automated application as the modified field may include unphysical artifacts (see, e.g., Kubyshkina et al., 2009, Sergeev et al., 2007).

An alternative approach to event-oriented magnetic field modeling is to treat the driving parameters (solar wind or geomagnetic indices) as free and independent parameters and simply adjust them to better match the data. In essence, this is the approach taken by Sergeev et al. (1990) when testing the different *Kp* parametrizations for the T87 model. Kubyshkina et al. (2008) used three models for event-oriented modeling of a sawtooth interval; two models modified current systems as described above, and one used the T02 model (Tsyganenko, 2002a, 2002b) with the input parameters treated as free parameters at each time step. All models, including the optimized T02 model were found to reproduce the large-scale features of the interval, although differences were also noted. More detail on this approach was provided by Kubyshkina et al. (2009) (see also Kubyshkina et al., 2011), as they introduced three variants on this optimized parameter approach. The AM01 model of Kubyshkina et al. (2009, 2011) used the T96 magnetic field model and fitted the four input parameters to minimize a weighted root-mean-squared difference between observed and modeled magnetic field components.

The present study builds on that of Kubyshkina et al. (2008, 2009) and uses magnetic field data from as many satellites as possible to optimize the input parameters for a range of semiempirical magnetic field models to try to improve their accuracy. Though the concept is essentially that of model AM01(e.g., Kubyshkina et al., 2009), we adopt a different approach to the fitting and use a wider range of data sources. We apply our methodology to four intervals selected by the Geospace Environment Modeling (GEM) focus group "Quantitative Assessment of Radiation Belt Modeling."

2. Methodology

2.1. Magnetic Field Models

The current study uses five empirical magnetic field models: T96, T01S, T02, TS04, and TS07, for which brief descriptions are presented in this section. All of them are widely used by the space physics community for a variety of applications as mentioned previously. Of these models, TS07 has a significantly different approach of calculating magnetic field values, making implementation of the optimization approach used for the other models problematic. Therefore, it is used here simply as a way to compare its results with those from the other models.

All models referred to above use a modular approach regarding the current systems driving them, where the final magnetic field vector at any given point inside the domain is calculated from the sum of the contributions from each current (see Tsyganenko, 2013). The current systems are parametrized by solar wind parameters, which are determined by spacecraft data gathered at the time of each publication (e.g., Tsyganenko, 2002b).

Table 1
List of Models and Their Input Parameters

	Model				
	T96	T01S	T02	TS04	TS07
Parameters	IMF B_Y, IMF B_Z, P_{dyn}, *Dst*	Same as T96 +G2 and G3	Same as T96 + G1 and G2	Same as T96 + W1 thru W6	multiple parameters
No. of params.	4	6	6	10	101

The T96 model (Tsyganenko, 1995; Tsyganenko & Stern, 1996) represented a significant improvement on the previous models in terms of both the magnetopause representation and the parametrization based on solar wind quantities. The model assumes a fixed shape for the magnetopause, which determines the boundary of the magnetosphere. This shape is based on an intermediate range of the solar wind dynamic pressure. It uses cylindrical harmonic functions for the magnetopause currents contribution with the purpose of shielding the dipole field, taking the Earth's dipole tilt into account. The model is parametrized by the *Y* and *Z* components of the interplanetary magnetic field (IMF B_Y, IMF B_Z), the solar wind dynamic pressure (P_{dyn}) and by the *Dst* index, see Table 1.

The general approached used in the T02 model (Tsyganenko, 2002a) is the same as in the T96 model, but it used newly developed methods to evaluate the magnetospheric field from all the current sources. Among the most important advances are the use of the field deformation technique to realistically represent the fields generated from the current systems, the use of two modules for the cross-tail current with different current densities, the addition of a partial ring current with field-aligned closure currents, and the variability of the Regions 1 and 2 Birkeland currents based on interplanetary conditions. The parameters used by this model are those used by T96 plus derived quantities which are called G1 and G2 (see Table 1).

The T01S model (Tsyganenko et al., 2003), also called TSK03 (e.g., McCollough et al., 2008), is based on the previously described T02 model, but it was constructed to be a storm time model. The data used for its parametrization contained only events with $Dst \leq 65$ nT with the goal of representing strongly disturbed geomagnetic configurations of the inner magnetosphere and their evolution during the storm cycle. The parameters used by this model are those used by T96 plus derived quantities G2 and G3 (see Table 1).

Building up on the previous work, the TS04 model (Tsyganenko & Sitnov, 2005) is a dynamical model of storm time geomagnetic fields based on data from 37 major events during the years of 1996 to 2000. This model aims at evaluating the temporal variation of the current systems using a simple model of growth and decay of the storm cycle from the satellite data. The current systems are driven by a variable calculated based on the time integral of a combination of solar wind parameters. Different relaxation time scales and baseline quiet-time intensity are used for each magnetic field source, which their partial contribution also depends on the history of the external driving. This model also uses the standard T96 parameters plus six derived quantities denominated W1 through W6 (see Table 1).

The last model used was the TS07 model (Tsyganenko & Sitnov, 2007), which is different from the others in the sense that it uses a large number of parameters (101) to calculate the geomagnetic fields, which makes it virtually impossible to use the optimization methods used for the other models. Despite that it uses a similar approach based on the same current systems as the previous models. It introduces a high-resolution expansion for the field of equatorial currents, which is capable of reproducing small variations in the magnetic field, and it is also able to better couple the equatorial and field-aligned currents. Additionally, a new database of spacecraft magnetometers was compiled corresponding to different activity levels and solar wind conditions.

To run the semiempirical magnetic field models used in the current work, solar wind parameters are used as inputs. As described above, the common parameters among all the models are the B_Y and B_Z components of the interplanetary magnetic field (IMF), the *Dst* index and the solar wind dynamic pressure (P_{dyn}). The solar wind parameters are obtained from spacecraft located at L1 (ACE and WIND) from the https://omniweb.gsfc.nasa.gov/ OMNIWEB website. There are some gaps in the data, especially before the launch of WIND; therefore, we used the formulation provided by Qin et al. (2007). This study created a database of solar wind parameters to be used by magnetic field models with a methodology to both fill these data gaps and calculate other input parameters required by some of the models, which are also derived from solar wind quantities. Specifically, the T02 and T01S models use the *G* parameters and the TS04 model uses the

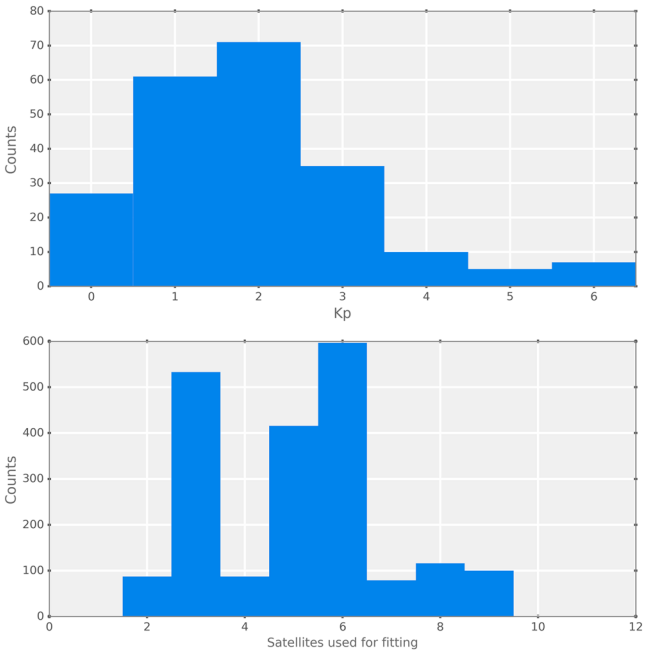

Figure 1. *Kp* distribution (top) and satellite counts (bottom) for the period of interest encompassing 21 days in 2013. For the *Kp* index, each count corresponds to a 3 h period, while for the satellites, the interval is 15 min. The satellites counted (bottom) exclude those used for validation.

W parameters. These set of parameters depend on the time history of other solar wind parameters, more specifically IMF-B_Y, IMF-B_Z, and P_{dyn} and therefore are not completely independent from them. Details about the calculation of these parameters are described in Qin et al. (2007) and can also be found in the references corresponding to each model mentioned above. All these input parameters are commonly referred to as Qin-Denton (Q-D) parameters, after the two main authors of this work. In terms of the number of parameters used by each model, T96 uses 4, T02 and T01S use 6 (not the same parameters) and TS04 uses 10. This information is summarized in Table 1.

2.2. Observational Data

To obtain the solar wind parameters needed to run the models, we used data from both the ACE and WIND spacecraft acquired from OMNIWEB, as well as the *Dst* index provided by the University of Kyoto. Magnetic field data from the following sources were used: the Electric and Magnetic Field Instrument Suite and Integrated Science (Kletzing et al., 2013) on the twin Van Allen Probes (also known as Radiation Belt Storm Probes—RBSP) (Mauk et al., 2013); the fluxgate magnetometers on the Geostationary Operational Environmental Satellite (GOES) platforms (Singer et al., 1996); the fluxgate magnetometers (FGM) (Auster et al., 2009) on the Time History of Events and Macroscale Interactions during Substorms (THEMIS) mission (Angelopoulos, 2008); and the Cluster Magnetic Field Investigation (FGM) (Balogh et al., 2001) flown on the Cluster constellation (Escoubet et al., 2001).

The extent of time chosen to perform the comparisons and optimizations done in the current study consisted of four different periods during the year of 2013 totaling 21 days. These periods were 16–20 March, 1–5 June, 17–21 September, and 22–27 September 22-27. These periods were chosen because they cover the Geospace Environment Modeling (GEM) Challenge events selected by the Quantitative Assessment of Radiation Belt Modeling (QARBM) Focus Group and because they contained different levels of activity covering both quiet-time and storm-time periods. During this 21 day period, the models were evaluated every 15 min, yielding 96 model evaluations per day. Therefore, the entire period contains 2,016 points in time where magnetic field data from each model at each satellite position were collected. Figure 1 shows the distribution of observed *Kp* values and a histogram of the number of satellites available to be used in the optimization process for this entire period. The mode number of satellites used in our optimization is six and the intervals used cover a wide range of magnetic activity level (as measured by the *Kp* index).

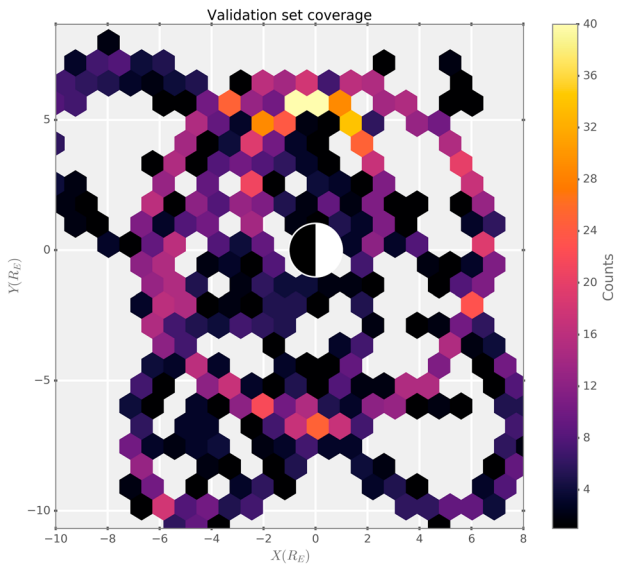

Figure 2. Spatial coverage of satellites used in the model validation during the period of interest, projected on the *X*-*Y* plane in GSM coordinates. This is a random sample of the full set of satellite data used in this study and so is similar to the coverage of the data used for fitting.

