Real-Time Validation of the Dst Predictor Model*

James P. McCollough and Shawn L. Young

Space Weather Center of Excellence
Space Vehicles Directorate
Air Force Research Laboratory

E. Joshua Rigler and Hal A. Simpson

Geomagnetism Program
U.S. Geological Survey

February 25, 2015

*This document partially fulfills the requirement for Task 2 in the FY2012 SWAFS SPO Statement of Work for the Air Force Research Laboratory

Table of Contents

List of Figures

List of Tables

1 Summary

The Dst Predictor model, which has been running real-time in the Space Weather Analysis and Forecast System (SWAFS), provides 1-hour and 4-hour forecasts of the Dst index. This is useful for awareness of impending geomagnetic activity, as well as driving other real-time models that use Dst as an input. In this report, we examine the performance of this forecast model in detail. When validating indices it should be noted that performance is only with respect to a reference index as they are derived quantities assumed to reflect a state of the magnetosphere that cannot be directly measured. In this case U.S. Geological Survey (USGS) Definitive Dst is the reference index (Section 3). Whether or not the model better reflects the actual activity level is nearly impossible to discern and outside the scope of this report.

We evaluate the performance of the model by computing continuous predictant skill scores against USGS Definitive Dst values as "observations" (Section 4.2). The two sets of data are not well-correlated for both 1-hour and 4-hour forecasts. The Dst Predictor Prediction Efficiency for both the 1- and 4-hour forecasts suggests poor performance versus the climatological mean. However, the skill score against a nowcast persistence model is positive, suggesting value added by the Dst Predictor model. We further examine statistics for storm times (Section 4.3) with similar results: nowcast persistence performs worse than Dst Predictor.

Dst Predictor is superior to the nowcast persistence model for the metric used in this study. We recommend continued use of the Dst Predictor model for 1- and 4-hour Dst predictions along with active study of other Dst forecast models that do not rely on nowcast inputs (Section 6).

The lack of certified requirements makes further recommendations difficult. A study of how the error in Dst translates to error in models and a better understanding of operational needs for magnetic storm warning are needed to determine such requirements. Nowcast persistence is often hard to beat for short term forecasts and specification and Dst Predictor clearly performs well against that standard (with 1-hour and 4-hour skill-scores of 0.233 and 0.485 respectively), although poor in absolute terms (with 1-hour and 4-hour prediction efficiencies of -64.6 and -43.1, respectively).

2 Introduction

This document presents a statistical assessment of the Dst Predictor model as implemented in the Space Weather Analysis and Forecast System (SWAFS). We use several different approaches to determine how accurate the model is at specifying the U.S. Geological Survey (USGS) Definitive Dst value, a recognized authoritative version. This performance is determined in absolute terms and also relative to a nowcast persistence model.

The format of this report is as follows. We first introduce modeling approaches employed for geomagnetic indices, followed by details of the Dst Predictor model. Next, we discuss the data and methodology for assessing the performance of Dst Predictor. This is followed by a presentation of the statistical results. Finally we present discussion and concluding remarks.

2.1 Modeling Geomagnetic Indices

The physical states that geomagnetic indices are taken to represent cannot be directly measured. Indices are derived quantities that are assumed to reflect certain properties of the magnetosphere. Thus, when validating indices it should be noted that performance is only with respect to a reference index (in this case USGS Definitive Dst; see Section 3). Whether or not this model better reflects the state of the global magnetosphere would require "measuring" the state directly, which is nearly impossible, even conceptually.

In this report we examine a model for predicting Dst. To compute Dst, a local disturbance value is computed at each of 4 low-latitude magnetometer stations by subtracting long-term variation in the time-domain from the horizontal magnetic field intensity. Short-term quiet-time variations are then subtracted from the signal the frequency domain. These local disturbance values for each station are finally weighted by magnetic latitude and averaged to produce Dst [1].

There are two main approaches to modeling geomagnetic indices: empirical modeling and approximation. Empirical modeling attempts to find a data-based relation between the quantity of interest, i.e. 3-hour Kp or Dst, and other measurable quantities, such as solar wind parameters and geomagnetic fluctuations. In contrast, approximation utilizes a subset of the raw data that contributes to the index in question to model its behavior. Dst Predictor is an empirical model, utilizing a neural network framework to forecast a Dst value for a given time period.

2.2 The Dst Predictor Model

As an empirical neural network (NN) model, Dst Predictor utilizes a mathematical model that mimics computation performed by neurons in the human brain [6]; it utilizes the same architecture as Kp Predictor (See [2]). Dst Predictor is a recurrent NN, where hidden layer outputs are fed back to model inputs. This is different than a feed-forward NN which only moves data forward through the network.

Dst Predictor includes two distinct NNs: a model that uses solar wind data and nowcast Dst values to produce 1-hour forecasts, and a model that takes the same input and produces 4-hour forecasts. In addition, a separate model produces 1-hour forecasts with only solar wind inputs for times when no nowcast is available or the nowcast is less than -200 nT. Dst Predictor produces the outputs from the NNs with nowcast inputs when nowcast values are available; otherwise, the 1-hour forecast is from the third model and no 4-hour output is produced. Dst Predictor was trained against the Dst index produced by the Kyoto World Data Center for Geomagnetism (WDC) from 1975 to 2001, with definitive data where possible and preliminary data elsewhere.

