PRediction Of Geospace Radiation Environment and Solar wind parameterS

Work Package 3
Forecast of the evolution of geomagnetic indices

Deliverable 3.3
Evaluation and verification of a set of selected existing models

P. Wintoft, M. Wik, R. Boynton, M. Balikhin
February 29, 2016

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 637302.

Document Change Record

Issue	Date	Author	Details
v1	Jan 29, 2016	P. Wintoft, M. Wik	Initial draft
v2	Feb 19, 2016	P. Wintoft, M. Wik	Updated sections 2.2, 2.3, and 3.
v3	Feb 29, 2016	P. Wintoft, M. Wik, R. Boynton, M. Balikhin	Various updates and added Sheffield Dst model.

Contents

Summary

The overall aim of WP 3 concerns improvement and new development of models based on data driven modelling, such as neural networks, support vector machines, and NARMAX. Existing models for *Dst* and *Kp* will be analysed and verified with the aim of finding weaknesses and to suggest improvements. Solar wind and geomagnetic indices shall also be analysed in order to develop models for the identification of features, such as (but not limited to) shocks, sudden commencements, and substorms. Such categorisation will aid the model development and verification, and can also serve as alternative approach to models providing numerical input-output mapping. In addition to the development of *Dst* and *Kp* models new models will be developed to forecast *AE*. The models will be implemented for real-time operation at IRF and data and plots will be provided on a web server.

This deliverable is targeted at the development of approaches to verify models prediction geomagnetic indices (*Kp*, *Dst*, *AE*) from solar wind data at L1. The methods will then be applied to existing forecast models available to the PROGRESS team. The results can be used for the further development of the models in WPs 3.4 and 3.5.

Acronyms

ACE	Advanced Composition Explorer
DSCOVR	Deep Space Climate Observatory
GFZ	GeoForschungsZentrum
GSFC	Goddard Space Flight Center
NASA	National Aeronautic and Space Administration
NCEI	National Centers for Environmental Information
NOAA	National Oceanographic and Atmospheric Administration
SWPC	Space Weather Prediction Center
WDC	World Data Center

1 Introduction

In order to monitor the progress of model development, to compare models, and to judge the validity of a model, methods have to be defined that capture differences in a comprehensive way. The meteorological community has a long history of developing methods to verify weather forecasts. Forecast verification is *the process of assessing the quality of a forecast* and a web site exists at the *Centre for Australian Weather and Climate Research* (CAWCR) devoted to this subject[1]. The site contains a comprehensive list of metrics (methods) to verify dichotomous (yes/no), multi-category, and continuous valued forecasts. To assess forecasts three types of goodness, in a general sense, can be identified (Murphy 1993): 1) Consistency, 2) Quality, 3) Value. *Consistency* is concerned with the relation between the forecaster's knowledge base and provided forecast. In weather forecasting the forecaster receives input from many sources, like numerical weather models

[1] http://www.cawcr.gov.au/projects/verification/

and observations, to provide a forecast. *Quality* is concerned with the degree of correspondence between forecasts and observations. *Value* is concerned with the value of the forecasts to the users. Although quite obvious, but still fundamental, it is clear that a forecast in itself has no value; it is the user who adds value. Cost/loss or cost/benefit analysis can be used in determining the value of forecasts. Two examples within space weather are reports produced by two different teams within the *ESA Space Weather Programme Feasibility Studies*[2]. In this work we will primarily address how to assess the degree of correspondence between forecast model outputs and observations, i.e. the *quality* of the forecasts.

2 Verification approaches for models driven by L1 measurements

The prediction lead time from L1 to geomagnetic indices is very short, in terms of sampling frequency only one or a few samples ahead. As described in D3.1, the L1-magnetopause travel time provides a lead time of 10 to 80 minutes depending to some degree on spacecraft location but mainly on solar wind speed. When the solar wind disturbance reach the magnetopause the interaction starts with the development of substorms and storms, adding about another hour of lead time. The three indices have different time resolutions: *Kp* 3 hours, *Dst* 1 hour, and *AE* 1 minute. Predicting *Kp* one sample ahead is on the limit considering the physics involved, while it is possible to predict *Dst* one or two samples ahead. For *AE* some summary measure with reduced temporal resolution will be predicted, e.g. 10 minutes, but still with higher resolution than *Kp* and *Dst*, thus the number of samples ahead will be higher. If it assumed that real-time observations of the indices are available then a persistence model, or variants thereof, will receive very high verification scores for most measures as the storm dynamics is longer that the forecast lead time.

As the indices work on different time scales and different physical processes are involved in their determination we will look at the indices individually to understand differences and commonalities in the verification approaches.

2.1 *Dst*

The *Dst* index "is probably the one that monitors and records with the greatest accuracy the phenomenon for which it was designed" (Mayaud 1980), namely the equatorial ring current. The *Dst* index is derived from the horizontal magnetic field from four observatories equally distributed in longitude and 20° to 30° from the magnetic equator. The official values are provided with hourly resolution, although it can be calculated with higher cadence. The index is given in units of nT and can be both positive and negative with no bounds. Under quiet conditions *Dst* is close to zero. The typical *Dst* storm goes through three phases: the initial phase, the main phase, and the recovery phase.