To perform the optimization, or *fitting*, we selected spacecraft based on the following criteria for radial distance (R), magnetic latitude (MLAT), and distance between satellites (d): $R > 1.5\,R_E$ and $R < 12\,R_E$; MLAT > -60 and MLAT < 60; $d > 0.3\,R_E$. In addition to that we selected one satellite at random at each point in time, excluding it from the fitting process, to be used in the *validation set*. This set, therefore, contains 2,016 data points—one for each time step in the entire period of investigation—from randomly chosen satellites that fit these criteria but were not used in the fitting process. By randomly selecting the validation satellite at each time step, we mitigate the effect of serial correlation in the state of the magnetic field model. If the model configuration is varying slowly relative to our 15 min cadence then by choosing a random satellite for validation we sample a range of locations within the model domain; subsequent measurements and evaluations at any given satellite are likely to vary slowly, such that the observations and errors would both be autocorrelated. All data included are within the published ranges of model validity and are confined to $R < 12\,R_E$. Figure 2 is a 2-D histogram showing the spatial coverage of the validation set projected on the *X*-*Y* plane for the entire

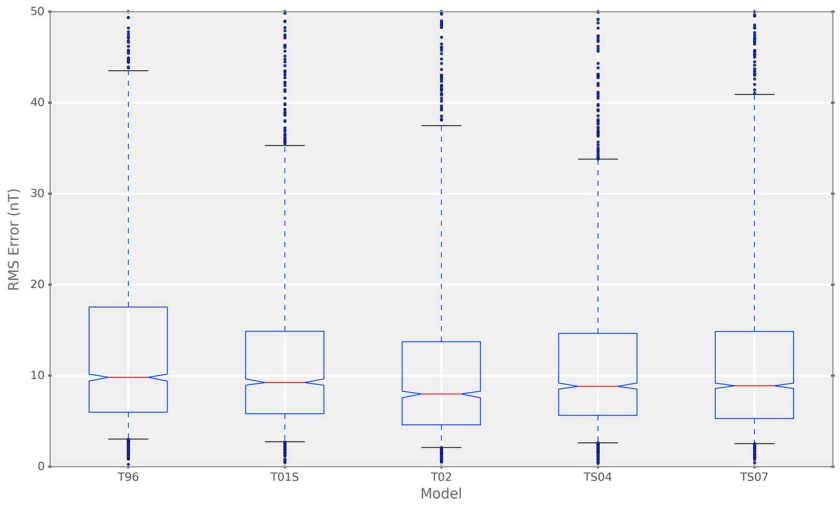

Figure 3. Comparison of the distribution of the RMS error for the *validation set* among the magnetic field models using the standard solar wind input parameters for each model. Red line inside the box shows the median of the distribution, with the notches resolving the median error. The box is defined by the first and third quartiles of the distribution and the whiskers are given by the 5th and 95th percentiles.

period of interest. As the validation set is a random sample of the full data set, the spatial coverage is similar for the data used in the optimization.

3. Results

3.1. Comparison Between Unoptimized Models

The data set described above provides a valuable way of comparing the performance of all the models used in this work. Many kinds of metrics could be used to compare the accuracy of models (e.g., Huang et al., 2008; McCollough et al., 2008); however, one common metric that has been used in other model fitting (and validation) studies is the root-mean-square (RMS) error (Kubyshkina et al., 2009; McCollough et al., 2008; Tsyganenko, 2002b).

$$\epsilon_{\mathrm{RMS}} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2\right)}, \tag{1}$$

where n is the number of measurements, y_i is the ith model value for the quantity and x_i is the ith observation value for the quantity. In the case of a three-dimensional quantity such as the magnetic field, for each observation point the error in each component is squared before the square root is applied:

$$\epsilon_j = \sqrt{\left(B_x^{\mathrm{mod}} - B_x^{\mathrm{obs}}\right)^2 + \left(B_y^{\mathrm{mod}} - B_y^{\mathrm{obs}}\right)^2 + \left(B_z^{\mathrm{mod}} - B_z^{\mathrm{obs}}\right)^2} \tag{2}$$

and the final error is the average of all the individual RMS errors from each observation point.

$$\epsilon_T = \frac{1}{n}\sum_{j=1}^{n}\epsilon_j \tag{3}$$

Figure 3 shows the distribution of the RMS errors in nT for all models using the validation set. The distributions are quite skewed, having a long tail for all the models. The box plot shows the median of the distribution as a red line across the box, with the 95% confidence interval of the median denoted by notches in the box, the height of the box itself is given by interquartile range (IQR) of the distribution and the whiskers mark the 5th and the 95th percentiles of the distribution. Each point shown outside the whiskers are actual error values in the distribution. The plot shows a lot of points outside the whiskers, which indicates a large number of outliers (i.e., a heavy-tailed error distribution). There are in fact a few dozen points (of order 1% of the validation set) for each model that are >50 nT and are thus not shown in the plot. Using the median of the error distribution as a metric to compare the models, the plot shows that the T02 model has the best performance. That is,

Table 2
Summary of the Quartiles, 95th Percentile and the Mean of the Distribution of RMS Errors for Each of the Tested Magnetic Field Models

	Model				
	T96	T01S	T02	TS04	TS07
25th p.	5.99	5.82	4.60	5.65	5.29
Median	9.83	9.24	7.99	8.82	8.90
75th p.	17.55	14.89	13.73	14.66	14.86
95th p.	43.60	35.36	37.63	33.80	40.95
Mean	17.54	15.45	14.69	15.13	16.08

50% of tested evaluations had an RMS error of 7.99 nT or lower (see Table 2). It also shows that T02 has the smallest values of all the percentiles analyzed except the 95th. While the median RMS error is smallest for T02 for our validation set, the differences between the other models are not significant; this can be seen by inspection of the plotted confidence intervals. The median RMS errors are of similar magnitude for all tested models, and they all are close to 9 nT. Models T01S, T02, TS04, and TS07 have roughly the same IQR indicating that the spread of the error is similar, although T96 has a slightly larger IQR. The quartiles, 95th percentile, and the mean of the RMS errors for each model are summarized in Table 2.

Even though the median values of the RMS error distribution for all models are similar, the median value of the RMS error is limited as a metric for comparing these models. For instance, it is not possible to know which of the three components of the magnetic field has the largest contribution or if the error is comparable among the components. Technically, the RMS error in this case can be interpreted as the magnitude of a vector ($\vec{E}$) that would correct the model vector to be equal to the observed vector ($\vec{B}^{\text{mod}} + \vec{E} = \vec{B}^{\text{obs}}$). However, no information can be extracted about the direction of this vector $\vec{E}$ from the magnitude of the error. Furthermore, the RMS error is scale dependent, inferring that errors of the same order (e.g., factor of 2 difference) are not given equal importance. In other words, the RMS error does not take the observation (base) value in consideration. When referring to magnetic fields in the magnetosphere, where values can vary across a few orders of magnitude, we are more interested in the relative error of the model, and so it is meaningful to know the relative importance of an absolute value, that is, a 5 nT error has very different significance depending on the value of the related observation. The value of the RMS error itself contains no information about the relative significance with respect to the observed value.

Other work has used a modified RMSE when fitting. For example, Tsyganenko and Sitnov (2005) introduced a weighting scheme to normalize the data coverage during fitting to account for regions with limited data coverage. Their weights were calculated as a function of the radial distance, such that weight on measurements in the tail were typically increased (but were constrained to be weighted by a factor $\leq$5). This weighting normalized the effective number of data points per radial bin but also likely mitigated the shortcomings of the RMS error described above. Kubyshkina et al. (2009) also used a modified RMSE to deemphasize regions that had multiple satellites close to each other. They weighted each ϵ^i_{RMS}, where the weight was reduced for spacecraft that were less than 3 R_E apart, and used the weighted mean of the RMS errors as their fit metric.

As described in section 2.2 we exclude data that are within 0.3 R_E of an included data point in our fitting procedure, which will be described in section 3.2; given the very small separations between the Cluster satellites in 2013 (Escoubet et al., 2015), this means that typically only a single Cluster satellite is used for fitting at this time. The different orbital characteristics of the other satellites used in this ensure that the data used for fitting are not usually clustered near any particular location and we therefore do not weight the data. We do, however, explicitly account for the changing importance of a given magnitude of error with the magnitude of the field using a relative error on the magnitude. We also explicitly measure the directional error, as will be discussed below.

To characterize the error in the predicted field magnitude, we use a relative error which we define as

$$\epsilon_{\text{mag}} = \left| \log_2 \frac{|\vec{B}^{\text{mod}}|}{|\vec{B}^{\text{obs}}|} \right| \tag{4}$$

This metric has several advantages over using the RMS error. First, it is an order-dependent error metric; in other words, it is essentially a fractional error relative to the magnitude of the observed magnetic field (see, e.g., Morley, 2016). Second, by log transforming the ratio of the predicted and observed magnitudes this metric is symmetric with respect to overprediction or underprediction, that is, a factor of 2 difference above or below the observed value produces the same error value. One drawback of this metric is that it is not easy to interpret, although a magnitude percentage error can be derived from it (Morley, 2016). The accuracy of the predicted magnitude over the set of validation points is shown in Figure 4. One box plot is shown for each tested model, and the *Y* axis shows the magnitude error converted to a percentage error (cf. Morley, 2016).

$$\varepsilon = 100\,(2^{\epsilon_{\text{mag}}} - 1) \tag{5}$$

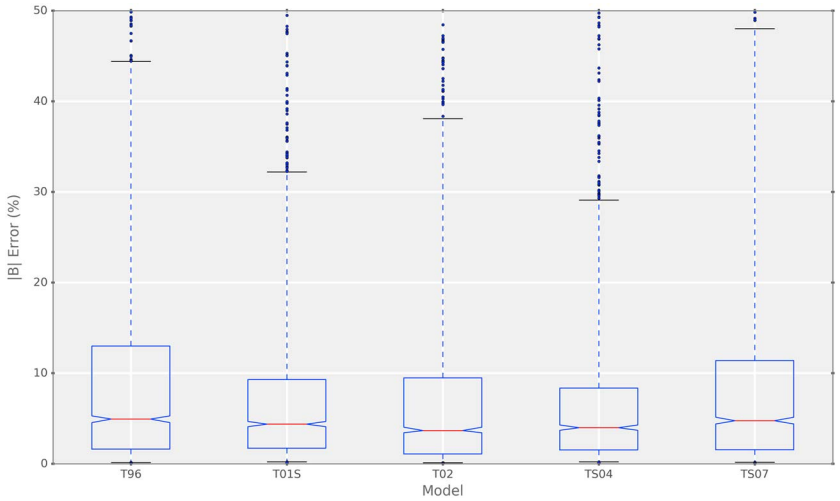

Figure 4. Similar to Figure 3. Comparison of the distribution of the error in the magnitude of the magnetic field for the *validation set* among the magnetic field models using the standard solar wind input parameters for each model.

Comparing Figures 4 to 3 shows that the IQR (box sizes), skewness, and number of outliers are similar. It also shows that based on the median of the distribution, again T02 and TS04 have the best performance among the models analyzed. In fact, TS04 has a better performance among the models with this metric compared to the RMS error. It is interesting to note that when using the RMS error, the more recent models (TS04 and TS07) are found to be similar in their accuracy; when using ϵ_{mag} to measure the accuracy, TS04 is notably better, as the median error is low and the IQR is small.