AFRL vs. SWAFS Instances

AFRL does not have the capability to stand-up and maintain an instance of the SWAFS system. The Dst Predictor model used in this study was received from the developer in the form it was delivered to SWAFS, including a critical post-delivery patch. In order to perform a validation against a solar cycle of historical data, further modification was performed by AFRL to specify the times for which to produce forecasts. This modification was as noninvasive and minimal as could

be accomplished, and mirrored modifications performed for the Kp Predictor validation [3]. A copy of the modified code as executed for validation is included in the report package. The AFRL instance was executed on a CentOS Linux workstation (hostname: `swfl_alpha`) designated as a model testbed and validation environment.

3 Methods, Assumptions, and Procedures

We have implemented a script to run Dst Predictor from 1 Jul 2003 to 30 Dec 2012. It executes a loop which updates the `nn_ctime.dat` file with the date and time of interest, runs the model, and archives the output.

We use real-time 1-minute Dst data (RT Dst) from the USGS Geomagnetism Program [1] as the nowcast model values. The RT Dst product has only been in production since 2011, so there is no archive of RT Dst data from which to draw that would support this study's goals. However, the USGS does archive so-called "preliminary" magnetic observatory data for their Honolulu and San Juan observatories, as do Kakioka and Hermanus through INTERMAGNET. These preliminary data were then fed into the USGS RT Dst algorithm, without modification, in a manner consistent with the USGS daily operations to generate a data set that resembled as closely as possible what would have been generated and archived had the USGS RT Dst algorithm been in production during the entire interval of interest. These magnetometer data are effectively uncalibrated, so the absolute baselines differ, sometimes substantially, from the absolute baselines used for definitive measurements.

For observational data, a modified version of the 1-minute Dst algorithm described by *Gannon et al.* [1] was used. These modification include: (1) definitive magnetic observatory data obtained only for the interval considered in this study were used; (2) secular variation was treated as a linear trend; (3) quasi-periodic solar cycle variations were not removed, since only $\sim$1 solar cycle of data was analyzed; and (4) geomagnetic storm intervals were identified according to a user-defined threshold before being interpolation prior to SQ calculation. This does not change the output substantially from those of *Gannon et al.* [1], and are based on a more justified approach. For the remainder of this discussion, we refer to these results as "Definitive Dst." We did not use the nowcast or definitive Dst provided by the Kyoto WDC due to a lack of data for a suitable time period for this study. In addition, foreign data sources for driving operational models by AFWA are discouraged for national security reasons.

3.1 Data Handling

Both RT and Definitive Dst have a 1-minute cadence and the Dst Predictor executes every 15 minutes. Thus, the nearest RT Dst value at each Dst Predictor run time is used as the "nowcast" input and the nearest Definitive Dst value at each output time becomes the "ground truth" for comparison purposes.

3.2 Statistical Metrics

There are many approaches to assess the performance of predictive models (see Chapter 8 of [5]). Given the continuous nature of Dst, we examine the performance of the model by computing continuous predictant skill scores. We will do this for all data as well as for when the RT Dst is below -50 to examine performance during storm-time [4]. For all the computations, times for which there is no value for either the model or the Definitive Dst were omitted from this study, since no comparison can be made between them. This does not bias the results since these times are rare and independent of activity level.

For an examination of how well the Definitive value is predicted, we produce the following quantities: correlation coefficient CC (or Pearson's r), prediction efficiency PE, and the skill score against a reference model, denoted by $\mathsf{SS}_{\mathsf{ref}}$. The correlation coefficient is computed in the standard manner, as:

$$\mathsf{CC} = \frac{\sum \left(x_i - \bar{x}\right)\left(\hat{x}_i - \bar{\hat{x}}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}\sqrt{\sum \left(\hat{x}_i - \bar{\hat{x}}\right)^2}}, \tag{1}$$

where x_i represents the forecast values, $\hat{x}_i$ the Definitive values, $\bar{x}$ the mean value of x_i, and the sums are over all data points i. The CC expresses how well correlated the model is with the observed values, but it does not describe how accurate the model is. The standard measure of accuracy is the mean squared error MSE:

$$\mathsf{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2, \tag{2}$$

and from this we can determine a skill score,

$$\mathsf{SS}_{\mathsf{ref}} = 1 - \frac{\mathsf{MSE}}{\mathsf{MSE}_{\mathsf{ref}}}, \tag{3}$$

where the "ref" subscript indicates a reference model. PE is the skill score where the reference model is the climatological mean, $\bar{\hat{x}}$:

$$\mathsf{PE} = 1 - \frac{\mathsf{MSE}}{\sigma^2}, \tag{4}$$

with σ^2 denoting the variance of the observed data. The $\mathsf{SS}_{\mathsf{pers}}$ quantity similarly uses a nowcast persistence reference model described by:

$$x_i = x^N_{i-\ell}, \tag{5}$$

where x^N_j is the nowcast value, i is the index of time for which the forecast is made, ℓ is the forecast time index (corresponding to 1 or 4 hours) and $i - \ell$ the index of time when the forecast is made. For this study, persistence utilizes the RT Dst. Skill scores have a range of $(-\infty, 1]$, where a score of 1 represents perfect model performance that forecasts the observed values

Table 1: Dst Predictor Performance: 01 Jul 2003–30 Dec 2012

Forecast	CC	PE	SS_{pers}	PE_{pers}
1-hour	0.101	-64.6	0.233	-83.2
4-hour	0.159	-43.1	0.485	-83.4

flawlessly, a score of 0 means the model is as accurate as the reference model, and a negative score means the model is less accurate than the reference model.