It is also important to analyze the bias of the magnitude error distributions. Figure 5 captures that information showing a comparison of the residual error among all tested models ($e = |\vec{B}^{\text{mod}}| - |\vec{B}^{\text{obs}}|$) as a function of $|\vec{B}^{\text{obs}}|$, where the Y axis is $\log_{10} e$, the X axis is $\log_{10} B_{\text{obs}}$, and the color scale denotes the number of observation points at that location. Positive values on the Y axis imply overprediction, while negative values imply underprediction. It is clear that the relative residual error is, in general, greater for values below $X \sim 2$, or $|\vec{B}^{\text{obs}}| \sim 100$ nT. This corresponds to points further out in the magnetosphere where the fields fluctuate more in response to variations in the solar wind, so a higher error is indeed expected in this region. All models seem to show a trend of over predicting for very small values of $|\vec{B}^{\text{obs}}|$ ($X \simeq 1$) and under predicting for intermediate values of $|\vec{B}^{\text{obs}}|$ ($1.5 < X < 2$).

We now introduce a metric for explicitly examining the error in the direction of the magnetic field vectors. The angular error is defined to be

$$\epsilon_\theta = 1 - \left(\frac{\vec{B}^{\text{mod}} \cdot \vec{B}^{\text{obs}}}{|\vec{B}^{\text{mod}}||\vec{B}^{\text{obs}}|} \right)^2 \tag{6}$$

Using the definition of cosine between two vectors

$$\cos\theta = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|} \tag{7}$$

and the fundamental law of trigonometry

$$\sin^2\theta + \cos^2\theta = 1, \tag{8}$$

we can see that ϵ_θ is equivalent to $\sin^2(\theta)$, where θ is the angle between $\vec{B}^{\text{mod}}$ and $\vec{B}^{\text{obs}}$ that is less than 180°. This is a straightforward metric for the angular error, and the actual angle θ can be directly calculated from it. In the case when θ is greater than 90°, the value of ϵ_θ would be decreasing with increasing θ. Despite this being a rare occurrence, to correct it, in this circumstance we construct ϵ_θ differently: $\epsilon_\theta = 2 - \sin^2(\theta)$. That way, the ϵ_θ function is continuous and monotonic for any value of θ between 0° and 180°. Figure 6 compares the angular

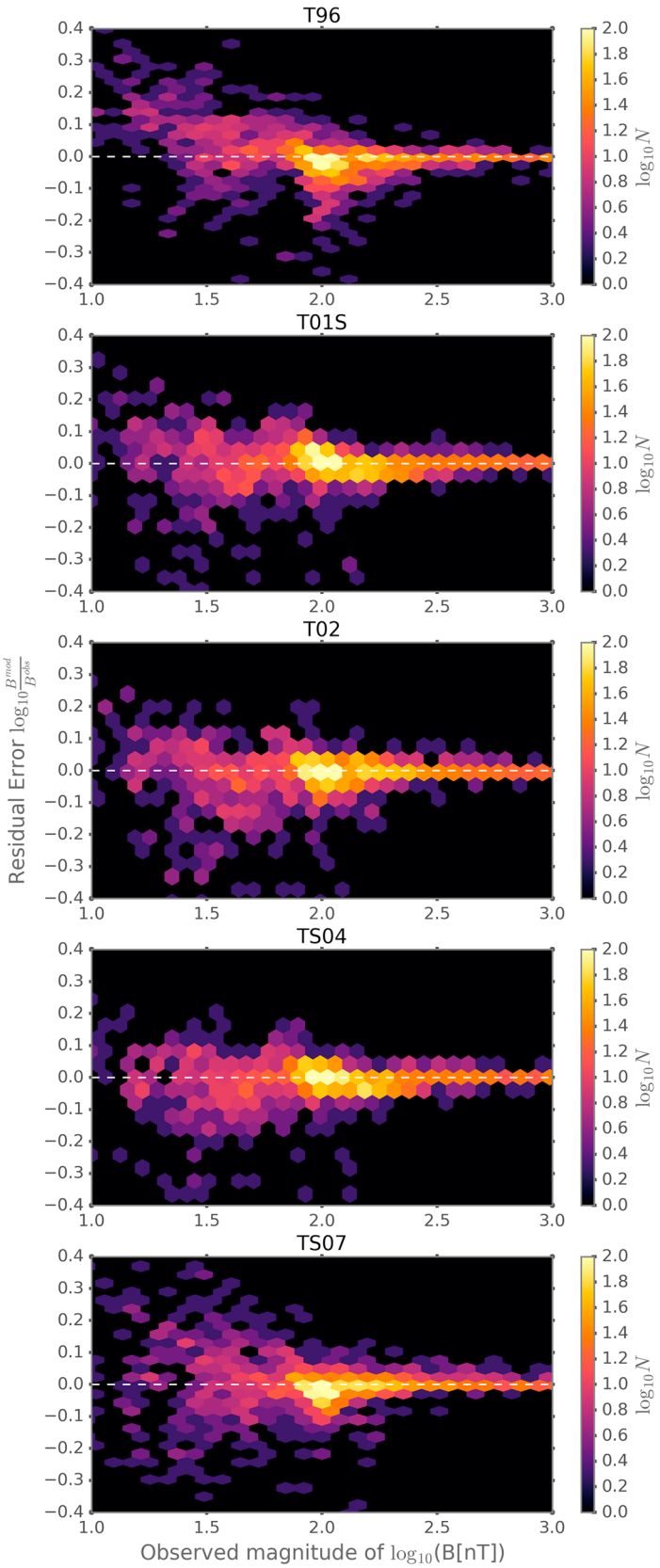

Figure 5. Comparison of the bias of the magnitude error of the magnetic field for the *validation set* among the magnetic field models using the standard solar wind input parameters for each model.

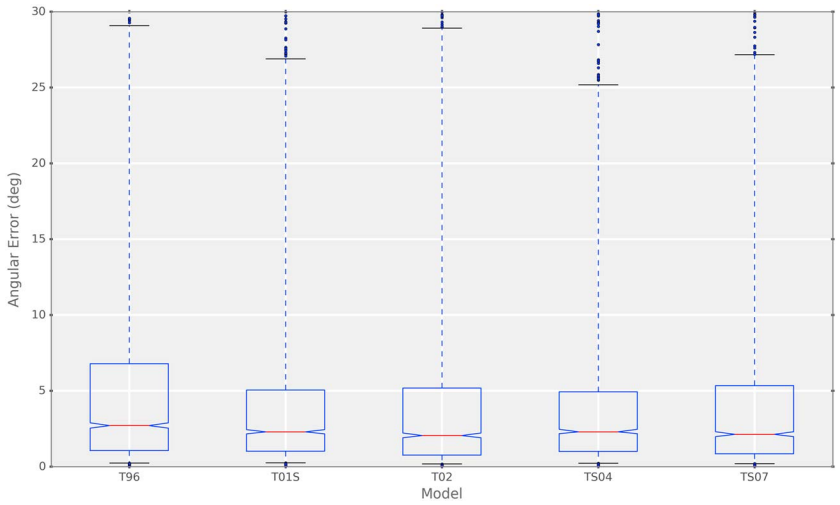

Figure 6. Similar to Figure 3. Comparison of the distribution of the angular error of the magnetic field for the *validation set* among the magnetic field models using the standard solar wind input parameters for each model.

difference θ between all models. It shows that using this metric, all models perform in a similar way with the 75th percentile being close to 5° for all of the models except T96, which performs notably worse. The relative performance of each model in terms of directional error is remarkably similar to the relative performance in terms of the error in field magnitude.

3.2. Using Data to Improve Model Performance

It is possible to use the magnetic field observation data from the satellites available in the magnetospheric region to improve the accuracy of these semiempirical magnetic field models by trying to fit the model to the data. This strategy has been used previously in other studies (e.g., Kubyshkina et al., 2008, 2009). A *cost function* can be implemented to that end based on the previously used error measurements of the magnitude of B (ϵ_{mag}) and the angular difference (ϵ_θ):

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\left(\epsilon_{\text{mag}}^{(i)} + \alpha\epsilon_\theta^{(i)}\right), \tag{9}$$

where N is the number of satellites available on a particular date and time that satisfy the selection criteria (previously described in section 2.2), i is the summation index over the satellites, and α is a weighting factor for the angular term. This cost function τ can be used as a measure of the combined error from both the magnitude part and the angular part for all the available magnetic field measurements at a particular date and time.

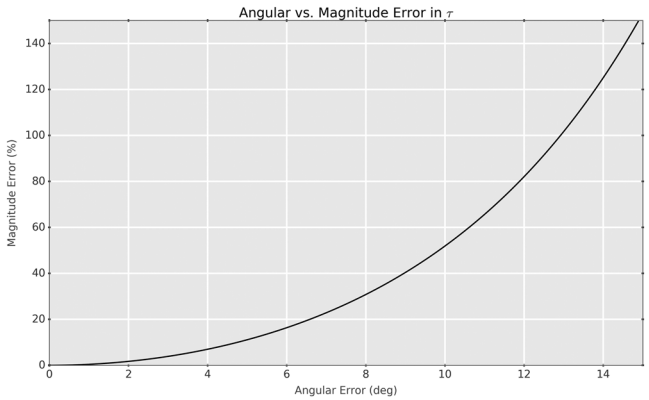

Figure 7. Functional relationship between the angular error and the magnitude error in the cost function τ valid for the results in this study.

Throughout the paper we used $\alpha = 20$, which was chosen so that the value of the angular term is close to the value of the magnitude term on average over the entire set, causing both terms to have roughly the same contribution to the total value of the τ function. For reference, with this value of α, a 2% error in the magnitude has the same weight as a 2.2° difference in the angle between model and observed magnetic field vectors, and a 20% magnitude error corresponds to an angle difference of 6.6°. Figure 7 shows the functional relationship between the angular error and the magnitude percentage error present in τ that is valid for the results presented here. The simple way of interpreting this figure is to consider that for every point in the curve, $\epsilon_{\text{mag}} = 20 \times \epsilon_\theta$, and therefore, the two components of the cost function have the same weight. For errors of up to ≃10% in magnitude and ≃5° in direction the relationship can be considered approximately linear and as the angular error increases it is weighted more heavily than the magnitude error.

Before using the τ metric for fitting, it is instructive to do a comparison between all models based on it, similar to the previous comparisons. Figure 8

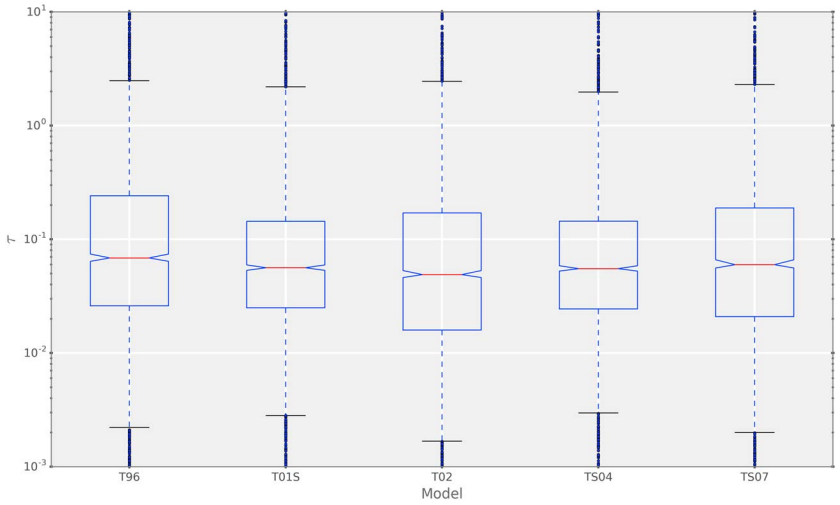

Figure 8. Comparison of the distribution of the τ metric, represented in log scale, based on the angular and magnitude errors of the magnetic field for the *validation set* among the magnetic field models using the standard solar wind input parameters for each model. Box markers and whiskers are defined similarly to Figure 3.

shows the results of this comparison. Here the same validation data set was used as in Figure 3, where one satellite per time, chosen randomly, was reserved for validation purposes. This time, the figure shows the *Y* axis in log scale for better visualization. Again, T02 had the best performance of all the models when examining the median of the error distribution, as well as the 25th and 5th percentiles. Both T01S and TS04 outperformed T02 when comparing the 75th and 95th percentiles, as well as having a smaller spread in the error distribution than T02. All tested models show similar performance, with some small but interesting differences. The lower quartile of the error distribution from the T02 model is much lower than for the other models; however, T01S and TS04 have similar medians and lower upper quartiles and 95th percentiles.

3.3. Optimization Process

The cost function τ may also be used to optimize the models using the satellite data available, since it depends implicitly on the model parameters (discussed in section 2.1). The number of parameters to be optimized depends on the model itself (these are summarized in Table 1). As discussed previously, the TS07 model cannot easily be optimized in this manner since, by construction, it depends on a large number of parameters.

The objective of the optimization process is to find values of the input parameters for each model that provide the best answer based on the observational data available. The best answer in this case is given where the cost function has its minimum value. As discussed in section 2.1, some of the input parameters (Gs and Ws) are not completely independent from others (IMF-B_Y, IMF-B_Z, and P_{dyn}). However, during the optimization process, they are treated as free parameters. Any interdependency between parameters is irrelevant for the purpose of finding the minimum value of the cost function in the parameter space. Interdependency between parameters often translates into a smoother cost function, which is a desired property in terms of the stability of the optimization algorithm.

To find the parameters that give a minimum in τ in the model parameter space we use a multidimensional search procedure to minimize τ. To improve the stability and robustness of the optimization procedure, we perform the optimization after first normalizing the parameters such that they lie in the interval (0,1). Once the optimal normalized parameters are found, the normalization is undone and we recover the optimal model input parameters.

The procedure for finding the optimal model parameters for each specific date and time, and for each model, are as follows and were repeated for the entire validation set in the present study:

1. apply selection criteria to choose which satellites to use
2. randomly select one satellite to set aside for validation
3. find the model input parameters to be used for the specific date and time
4. find the value of the three components of the magnetic field observed by the satellites selected ($\vec{B}^{\mathrm{obs}}$)
5. evaluate this model to find the magnetic field components at the locations of the satellites ($\vec{B}^{\mathrm{mod}}$)

6. calculate the sum of τ for the specific date and time based on $\vec{B}^{obs}$ and $\vec{B}^{mod}$
7. apply the optimization algorithm to τ using the model parameters as variables
8. find model parameters that minimize τ
9. evaluate model field using optimized parameters to test improvement

There are many optimization algorithms that one could use to find the minimum point in the τ function. In this study, we chose to use the Nelder-Mead algorithm (Nelder & Mead, 1965). This "direct search" method is based on a simplex (a triangle in 2-D space, a tetrahedron in 3-D space, and so on), where function values are evaluated at the vertices. The method extrapolates the behavior of the function based on the value at the vertices and takes a series of steps trying to converge to a global minimum by performing changes in the shape of the simplex at each step such as a reflection, an expansion, or a contraction (see, e.g., Kiusalaas, 2010). This method has several advantages: it is efficient for low-dimensional problems (Lewis et al., 2000); derivatives of the cost function do not need to be calculated (e.g., Han & Neumann, 2006); it is generally robust (Kiusalaas, 2010) (but see also Lewis et al., 2000); and it is straightforward to apply to a wide range of optimization problems (Olsson & Nelson, 1975).

It can be difficult to find the global minimum of a function if it has multiple local minima or a noisy cost function. We assume that the standard input parameters for each model are likely to be close to the global minimum and use these to determine the initial simplex. However, to ensure that we were finding a global minimum, we tested 100 random sets of initial parameters for the simplex for each optimization. This testing suggested that the cost function could be noisy but that the optimal parameters found were generally tightly clustered. Our testing showed that this method consistently found a global minimum with little variation in the optimized parameters depending on the starting position, so we recommend using the original model input parameters to determine the starting simplex.

3.4. Comparison Between Optimized Models

The optimization process described above was performed for the entire validation set and for four models (T96, T01S, T02, and TS04). It is worth stressing that the randomly chosen validation set does not contain any of the satellites used for fitting. In that sense, our optimization results presented in this study is a tentative approach to improve model results in the magnetosphere within the preestablished constraints described earlier. Using different constraints would likely produce different results since the set of satellites available for fitting would be different at times.

Figure 9 shows a comparison between the original models (i.e., models where the standard Qin-Denton parameters were used) and the optimized models in terms of the distribution of the cost function τ, shown in a box plot format. As described previously, the median is given by a red line across the box, the height of the box itself is given by the interquartile range of the distribution. In this and following box plot figures, the whiskers mark the 10th and the 90th percentiles of the distribution, differently from what is shown in previous box plots. The distributions for all four models have a long tail indicating that all models produce a significant number of outliers in their magnetic field estimates (either in the magnitude term, the angular term, or both). Using the median as a form of comparison, we notice that the optimization procedure improved the T96 model, while the results for the TS01 model were slightly worse. The other two models showed very little difference using this metric. Utilizing the 90th percentile as a parameter for the comparison provides a different picture. All models had a noticeable improvement from the optimization, using this parameter. The τ value for this metric was on average 20% smaller. The ability to significantly reduce the number of large errors seems to be an important feature of the optimization process.

During the fitting process, the number of satellites available that satisfy the selection criteria varies for each date and time, as shown in Figure 1. The number of satellites used for fitting is a critical parameter in the optimization process (e.g., Kubyshkina et al., 2009). Therefore, as an additional way of testing the optimized models, we designate a minimum number of satellites to be used for fitting. We chose to consider points in time where at least six satellites are available to be used for fitting (i.e., seven satellites are available, since one is set aside for validation). The comparison between original and optimized models when using at least six satellites for fitting is shown in Figure 10. Again, using the median as a parameter of comparison, an improvement is noticeable in the T96 and TS04 models; TS04 had not shown improvement when using the entire validation set. T01S shows in this figure no improvement from optimization while it was slightly worse in the previous figure, and T02 again showed no change from optimization. Using both the 90th percentile, denoted by the top whisker mark, and the 75th percentile, denoted by the top line of the box, the plots show again

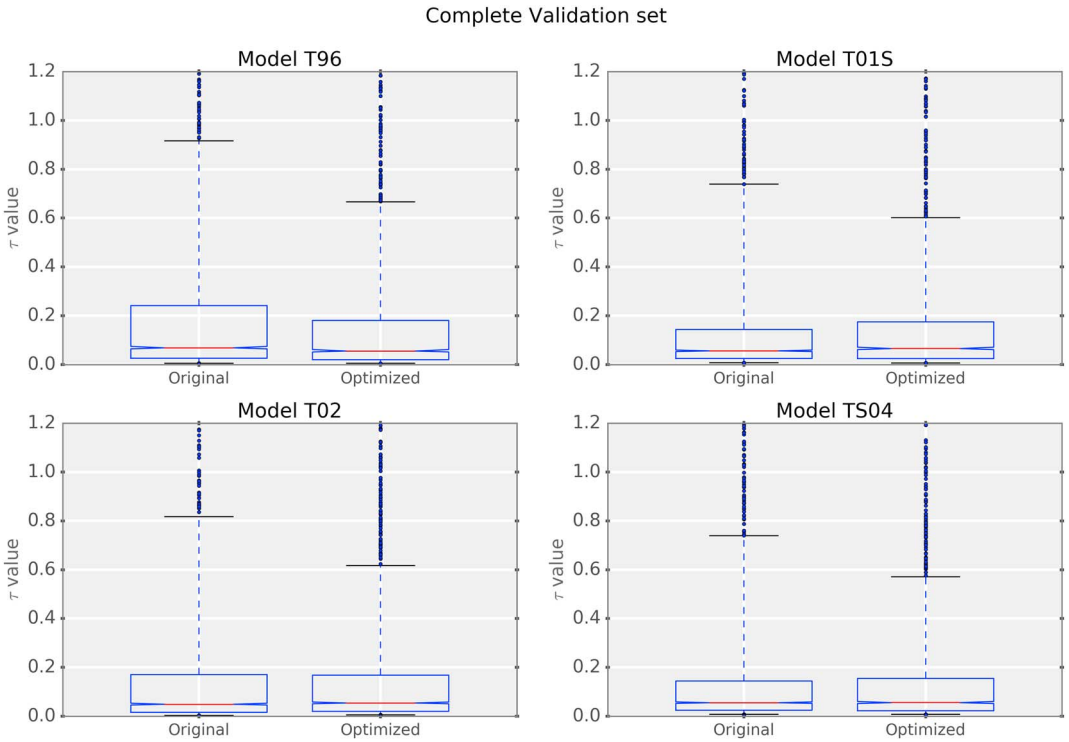

Figure 9. Comparison between the error distributions, based on the τ metric, of the optimized and unoptimized results from four magnetic field models (T96, T01S, T02, and TS04) using the full validation set. The box plots are similar to those in Figure 3, except the whiskers are given by the 10th and 90th percentiles.

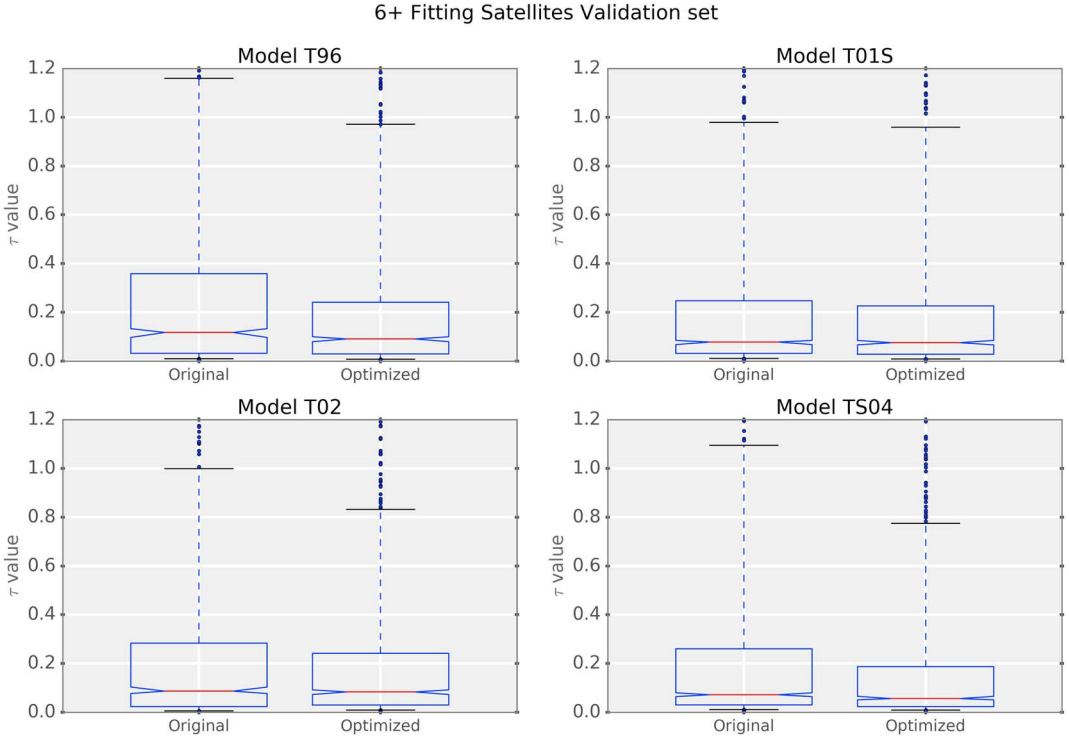

Figure 10. Comparison between the error distributions, based on the τ metric, of the optimized and unoptimized results from four magnetic field models (T96, T01S, T02, and TS04) using the 6 + *validationset*. The box plots are similar to those in Figure 9.