4 Results and Discussion

Here we present the results for continuous predictant performance measures. We include persistence model results for context and comparison purposes. An extended discussion of the nowcast inputs is included to aid interpretation of performance measures.

4.1 Absence of Certified Requirements

Certified requirements were not available at the time of this report. In order to generate these, a study of how the error in Dst translates to error in models to understand model sensitivity to this quantity is required. In addition, a better understanding of operational needs for magnetic storm forecasting is needed to understand operational sensitivity of this quantity.

4.2 General Performance

We utilize the continuous predictant quantities and persistence as a baseline model to provide context for the performance of Dst Predictor. These data were calculated for the time period of 01 Jul 2003–30 Dec 2012.

Table 1 shows the CC, PE, SS_{pers} and PE_{pers} for Dst Predictor. The data are not well correlated, and the Dst Predictor PE suggests a majority of the variance in the Definitive Dst is not captured by the forecasts. The positive skill scores against the nowcast persistence model suggest significant added value by the Dst Predictor model, while the very negative PE values of the persistence model suggest that the RT Dst is often very different than the corresponding Definitive value. We suspect the latter dominates Dst Predictor performance. It is thus straightforward to conclude that using a nowcast persistence model (Equation 5) is not preferred over the Dst Predictor in predicting the correct Definitive Dst value. The lackluster performance of the nowcast persistence model will be discussed in Section 4.4.

4.3 Storm-time Performance

We now examine the continuous predictant quantities for storming periods, defined as Dst less than -50 nT [4]. Performance for this subset of data is of particular importance for operators. Table 2 shows the CC, PE, SS_{pers} and PE_{pers} for the Dst Predictor during storms, where the

Table 2: Dst Predictor Performance: Storm time (USGS RT Dst < -50 nT)

Forecast	CC	PE	SS_{pers}	PE_{pers}
1-hour	0.007	-127	0.287	-175
4-hour	0.241	-59.8	0.650	-189

storm determinant is the RT Dst (as that is what an operator would use for decision-making). The 1-hour data are less well-correlated and the 4-hour better, but the conclusions are similar to Section 4.2. The Dst Predictor PE suggests a majority of the variance in the definitive Dst is missed by the forecasts. The positive skill scores against the nowcast persistence model suggest significant added value by the Dst Predictor model, particularly for 4-hour values. This is also suggested by the lower PE for persistence.

4.4 USGS Nowcast Issues

The results presented above are startlingly poor, and a significant factor for this model is the quality of the nowcast input (i.e., the USGS RT Dst). As can be surmised from the poor performance of the nowcast persistence, the RT Dst used as nowcast has significant discrepancies with the Definitive Dst. Figures 1 and 2 show how there are times where the RT Dst is significantly offset from the Definitive, particularly during the early part of the study period.

4.4.1 USGS RT Dst Algorithm

A high-level description of the USGS RT Dst algorithm will help to understand why nowcasts can be so poor: (1) the most recent year's worth of magnetic observatory data is ingested; (2) a simple average is subtracted from the time series; (3) a Fourier transform is performed on this detrended series; (4) the power at specific frequency bands is set to zero; and (5) the resulting spectrum is inverse Fourier transformed to produce a band-stop filtered series with periodic SQ removed. Finally, this "disturbance" time series is estimated for each of the observatories used for Dst, normalized to 0 degrees magnetic latitude, and averaged to produce Dst.

If a semi-permanent artificial offset arises in the observatory time series, the baseline calculated for that observatory will be impacted, introducing an artificial offset to the detrended series, the band-stop filtered time series, and ultimately Dst. It takes a significant amount of time for the actual and calculated baseline to realign. Meanwhile, Dst will exhibit an offset with the same sign from the expected zero-mean that decays linearly back to zero. This is especially apparent in Figure 1, in years 2005 and 2006. Short-lived offsets that return to the original baseline (i.e., a spike) cause something similar, but the offset has an opposite sign from the spike and remains constant for the entire realignment interval. These would not be so bad, except that such spikes are usually non-physical, and can be several tens of thousands of nT, leading to significant and constant offsets despite the vast majority of non-spike observations included in the average. An example of this is seen in Figure 1, straddling years 2003 to 2004.

These observatory baseline offsets and spikes are always cleaned up in definitive data processing, which has historically been the primary focus of the USGS Geomagnetism Program.