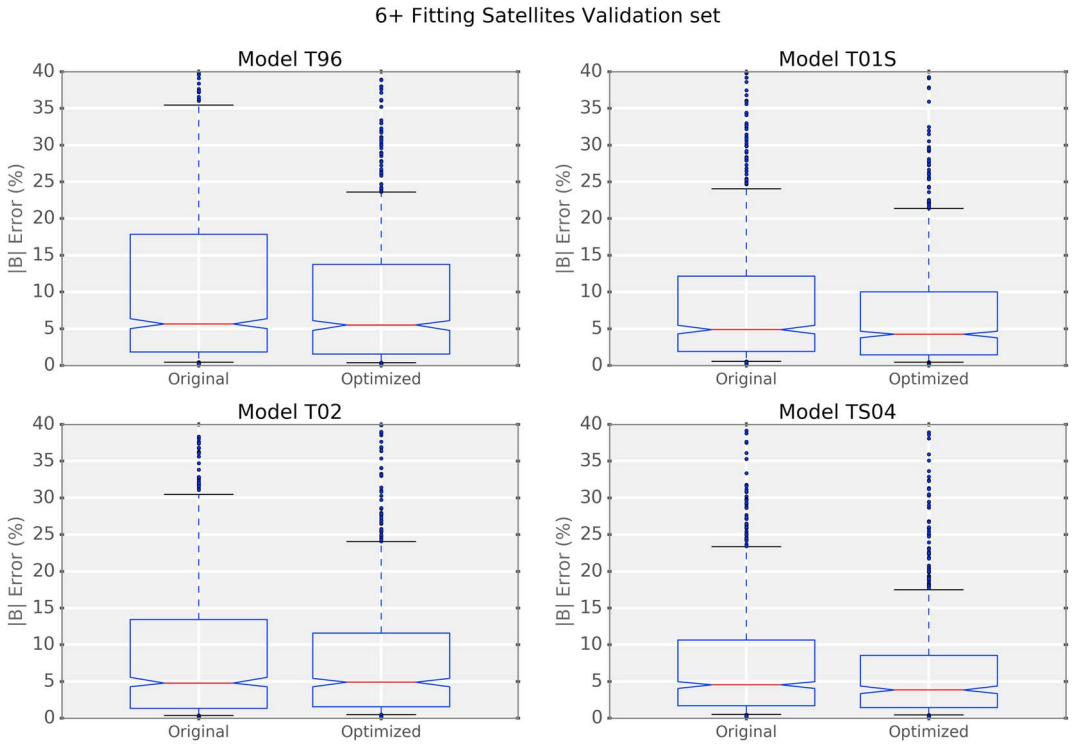

Figure 11. Comparison between the magnitude error distributions of the optimized and unoptimized results from four magnetic field models (T96, T01S, T02, and TS04) using the 6 + *validationset*. The box plots are similar to those in Figure 3.

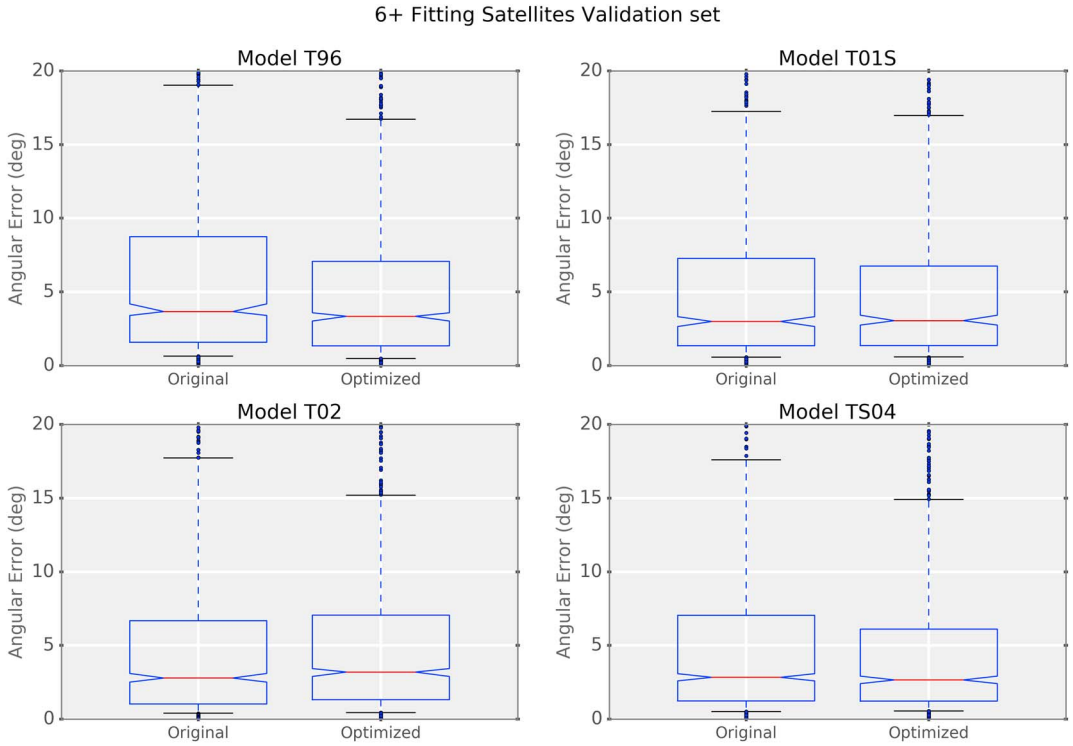

Figure 12. Comparison between the angular error distributions of the optimized and unoptimized results from four magnetic field models (T96, T01S, T02, and TS04) using the 6 + *validationset*. The box plots are similar to those on Figure 3.

Table 3
Comparison Between the Optimized and Unoptimized (Original) Versions of the Four Magnetic Field Models Analyzed

	Model							
	T96org.	T96opt.	T01Sorg.	T01Sopt.	T02org.	T02opt.	TS04org.	TS04opt.
τ med.	0.1174	0.0908	0.0781	0.0755	0.0866	0.0832	0.0717	0.0559
τ 90th p.	1.162	0.971	0.994	1.010	1.006	0.837	1.113	0.782
|B|% med.	5.64	5.51	4.88	4.25	4.78	4.90	4.54	3.84
|B|% 90th p.	36.0	23.6	24.6	21.4	31.0	24.1	23.4	17.7
Ang. med.	3.66	3.34	2.99	3.05	2.79	3.19	2.84	2.66
Ang. 90th p.	19.1	16.8	17.6	17.0	17.7	15.2	17.8	14.9

Note. The table summarizes the results shown in Figures 10–12 comparing the median and 90th percentiles of the τ, magnetic error, and angular error distributions using the 6 + *validationset*.

significant improvement in performance, except for the T01S model. It is not clear why T01S had a very similar performance in this metric; perhaps the fact that it is constructed as a storm model while our set contains mostly quiet time periods is relevant.

The same analysis done for the two previous figures can be done for the two terms of the τ cost function separately, the magnitude term and the angular term. Figures 11 and 12 show the breakdown of τ in terms of the distribution of its two terms again only considering the set with six or more observations per time step. The magnitude term is converted to a magnitude percentage error with respect to the observational value and the angular term is converted to the angle between observed and model magnetic field vectors that is less than 180° .

Figure 11 shows that all models have a significant improvement in terms of the 90th percentile, with T96 showing the biggest improvement and TS01 showing the smallest. In terms of the median, models T01S and TS04 show some improvement, while models T96 and T02 show no change. Overall, all models showed some improvement in the magnitude term.

Figure 12 shows that all models had a significant improvement when using the 90th percentile for comparison. TS04 showed the biggest improvement, while TS01 showed the smallest. Using the median, models T96 and TS04 show some improvement, model T01S shows no change, while model T02 shows a worse performance. The results from the three previous figures are summarized in Table 3.

Another way of comparing the results from all the optimized models and with the results from the original models—that is, using the standard model input parameters—is to track the results from each time step in the set and to evaluate the percentage of elements in the set where the optimized model produces a better result than the original model. Table 4 shows these results with both the entire validation set and only with the set where there are at least six satellites available for fitting (6+ set). As expected, the optimization provides better results in general with the 6+ set, with only a slightly worse result for the T96 model, interestingly. With this set, all optimized models produce better results than their unoptimized counterparts more than 50% of the time demonstrating that the optimized models have more skill. We can also compare the optimized models with the best unoptimized model (T02). Table 5 presents that comparison showing that only the optimized TS04 model using the 6+ set produces a better result than the unoptimized T02, proving the need for sufficient data when fitting to observations. The former produces a more accurate result 56.2% of the time in the 6+ set. The other two optimized models (T96 and T01S) produced slightly inferior results with this set when compared to T02.

Table 4
Comparison of Percentage of Validation Points Where the Optimized Model Improves on the Unoptimized Model

	Model			
	T96	T01S	T02	TS04
Full set	59.7%	48.8%	48.3%	51.2%
6+ set	58.2%	52.1%	52.2%	57.3%

Note. A score of more than 50% means that the optimized model is outperforming the unoptimized model, suggesting that the optimization procedure has improved the model skill.

Conversely, an analysis of the cases where the optimization produced much worse results was also done. Consider a "failure" as a case where using the original parameters resulted in the value of the cost function being within the 75th percentile of the error distribution, while using the optimized parameters resulted in its value being greater than the 95th percentile of the distribution. Using this assumption, all four models where the optimization was performed, using both sets (complete and 6+), produced an

Table 5
Comparison of Percentage of Validation Points Where the Optimized Model Improves on the Unoptimized T02 Model

	Model		
	T96	T01S	TS04
Full set	47.3%	42.8%	46.2%
6+ set	48.4%	49.9%	56.2%

Note. A score of more than 50% means that the optimized model is outperforming the unoptimized T02 model.

insignificant percentage of "failures," although nonzero, varying between 0.6% in the worst case and 0.1% in the best case. One notable point from this analysis is that TS04 was the worst model when using the complete set (0.6% failures) and became the best one when using the 6+ set (0.15% failures).

To illustrate the differences in mapping between the original and the optimized parameters, a brief comparison of foot points using the two versions of the TS04 model was done. Figure 13 provides a summary of this comparison. A point in the nightside SM equatorial plane ($-8\,R_E, 0, 0$) was used as reference, which is expected to be in the transition region where the stretched tail-like configuration is dominant (e.g Murphy et al., 2014). For each time step of the 21 day period, field lines were mapped from that point to the Northern Hemisphere, if the field line was closed. Then, the magnetic latitude (Λ) and magnetic local time (MLT) of the corresponding magnetic foot point was calculated. The two panels on the left show the histogram of the differences in Λ ($\Lambda_{opt.} - \Lambda_{orig.}$) and MLT ($MLT_{opt.} - MLT_{orig.}$) for the entire data set, considering only the closed field lines. In both cases, the distribution is fairly symmetric, with the mean close to zero, showing that there is minimal systematic bias between the optimized model and the unoptimized model. The standard deviation of both distributions is relatively small showing that the foot point locations for original and optimized models are close on average. The spread in magnetic latitude is about 1.1°, meaning that just over 1/3 of comparisons show a latitudinal difference greater than a degree. The two panels on the right show a more direct comparison between the foot points calculated with both versions of TS04 for quiet and storm times. Both panels show foot point locations during a 4 h period (time is color coded) when six or more satellites were available for fitting. The bounds of both *X* and *Y* axis are different in both plots, but the range is the same (8° for Λ and 1.6 h for MLT). The predominantly northward motion of the foot point during this quiet time interval and westward motion during this storm time interval are not necessarily representative of these activity levels. It can be noticed in both panels that using the original parameters produces a slightly smoother foot point movement. However, both versions produce similar results where the foot points show the same tendency of movement and with generally small differences. The early part of the storm time interval shows a more systematic offset in the mapped magnetic latitude

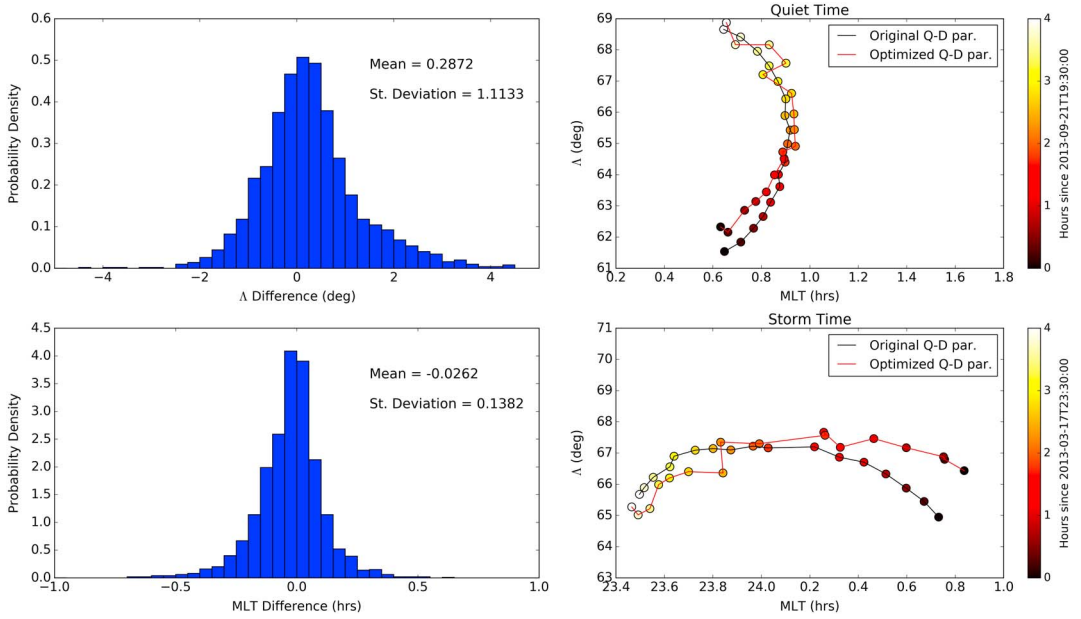

Figure 13. Comparison of mapped foot point locations starting from a point in the near tail at (−8,0,0) R_E (SM) using the original and optimized versions of TS04. Top left panel shows the histogram of the magnetic latitude (Λ) difference between both model versions. Bottom left panel shows the histogram of the MLT differences. Top right panel shows foot point locations during a 4 h quiet period for both versions where at least six satellites were available for fitting the optimized model. Bottom right panel shows foot point locations during a 4 h storm period for both versions where, again, at least six satellites were available for fitting the optimized model.

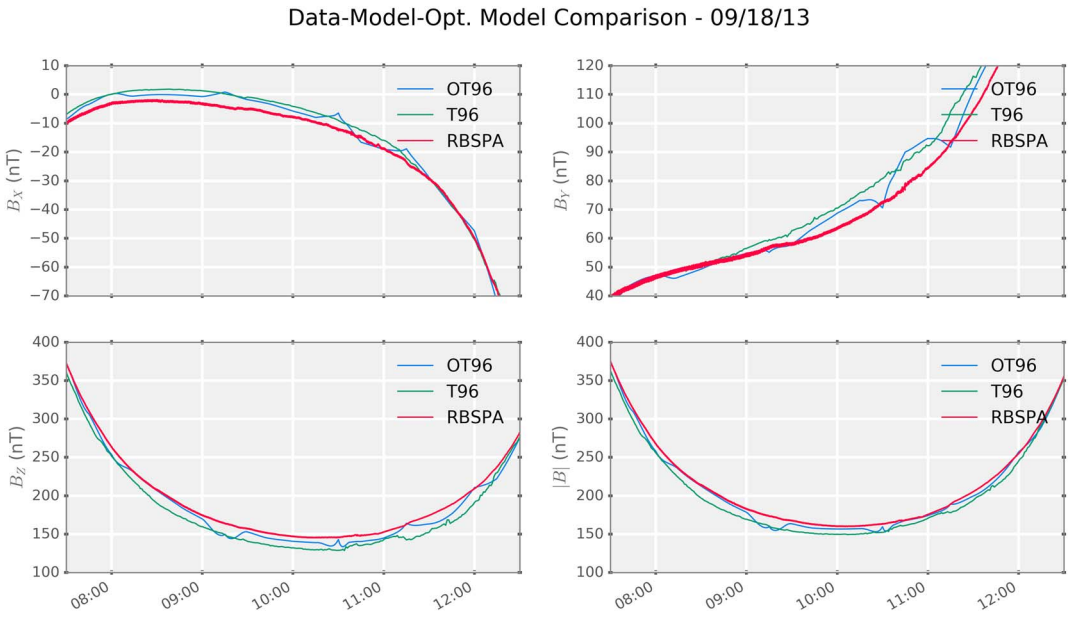

Figure 14. Comparison between observations and the optimized and unoptimized versions of the T96 model. A pass of apogee of the RBSP-A satellite on 18 September 2013 is used for comparison. T96 refers to the output from the model using the standard solar wind parameters and OT96 refers to the output using the optimized parameters.

that represents a differently stretched magnetotail in the optimized model. A more detailed study involving a point or region that could be mapped using other methods, such as the ion isotropic boundary (e.g., Sergeev et al., 1993), would be required to quantitatively assess whether the optimized models are more accurate than their original counterparts for mapping purposes.

As another example of the changes provided by the optimization process, we show a comparison of the magnetic field measured along the trajectory of one of the Van Allen Probes satellites and modeled with the original and optimized input parameters using the T96 model. Figure 14 shows that all the components of the optimized magnetic field (blue line) are almost always closer to the observed value than the original magnetic field (green line). However, the figure also shows that the optimized curve is less smooth than the original. The reason for this behavior could be attributed to the fact that the satellite used in the plot for comparison (RBSP-A) could have been chosen to be on the fitting set during part of the time, due to the selection criteria, and could have a larger weight at times on the calculation of the optimized parameters, pulling the value of the magnetic field closer to the observation.

4. Discussion

The difference between all the models analyzed, when compared in terms of the median of the error distributions, is not very large. However, they are still relevant, and moreover, it is worth comparing the different field models available to the community to find out how their performances compare. It is important to note that there are significant differences when it comes to the tail end of the error distributions examined. This means that some models are more prone to producing large errors than others, depending on the parameters. The typical error is small in both optimized and unoptimized models. However, the unoptimized models can have very large errors and these are much reduced after the optimization procedures. The error distributions are all very skewed, so finding the difference in, for instance, the 95th percentile of the distribution or the number of outliers is sometimes more important than looking at the mean or the median of the distribution.

When comparing all the models before the optimization (Figures 3–6 and 8), the metric used to make the comparison makes a difference in the performance outcome. Figure 3 compares the models using the RMS error and shows that the T02 model performs clearly better than the others, while the TS04 and TS07 models come in second place and TS01 is slightly worse. This comparison changes a little if we look at Figure 8. Here T02 is still the best model but not by a large margin, and TS01 performs slightly better than TS04 and TS07. The same

analysis can be done when comparing the magnitude and angular parts of the magnetic field error separately. All models, except T96, have similar accuracy when comparing the angular error, with T02 again performing slightly better than the others (Figure 6). TS07 has a noticeable worse performance, out of the main four, in terms of the magnitude error (Figure 4). It is interesting to note that the most modern model—TS07—does not perform better than its recent predecessors in any metric during the specific period of analysis.

The present study is an improvement on previous studies where fitting with in situ data was used (Kubyshkina et al., 2008, 2009) (Model 2 and Model AM-01, respectively). These studies used the RMS error of the magnetic field to optimize the parameters of the T01S and T96 models, respectively. Although Kubyshkina et al. (2008) refer to the T01S model throughout their paper, calling it a storm time model, they cite Tsyganenko (2002a), which corresponds to the T02 model. The present study differs from these previous studies in several ways besides the different metric: this study uses more satellites for the fitting than previously published event-fitted modeling; we apply the optimization procedure to a range of models instead of just one; we performed an out-of-sample validation test of the final results. Previous studies have not described an out-of-sample validation of the magnetic field vectors; results were only compared with results from other model versions. The independent validation performed by Kubyshkina et al. (2009), for example, compares the isotropy boundaries estimated from low-Earth orbiting satellites to isotropy boundary locations computed from the magnetic field model. We have restricted our study to the local magnetic field vectors, but other validation studies including comparing the isotropy boundary location (e.g., Kubyshkina et al., 2009; Sergeev et al., 1993) and performing phase space density matching (e.g., Morley et al., 2013) would provide useful tests of the global morphology and mapping accuracy of the optimized models.

It is also important to note that this study focuses on improving the accuracy of the magnetic field without any concern for spatial gradients and time derivatives. As seen in Figures 13 and 14, the time evolution of magnetic field foot points and magnetic field values along a satellite trajectory using the optimized parameters are more "choppy" when compared to the original parameters. Therefore, it is safe to assume that using optimized parameters will probably lead to greater and more variable time derivatives in general and derived quantities that use gradients or time derivatives may be less accurate using this optimization scheme. For future development of this method we aim to investigate methods for preserving some "memory" in the field model. For example, the W and G parameters could be given initial values from the previous time step and given stricter bounds on how far away from their previous value the optimization could allow them to be.

While the optimization process does find a set of input parameters that minimize the differences between the model field and the measurements used for fitting, it is unlikely that the optimal driving parameters represent a realistic estimate of the observables that the models are parameterized by. However, most of the model parameters estimated during the optimization procedure show broad, qualitative similarity with the corresponding Qin-Denton parameter; one important difference is that the original Qin-Denton parameters are mostly autocorrelated, but the optimized parameters tend to show significant variability from time step to time step. Smoothing the optimized parameters generally gives qualitative agreement with the original Qin-Denton parameters, though magnitudes still differ. The biggest exception to this is IMF-B_Z: Smoothing the optimized IMF-B_Z parameter does not make it qualitatively similar to the observed B_Z for all of the models.