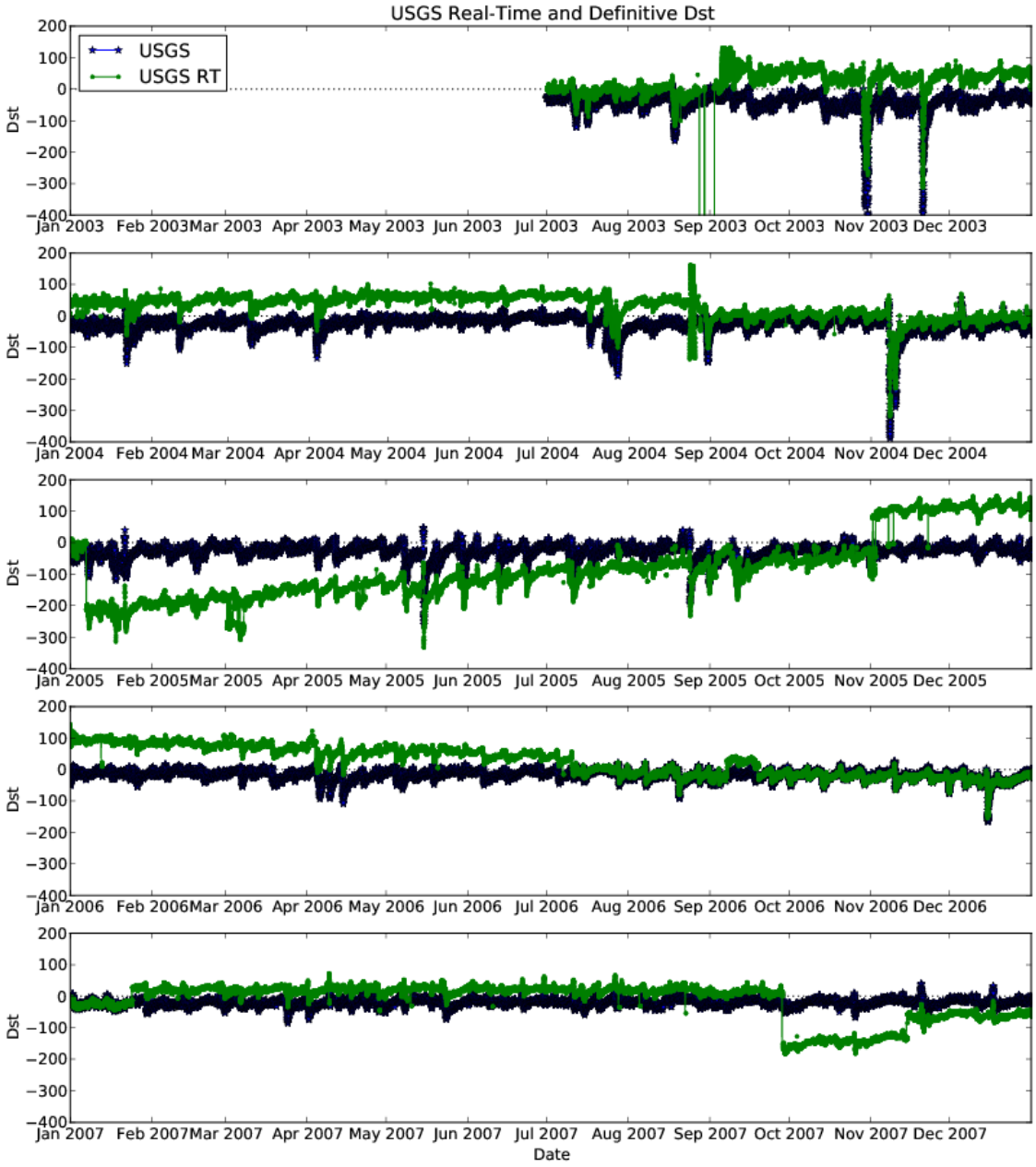

Figure 1: USGS Real-Time and Definitive Dst for the time period 2003–2007. While there are times with good agreement, there are frequently significant offsets in the RT data that will affect Dst Predictor performance.

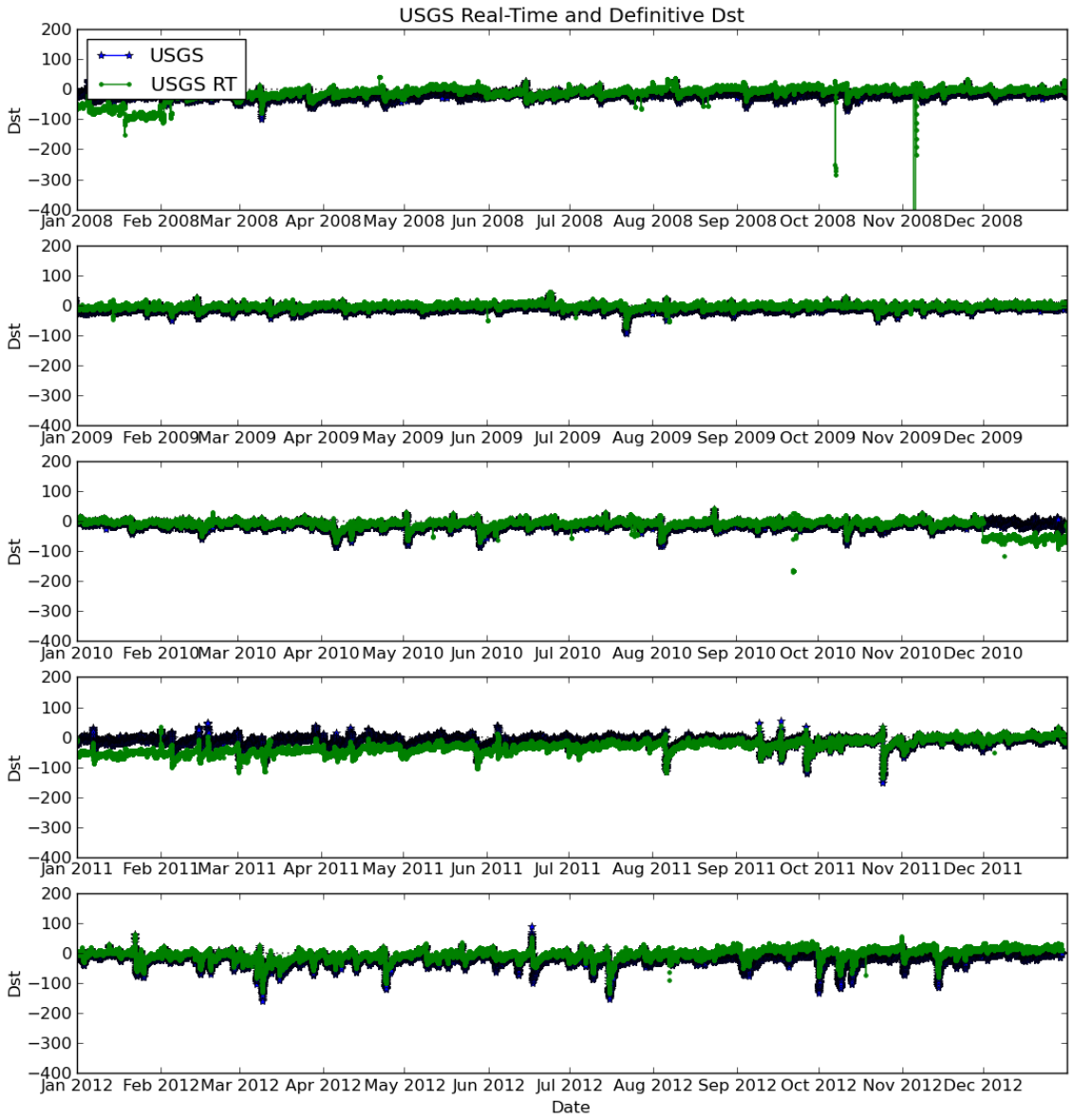

Figure 2: USGS Real-Time and Definitive Dst for the time period 2008–2012. While there are still offsets, the two datasets agree much more in this time period than that of Figure 1.

Table 3: Dst Predictor Performance: Figures 3 and 4

Forecast	CC	PE	SS_{pers}	PE_{pers}
1-hour	0.898	0.524	-0.261	0.622
4-hour	0.802	0.421	0.0665	0.381

However, a push has been made in recent years to improve the quality of USGS preliminary data for use in space weather applications, as evidenced by the reduced frequency and magnitude of RT Dst offsets in more recent years seen in Figure 2. However, a number of factors (including early results from this study) have motivated the USGS Geomagnetism Program to investigate alternative algorithms that are more suited for real-time data processing. Specifically, aims are to adjust calculated baselines more quickly to offsets and to more effectively filter spikes. Beta testing will precede transition to regular operations of such improvements.

4.4.2 Dst Predictor Behavior

Examining behavior during particular time periods is insightful. First, March 2012 is a time period where the RT Dst is in close agreement with the Definitive Dst (See Figures 3 and 4). In this case Dst Predictor also follows the Definitive Dst, with postive PEs for both the 1-hour and 4-hour forecasts (See Table 3). Note, however, that while the 4-hour forecast performs slightly better than nowcast persistence, the 1-hour forecast does not. This is not surprising given the fact that the nowcast is a driver for Dst Predictor.

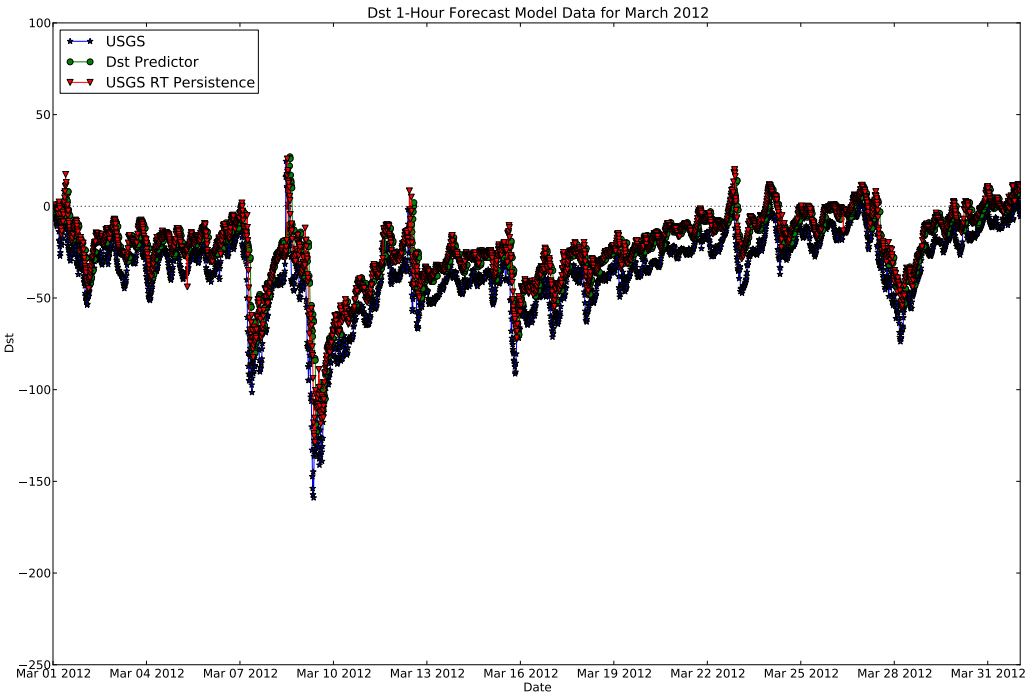

Figure 3: 1-hour forecasts of Dst Predictor and Persistence for March 2012. Definitive Dst is plotted for comparison.