The metrics comprising the cost function (τ) presented in this work, used both to compare model performance and to optimize them, are also different from the metrics used by other studies. Kubyshkina et al. (2009) used the RMS error, while previous works do not describe the cost function used for the fitting. Using the RMS error has two significant disadvantages, first, it is not scale dependent, as discussed previously, and thus, it will penalize similar absolute errors equally regardless of the observation value. It is clear that a 5 nT error, for instance, should be penalized differently if it is referring to a 100 nT observation or a 10 nT observation. Second, the RMS error is biased toward the B_Z component of the magnetic field, since most of the observations are taken at or near the equatorial plane where the value of the B_Z component dominates. For instance, if a measurement made at geosynchronous orbit near the equatorial plane has a 5% error in all three components, the error in the B_Z component will contribute much more to the final RMS error, since at this location the B_Z component is of the order of 100 times greater than the other two. These two problems are mitigated through use of the τ metric. However, τ has the disadvantage of not being an interpretable one, although it is easy to convert each of its two terms into something that is interpretable.

Using the τ metric has another advantage over using the RMS error. τ has two separate terms corresponding to the magnitude contribution and the angular contribution to the final value of the metric. Having two separate

terms is beneficial because it allows for the adoption of a greater weight in one of the two terms, if desired. Depending on the application, it could be better to give one of the two parts a greater weight in the calculation of the best fit. In addition to that this separation allows us to use data that otherwise would not be possible, such as the use of angular orientation of the magnetic field based on particle data done with the Los Alamos National Laboratory geosynchronous constellation (Chen et al., 2016; Thomsen et al., 1996).

Using the full 21 day period described here, all models presented an improvement in accuracy over their unoptimized counterparts using the 90th percentile of the error distribution, which gives information on the outliers, as a metric of comparison. The restriction imposed by using at least six satellites for the fitting process proved to be relevant to get even better results, which are more clearly seen when comparing the results for the T96 (four parameters) model. It imposes a greater burden on the optimization process to do it with fewer in situ measurements than parameters to be optimized. Therefore, we can infer that having more satellites will usually lead to better optimization results.

The final results when using at least six satellites for the fitting proved to be relevant. Sometimes, the improvement from the optimization procedure is more pronounced in the magnitude term, other times in the angular term. The improvement is not always clear in the median of the error distribution; with the exception of τ for T01S, all models showed a significant reduction in the 75th and 90th percentiles of the same distribution whether the comparison with the original models was done with τ, or the magnitude and angular parts separately. This improvement corresponds to a reduction in the number of outliers in the distribution. Some models are prone to large errors in some conditions, and application of this optimization method mitigates this problem.

Overall, the best model after the optimization was the TS04 model. Comparing all four optimized models when using six or more satellites for optimization, TS04 has the best performance both in terms of the median and in terms of the 90th percentile. It also has the best performance when separately comparing the magnitude and the angular term of the τ function. As the TS04 model has ten parameters and we only fit with data from up to nine satellites, we note that the optimized model could be underdetermined. However, the testing of the optimization procedure suggested that the fitting was robust, even with fewer data points than parameters. One possible explanation is that the input parameters are (to varying degrees) correlated with one another; thus, the number of effective free parameters is lower than the actual number of model parameters. While use of six or more satellites gave a significant improvement, we believe that including additional data would further improve the results from an optimized TS04 model.

For future applications of this optimization method, some remarks have to be made. In the results presented here all the errors were calculated on the validation satellite at each time step, which was not used for fitting. Depending on the application, the user may want to use all satellites available for fitting and even give one specific satellite a larger weight on the cost function. This method permits giving different weights for different points or regions of the magnetosphere besides giving different weights for the two different terms (magnitude and angular). For instance, if a user wants to find the magnetic foot point for a given satellite location, one of the optimized models can be used for that where the user can use all in situ measurements available at a particular time to do the fitting including the satellite at the location for which the foot point is being calculated. In this case the user may want to more heavily weight the angular term in the cost function.

5. Conclusions

We presented a method for comparing the accuracy of empirical magnetic field models as well as optimizing the empirical magnetic field model using in situ magnetic field measurements. The optimized model builds on the approach of Kubyshkina et al. (2009) and uses a direct search optimization algorithm to minimize a cost function—τ (see equation (9))—that explicitly includes both the relative error in predicted magnetic field magnitude and the angular error. By reserving one observation per time step for validation, we performed an out-of-sample validation for the fitted magnetic field models.

Four different periods during the year of 2013, totaling 21 days, were used in this study. These periods were 16–20 March, 1–5 June, 17–21 September, and 22–27 September. These periods were chosen because they cover the Geospace Environment Modeling (GEM) Challenge events selected by the Quantitative Assessment of Radiation Belt Modeling (QARBM) Focus Group. The selected intervals contained different levels of activity

covering both quiet time and storm time. As the results from this model evaluation were based on points in the magnetosphere that were not used for fitting, the results provide an independent validation of the method.

Comparison between five magnetic field models (T96, T01S, T02, TS04, and TS07) widely used by the community provided interesting results. Specifically, the T02 model was demonstrated to be on average the most accurate model out of the five when driven by the standard model input parameters — we note that this result is for the unoptimized models. This result is a little surprising as more recent models such as TS04 and TS07 might be expected to perform better in general than older models. The performance of the earlier (T96, T02, and T01S) models are consistent with the statistical comparison of McCollough et al. (2008), who concluded that T02 should be used in place of T96 for quiet to moderate conditions ($Kp < 6$) and that T01S "is a much more suitable choice during storm time conditions ($Kp > 6$), since it can handle extreme solar wind inputs." The quantitative assessment performed by Huang et al. (2008) concluded that TS04 had improved performance over its predecessors across a range of storm time activity levels. It is worth noting that both of these studies were limited to geosynchronous orbit, and so we might not expect the results to be generally applicable. Indeed, Morley et al. (2013) found that T96 better reproduced the magnetic field magnitude during the storm recovery phase that they studied, even though TS04 gave better results for phase space density matching.

We further presented results of fitting four of these magnetic field models (T96, T01S, T02, and TS04) to in situ data with the goal of improving the accuracy of the models. This optimization procedure improved the accuracy of all of the four models when using at least six satellites to provide data for fitting (see Figure 10 and Table 4). Using fewer than six satellites for the fitting did not provide a clear improvement in all metrics for the T01S and T02 models. While the median of the error distribution did not change significantly in some cases, this approach did significantly reduce the number of outliers in the error distribution for all models. That is, even when the typical error was not reduced, the number of large errors was substantially improved.

After applying the fitting procedure (and using at least six satellites in the fitting), the optimized TS04 model was verified to be the most accurate in terms of both magnitude and direction. Specifically, the optimized TS04 model gave a smaller error than its unoptimized counterpart 57.3% of the time, and the median percentage error in $|B|$ was reduced from 4.54% to 3.84%. The optimized TS04 model also had the smallest angular error, where the median angular error was 2.66° and 90% of all angular errors were smaller than 14.9°. It was also better than the best unoptimized model (T02) 56.2% of the time. More tests are required to further characterize the model improvements presented in this study. Additional forms of testing the optimized models are also important as a next step, such as comparing isotropy boundaries and performing phase space density matching.

As our methodology uses an optimized parameter approach, the results of this study include a set of optimized parameters that can be used to evaluate the models studied in this paper. These optimized parameters are included with this paper as supporting information so that the broader scientific community can use the optimized magnetic field models immediately, and without any additional code development, using any standard implementation of the magnetic field models tested here.

Acknowledgments
This work was performed under the auspices of the U.S. Department of Energy and funded by the Laboratory Directed Research and Development (LDRD) program award 20150127ER. T. V. B. wishes to thank Richard Denton for discussions about this paper. Analysis and plotting used the publicly available LANLGeoMag and SpacePy libraries. SpacePy is available from http://sourceforge.net/p/spacepy and LANLGeoMag is available from https://github.com/drsteve/LANLGeoMag. Magnetometer data from Cluster, GOES, THEMIS, and Van Allen Probes were obtained from the NASA CDAWeb website (https://cdaweb.sci.gsfc.nasa.gov). Orbital elements for all satellites were obtained from http://www.Space-Track.org.

References

Angelopoulos, V. (2008). The THEMIS mission. *Space Science Reviews*, *141*(1), 5. https://doi.org/10.1007/ s11214-008-9336-1

Antonova, E. E., Vorobjev, V. G., Kirpichev, I. P., Yagodkina, O. I., & Stepanova, M. V. (2015). Problems with mapping the auroral oval and magnetospheric substorms. *Earth, Planets and Space*, *67*(1), 166. https://doi.org/10.1186/s40623-015-0336-6

Apatenkov, S. V., Sergeev, V. A., Kubyshkina, M. V., Nakamura, R., Baumjohann, W., Runov, A., … Khotyaintsev, Y. (2007). Multi-spacecraft observation of plasma dipolarization/injection in the inner magnetosphere. *Annales Geophysicae*, *25*(3), 801–814. https://doi.org/10.5194/angeo-25-801-2007

Auster, H. U., Glassmeier, K. H., Magnes, W., Aydogar, O., Baumjohann, W., Constantinescu, D., … Wiedemann, M. (2009). The THEMIS fluxgate magnetometer. In J. L. Burch & V. Angelopoulos (Eds.) *The THEMIS mission* (pp. 235–264). New York: Springer. https://doi.org/10.1007/978-0-387-89820-9_11

Balogh, A., Carr, C. M., Acuña, M. H., Dunlop, M. W., Beek, T. J., Brown, P., … Schwingenschuh, K. (2001). The Cluster magnetic field investigation: Overview of in-flight performance and initial results. *Annales Geophysicae*, *19*(10/12), 1207–1217. https://doi.org/10.5194/angeo-19-1207-2001

Chen, Y., Cunningham, G., & Henderson, M. (2016). Determination of errors in derived magnetic field directions in geosynchronous orbit: Results from a statistical approach. *Annales Geophysicae*, *34*(9), 831–843. https://doi.org/10.5194/angeo-34-831-2016

Desorgher, L., Kudela, K., Fluckiger, E. O., Butikofer, R., Storini, M., & Kalegaev, V. (2009). Comparison of Earth's magnetospheric magnetic field models in the context of cosmic ray physics. *Acta Geophysica*, *57*, 75–87. https://doi.org/10.2478/s11600-008-0065-3

Escoubet, C. P., Fehringer, M., & Goldstein, M. (2001). Introduction the cluster mission. *Annales Geophysicae*, *19*(10/12), 1197–1200. https://doi.org/10.5194/angeo-19-1197-2001

Escoubet, C. P., Masson, A., Laakso, H., & Goldstein, M. L. (2015). Recent highlights from Cluster, the first 3-D magnetospheric mission. *Annales Geophysicae*, *33*(10), 1221–1235. https://doi.org/10.5194/angeo-33-1221-2015

Friedel, R. H. W., Bourdarie, S., & Cayton, T. E. (2005). Intercalibration of magnetospheric energetic electron data. *Space Weather*, *3*, S09B04. https://doi.org/10.1029/2005SW000153

Ganushkina, N. Y., Pulkkinen, T. I., Kubyshkina, M. V., Singer, H. J., & Russell, C. T. (2002). Modeling the ring current magnetic field during storms. *Journal of Geophysical Research*, *107*(A7), SMP 3–1–SMP 3-13. https://doi.org/10.1029/2001JA900101

Ge, Y. S., Zhou, X.-Z., Liang, J., Raeder, J., Gilson, M. L., Donovan, E., … Runov, A. (2012). Dipolarization fronts and associated auroral activities: 1. Conjugate observations and perspectives from global MHD simulations. *Journal of Geophysical Research*, *117*, A10226. https://doi.org/10.1029/2012JA017676

Han, L., & Neumann, M. (2006). Effect of dimensionality on the Nelder-Mead simplex method. *Optimization Methods and Software*, *21*(1), 1–16. https://doi.org/10.1080/10556780512331318290

Hones, E. W., Thomsen, M. F., Reeves, G. D., Weiss, L. A., McComas, D. J., & Newell, P. T. (1996). Observational determination of magnetic connectivity of the geosynchronous region of the magnetosphere to the auroral oval. *Journal of Geophysical Research*, *101*(A2), 2629–2640. https://doi.org/10.1029/95JA00418

Huang, C.-L., Spence, H. E., Singer, H. J., & Tsyganenko, N. A. (2008). A quantitative assessment of empirical magnetic field models at geosynchronous orbit during magnetic storms. *Journal of Geophysical Research*, *113*, A04208. https://doi.org/10.1029/2007JA012623

Iles, R. H. A., Meredith, N. P., Fazakerley, A. N., & Horne, R. B. (2006). Phase space density analysis of the outer radiation belt energetic electron dynamics. *Journal of Geophysical Research*, *111*, A03204. https://doi.org/10.1029/2005JA011206

Kiusalaas, J. (2010). *Numerical methods in engineering with Python* (2nd edn.). New York: Cambridge University Press.