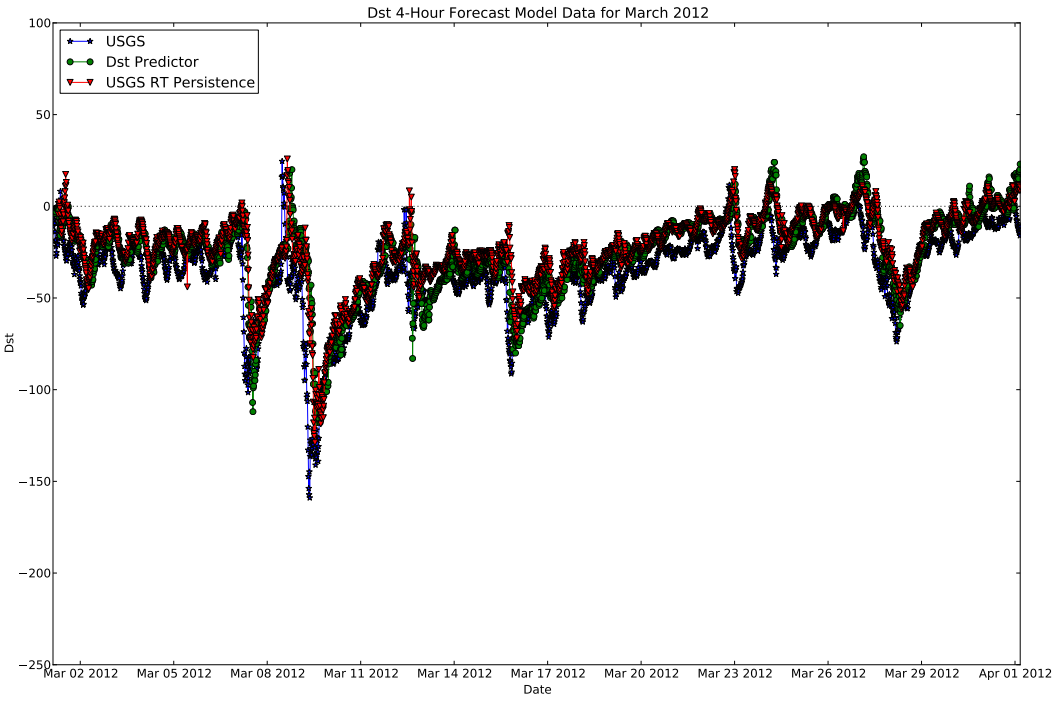

Figure 4: 4-hour forecasts of Dst Predictor and Persistence for March 2012. Definitive Dst is plotted for comparison.

Table 4: Dst Predictor Performance: Figures 5 and 6

Forecast	CC	PE	SS_{pers}	PE_{pers}
1-hour	0.827	-105	0.158	-125
4-hour	0.769	-102	0.181	-124

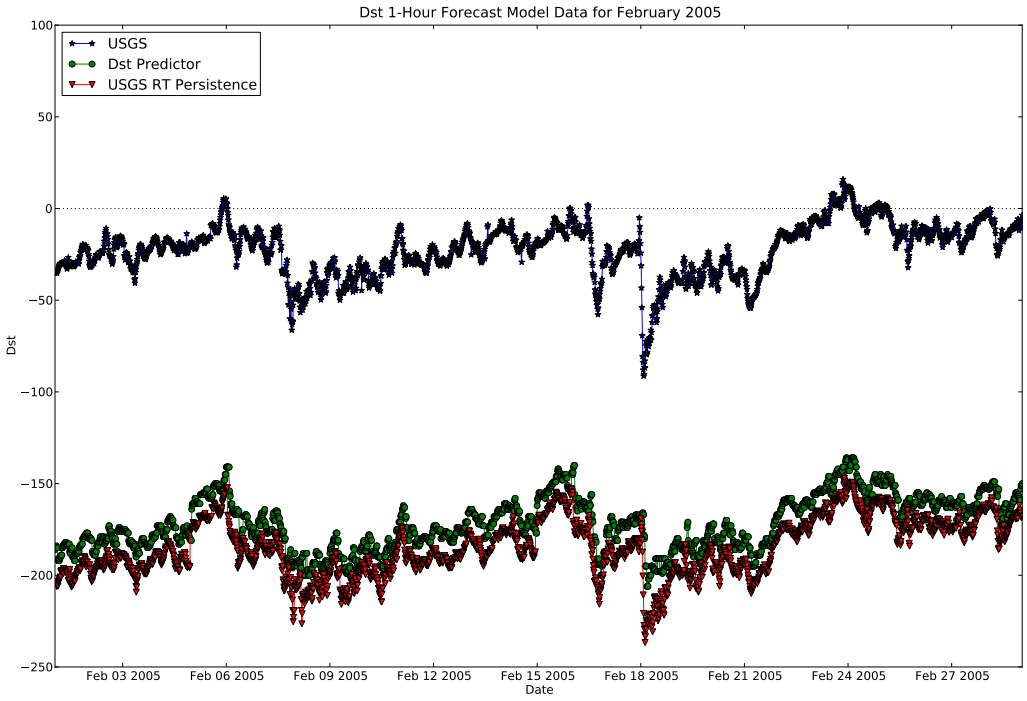

Figure 5: 1-hour forecasts of Dst Predictor and Persistence for February 2005. Definitive Dst is plotted for comparison.

Next, February 2005 is a time period where the RT Dst is significantly offset from the Definitive Dst (See Figures 5 and 6). The RT Dst is significantly more negative that the definitive, and Dst Predictor values are seen to be closer to Definitive than the nowcast persistence, suggesting that the model is being informed correctly by the solar wind inputs. This is reflected in the positive skill scores for both the 1-hour and 4-hour predictions (Table 4).

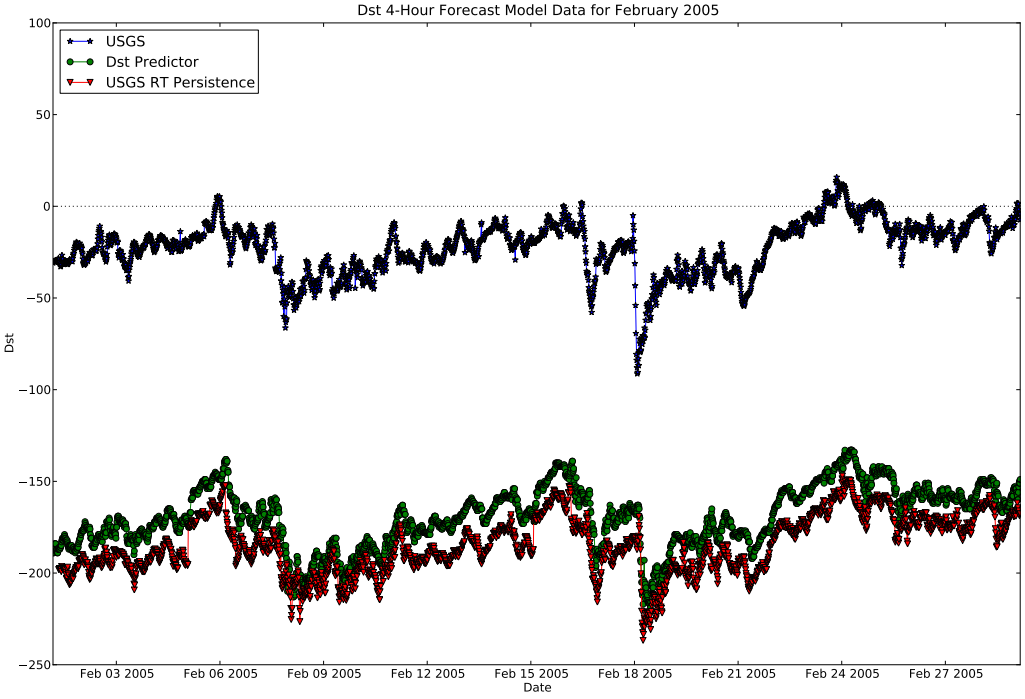

Figure 6: 4-hour forecasts of Dst Predictor and Persistence for February 2005. Definitive Dst is plotted for comparison.

Table 5: Dst Predictor Performance: Figures 7 and 8

Forecast	CC	PE	SS_{pers}	PE_{pers}
1-hour	0.908	-19	0.147	-22.8
4-hour	0.834	-60.5	-1.57	-22.8

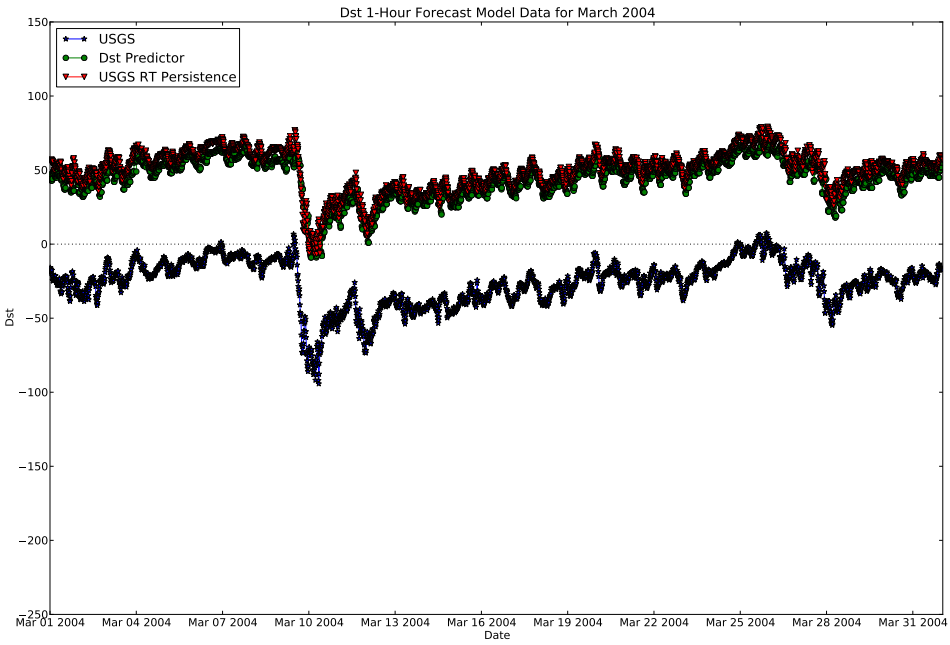

Figure 7: 1-hour forecasts of Dst Predictor and Persistence for March 2004. Definitive Dst is plotted for comparison.