Kletzing, C. A., Kurth, W. S., Acuna, M., MacDowall, R. J., Torbert, R. B., Averkamp, T., … Tyler, J. (2013). The electric and magnetic field instrument suite and integrated science (EMFISIS) on RBSP. *Space Science Reviews*, *179*, 127–181. https://doi.org/10.1007/s11214-013-9993-6

Kress, B. T., Mertens, C. J., & Wiltberger, M. (2010). Solar energetic particle cutoff variations during the 29–31 October 2003 geomagnetic storm. *Space Weather*, *8*, S05001. https://doi.org/10.1029/2009SW000488

Kubyshkina, M., Pulkkinen, T. I., Ganushkina, N. Y., & Partamies, N. (2008). Magnetospheric currents during sawtooth events: Event-oriented magnetic field model analysis. *Journal of Geophysical Research*, *113*, A08211. https://doi.org/10.1029/2007JA012983

Kubyshkina, M., Sergeev, V., Tsyganenko, N., Angelopoulos, V., Runov, A., Donovan, E., … Baumjohann, W. (2011). Time-dependent magnetospheric configuration and breakup mapping during a substorm. *Journal of Geophysical Research*, *116*, A00I27. https://doi.org/10.1029/2010JA015882

Kubyshkina, M., Sergeev, V., Tsyganenko, N., Angelopoulos, V., Runov, A., Singer, H., … Baumjohann, W. (2009). Toward adapted time-dependent magnetospheric models: A simple approach based on tuning the standard model. *Journal of Geophysical Research*, *114*, A00C21. https://doi.org/10.1029/2008JA013547

Lewis, R. M., Torczon, V., & Trosset, M. W. (2000). Direct search methods: Then and now. *Journal of Computational and Applied Mathematics*, *124*(1), 191–207. https://doi.org/10.1016/S0377-0427(00)00423-4

Mauk, B. H., Fox, N. J., Kanekal, S. G., Kessel, R. L., Sibeck, D. G., & Ukhorskiy, A. (2013). Science objectives and rationale for the radiation belt storm probes mission. *Space Science Reviews*, *179*, 3–27. https://doi.org/10.1007/s11214-012-9908-y

McCollough, J., Gannon, J., Baker, D., & Gehmeyr, M. (2008). A statistical comparison of commonly used external magnetic field models. *Space Weather*, *6*, S10001. https://doi.org/10.1029/2008SW000391

Morley, S. K. (2016). Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications (Tech. Rep. LA-UR-16-24592). Los Alamos, NM: Los Alamos National Laboratory. https://doi.org/10.2172/1260362

Morley, S. K., Henderson, M. G., Reeves, G. D., Friedel, R. H. W., & Baker, D. N. (2013). Phase space density matching of relativistic electrons using the Van Allen Probes: REPT results. *Geophysical Research Letters*, *40*, 4798–4802. https://doi.org/10.1002/grl.50909

Murphy, K. R., Mann, I. R., Rae, I. J., Walsh, A. P., & Frey, H. U. (2014). Inner magnetospheric onset preceding reconnection and tail dynamics during substorms: Can substorms initiate in two different regions? *Journal of Geophysical Research: Space Physics*, *119*, 9684–9701. https://doi.org/10.1002/2014JA019795

Nelder, J., & Mead, R. (1965). A simplex-method for function minimization. *Computer Journal*, *7*(4), 308–313.

Olsson, D. M., & Nelson, L. S. (1975). The Nelder-Mead simplex procedure for function minimization. *Technometrics*, *17*(1), 45–51.

Pulkkinen, T. I., Baker, D. N., Fairfield, D. H., Pellinen, R. J., Murphree, J. S., Elphinstone, R. D., … Nagai, T. (1991). Modeling the growth phase of a substorm using the Tsyganenko model and multi-spacecraft observations: CDAW-9. *Geophysical Research Letters*, *18*, 1963–1966. https://doi.org/10.1029/91GL02002

Pulkkinen, T. I., Ganushkina, N. Y., Tanskanen, E. I., Kubyshkina, M., Reeves, G. D., Thomsen, M. F., … Gjerloev, J. (2006). Magnetospheric current systems during stormtime sawtooth events. *Journal of Geophysical Research*, *111*, A11S17. https://doi.org/10.1029/2006JA011627

Pulkkinen, T. I., & Tsyganenko, N. A. (1996). Testing the accuracy of magnetospheric model field line mapping. *Journal of Geophysical Research*, *101*(A12), 27,431–27,442. https://doi.org/10.1029/96JA02489

Qin, Z., Denton, R. E., Tsyganenko, N. A., & Wolf, S. (2007). Solar wind parameters for magnetospheric magnetic field modeling. *Space Weather*, *5*, S11003. https://doi.org/10.1029/2006SW000296

Roederer, J. G., & Zhang, H. (2014). Dynamics of magnetically trapped particles. In *Foundations of the Physics of Radiation Belts and Space Plasmas* (2nd edn., Vol. 10, pp. 978–3). Berlin: Springer-Verlag.

Sergeev, V. A., Malkov, M., & Mursula, K. (1993). Testing the isotropic boundary algorithm method to evaluate the magnetic field configuration in the tail. *Journal of Geophysical Research*, *98*(A5), 7609–7620. https://doi.org/10.1029/92JA02587

Sergeev, V., Semenov, V., Kubyshkina, M., Ivanova, V., Baumjohann, W., Nakamura, R., … Lucek, E. (2007). Observation of repeated intense near-Earth reconnection on closed field lines with Cluster, Double Star, and other spacecraft. *Geophysical Research Letters*, *34*, L02103. https://doi.org/10.1029/2006GL028452

Sergeev, V. A., Tanskanen, P., Mursula, K., Korth, A., & Elphic, R. C. (1990). Current sheet thickness in the near-Earth plasma sheet during substorm growth phase. *Journal of Geophysical Research*, *95*(A4), 3819–3828. https://doi.org/10.1029/JA095iA04p03819

Singer, H., Matheson, L., Grubb, R., Newman, A., & Bouwer, D. (1996). Monitoring space weather with the GOES magnetometers. In E. R. Washwell (Ed.), *GOES-8 and Beyond, Proceedings of the SPIE* (Vol. 2812, pp. 299–308). https://doi.org/10.1117/12.254077

Thomsen, M. F., McComas, D. J., Reeves, G. D., & Weiss, L. A. (1996). An observational test of the Tsyganenko (T89a) model of the magnetospheric field. *Journal of Geophysical Research*, *101*(A11), 24,827–24,836. https://doi.org/10.1029/96JA02318

Tsyganenko, N. A. (1987). Global quantitative models of the geomagnetic field in the cislunar magnetosphere for different disturbance levels. *Planetary and Space Science*, *35*(11), 1347–1358.

Tsyganenko, N. A. (1989). A magnetospheric magnetic field model with a warped tail current sheet. *Planetary and Space Science*, *37*, 5.

Tsyganenko, N. A. (1995). Modeling the Earth's magnetospheric magnetic field confined within a realistic magnetopause. *Journal of Geophysical Research*, *100*(A4), 5599–5612. https://doi.org/10.1029/94JA03193

Tsyganenko, N. A. (2002a). A model of the near magnetosphere with a dawn-dusk asymmetry 1. Mathematical structure. *Journal of Geophysical Research*, *107*(A8), SMP 12–1–SMP 12-15. https://doi.org/10.1029/2001JA000219

Tsyganenko, N. A. (2002b). A model of the near magnetosphere with a dawn-dusk asymmetry 2. Parameterization and fitting to observations. *Journal of Geophysical Research*, *107*(A8), SMP 10–1–SMP 10-17. https://doi.org/10.1029/ 2001JA000220

Tsyganenko, N. A. (2013). Data-based modelling of the Earth's dynamic magnetosphere: A review. *Annales Geophysicae*, *31*(10), 1745–1772. https://doi.org/10.5194/angeo-31-1745-2013

Tsyganenko, N. A., Singer, H. J., & Kasper, J. C. (2003). Storm-time distortion of the inner magnetosphere: How severe can it get? *Journal of Geophysical Research*, *108*(A5), 1209. https://doi.org/10.1029/2002JA009808

Tsyganenko, N. A., & Sitnov, M. I. (2005). Modeling the dynamics of the inner magnetosphere during strong geomagnetic storms. *Journal of Geophysical Research*, *110*, A03208. https://doi.org/10.1029/2004JA010798

Tsyganenko, N. A., & Sitnov, M. I. (2007). Magnetospheric configurations from a high-resolution data-based magnetic field model. *Journal of Geophysical Research*, *112*, A06225. https://doi.org/10.1029/2007JA012260

Tsyganenko, N. A., & Stern, D. P. (1996). Modeling the global magnetic field of the large-scale Birkeland current systems. *Journal of Geophysical Research*, *101*(A12), 27187.

Turner, D. L., Angelopoulos, V., Morley, S. K., Henderson, M. G., Reeves, G. D., Li, W., … Rodriguez, J. V. (2014). On the cause and extent of outer radiation belt losses during the 30 September 2012 dropout event. *Journal of Geophysical Research: Space Physics*, *119*, 1530–1540. https://doi.org/10.1002/2013JA019446

Woodfield, E. E., Dunlop, M. W., Holme, R., Davies, J. A., & Hapgood, M. A. (2007). A comparison of Cluster magnetic data with the Tsyganenko 2001 model. *Journal of Geophysical Research*, *112*, A06248. https://doi.org/10.1029/2006JA012217

Yu, Y., Koller, J., Jordanova, V. K., Zaharia, S. G., Friedel, R. W., Morley, S. K., … Spence, H. E. (2014). Application and testing of the L^* neural network with the self-consistent magnetic field model of RAM-SCB. *Journal of Geophysical Research: Space Physics*, *119*, 1683–1692. https://doi.org/10.1002/2013JA019350

Zaharia, S., Jordanova, V. K., Thomsen, M. F., & Reeves, G. D. (2006). Self-consistent modeling of magnetic fields and plasmas in the inner magnetosphere: Application to a geomagnetic storm. *Journal of Geophysical Research*, *111*, A11S14. https://doi.org/10.1029/2006JA011619

Zhang, Q.-H., Dunlop, M. W., Holme, R., & Woodfield, E. E. (2010). Comparison of eight years magnetic field data from Cluster with Tsyganenko models in the inner magnetosphere. *Annales Geophysicae*, *28*(1), 309–326. https://doi.org/10.5194/angeo-28-309-2010