Finally, March 2004 is a time period where the RT Dst is significantly offset from the Definitive Dst in a positive manner (See Figures 7 and 8). The RT Dst is significantly more positive than the Definitive, and falls on values that are almost never seen in Definitive Dst. Because Dst Predictor has been trained on datasets without values this high, the model is not stable and can produce values even more positive. This is particularly noticeable with the 4-hour prediction (Figure 8), suggesting the greater lead-time allows for more varied model output. The “unstable” behavior could be mitigated with an appropriate filter applied to the nowcast data prior to ingest.

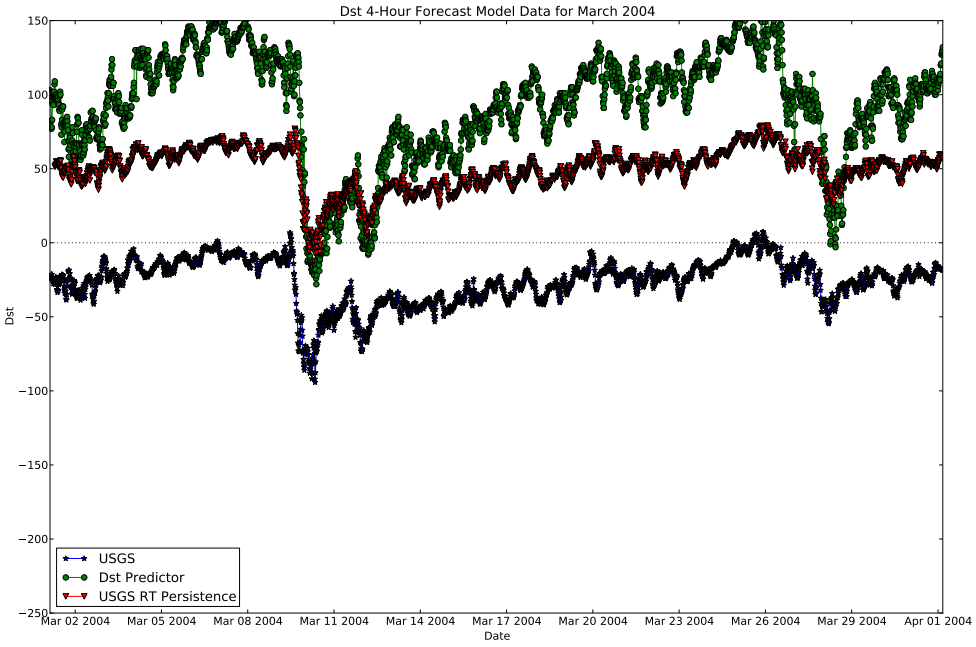

Figure 8: 4-hour forecasts of Dst Predictor and Persistence for March 2004. Definitive Dst is plotted for comparison.

5 Conclusions

The performance of the Dst Predictor against Definitive Dst values is dominated by the quality of the nowcast specification. Where the nowcast is a good approximation of the Definitive Dst, Dst Predictor performs well and can add value. Where the nowcast is significantly offset relative to the Definitive Dst, the model is not going to get much closer. In addition, several other factors may further inhibit Dst Predictor performance: insufficient NN training; "stale" NN weights that need to be updated; or discrepancies between the official and nowcast data used for training the model and the nowcast data as fed in by SWAFS. Regardless, given the overall performance results (Table 1), Dst Predictor cannot be said to meet a requirement of predicting the Definitive Dst with 1-hour and 4-hour lead times. On the other hand, persistence is often hard to beat for 1-hour forecasts yet the Dst Predictor clearly performs well against that standard.

The fundamental conclusion is this: Dst Predictor outperforms the nowcast persistence model. However, the large discrepancies that can occur between the RT and Definitive Dst values suggest investigating alternative models that do not rely on nowcast inputs.

6 Recommendation

We recommend continued use of the Dst Predictor model for 1-hour and 4-hour Dst predictions. While the RT Dst has already improved (Figure 2 versus Figure 1) and we expect further improvement, we recommend active study of other Dst forecast models that do not rely on nowcast inputs that can be used both to help identify bad nowcast value and as supplemental forecasts when bad data are identified.

REFERENCES

[1] Gannon, J. L., Love, J. J., Friberg, P. A., Stewart, D. C., and Lisowski, S. W., U.S. Geological Survey Near Real-Time Dst Index: U.S. Geological Survey Open-File Report 2011-1030, 2011.

[2] McCollough, J. P., and Young, S. L., Real-Time Validation of the Kp Analysis Model, Delivered to SWAFS SPO, 2013.

[3] McCollough, J. P., Young, S. L., and Frey, W. R., Real-Time Validation of the Kp Predictor Model, Delivered to SWAFS SPO, 2013.

[4] NOAA SWPC, GLOSSARY OF SOLAR-TERRESTRIAL TERMS, URL: http://www.swpc.noaa.gov/info/glossary.html#d, 2009.

[5] Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, vol. 100 of *International Geophysics Series*, 3rd ed., Academic Press, 2011.

[6] Wing, S., Johnson, J. R., Jen, J., et al., Kp forecast models, *Journal of Geophysical Research: Space Physics*, *110*, A04,203, 2005.