[2]http://www.esa-spaceweather.net/spweather/esa_initiatives/spweatherstudies/public_doc.html

The initial phase shows an increase in *Dst* and is caused by a dynamic pressure increase in the solar wind acting on the magnetopause. The only available lead time is that from the L1-magnetopause travel time as the magnetospheric response is immediate. The main phase is basically related to when the solar wind magnetic field component B_z turns negative, enabling reconnection, driving the magnetospheric storm with an increase of the equatorial ring current pushing the *Dst* index to negative values. The magnetospheric processes provides an additional hour of prediction lead time. The recovery phase sets in after the storm main phase during which the ring current decays. If there is no new storm during the decay the possible prediction lead time is several hours to days.

In an extensive model study by Rastätter et al. (2013) 12 different models, using different settings, resulting in 26 different model runs were verified against observed *Dst* for 4 different storms. Both numerical MHD models and empirical models were compared. The study showed that the empirical models generally scored highest including the NARMAX model used here (Boynton et al. 2011). Although the study included many models it should be noted that very few events were included which makes it difficult to come to a more general conclusion.

In the setup in the current *Dst* models do not distinguish between the different possible lead times, but instead always makes an one-hour prediction from solar wind data at the magnetospheric bow shock location.

The available models have been run on the OMNI dataset extending over the year 1998–2014, i.e. over 17 years consisting of about 149 000 hourly values. The results are summarised in Table 1 for the 4 models together with observed and persistence *Dst*

$$Dst_{\mathrm{PERS}}(t) = Dst(t-1). \tag{1}$$

Naturally, the persistence *Dst* has the same statistics as observed *Dst* but is included for consistency. The models included here are:

- BMR: Burton et al. (1975)
- OM: O'Brien & McPherron (2000)
- LUND: Lundstedt et al. (2001)
- SN_1: Boynton et al. (2011) and web site[3].

The models are described in D3.1.

The summary reveals that the mean and median observed *Dst* are slightly below zero (−13 and −9 nT, respectively), and that only 1% of the hourly *Dst* reach below −88 nT. The models show similar results. The strongest storm reach $Dst = -422$ nT and only the SN_1 model comes close to that value. The LUND model only reach -230 nT, while the BMR model overshoots by a similar amount. The OM model reaches −335 nT. However, the lowest predicted *Dst* for each model may not belong to the same storm. We will study this more in Section 3.1.

[3]http://www.ssg.group.shef.ac.uk/USSW/UOSSW.html

	Dst	LUND	SN_1	BMR	OM	PERS
count	149016	147374	146343	147383	147383	149015
mean	-13	-19	-13	-12	4	-13
std	21	17	20	21	26	21
min	-422	-230	-437	-650	-335	-422
1%	-88	-83	-76	-93	-84	-88
50%	-9	-16	-10	-7	10	-9
99%	20	7	18	7	36	20
max	95	31	62	32	48	95

Table 1: Summary statistics of observed and predicted *Dst* based on the years 1998 to 2014.

We regard *Dst* as a continuous variable and a few standard measures and skill scores (Déqué 2012) are suggested in the following. With x as the observed variable and y the predicted variable the error for sample i is

$$e_i = y_i - x_i \tag{2}$$

where positive errors correspond to predicted values above the observed.

The mean bias is defined as

$$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{3}$$

where n is the total number of samples. A model with perfect forecasts has BIAS = 0. However, a BIAS of zero does not imply perfect forecasts as errors with opposite signs may cancel.

The mean-absolute-error, the mean-square-error, and the root-mean-square-error are defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i| \tag{4}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{6}$$

In all cases a value of zero indicates perfect forecasts. Both MAE and RMSE have the same units as the observed variable. Both MSE and RMSE are more influenced by forecasts with large errors compared to MAE.

The linear correlation coefficient is another standard measure

$$\text{CORR} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{7}$$

with perfect forecasts CORR = 1.

A skill score is used to compare the performance of a model with a reference model. A standard skill score uses the MSE as the measure and is defined as

$$\text{MSESS} = \frac{\text{MSE} - \text{MSE}_{\text{ref}}}{\text{MSE}_{\text{perfect}} - \text{MSE}_{\text{ref}}} = 1 - \frac{\text{MSE}}{\text{MSE}_{\text{ref}}} \tag{8}$$

where MSE_{ref} is the MSE from the reference model and $\text{MSE}_{\text{perfect}}$ is the MSE for a perfect model. As $\text{MSE}_{\text{perfect}} = 0$ that term can be eliminated. A model with MSESS = 1 indicates perfect forecasts, MSESS = 0 forecasts comparable with the reference model, and MSESS < 0 poorer than the reference model.

The measures and skill score are computed for all models on all data and are summarised in Table 2. The MSE skill score is computed using two different referens models, the persistence model and the BMR model, respectively. Obviously the persistence model shows the best result for all measures, however, we also know that the persistence model can not provide timely forecasts. With that background knowledge we know the limitations of the persistence model, but based only on the measures it is not apparent. The LUND, SN_1, BMR, and OM models do not use past values of observed *Dst* as input, however, any model that does that will likely perform well using the above measures due to the high autocorrelation in *Dst*. But of course, if past values of *Dst* are available in real time then it is motivated to use it as input, however, high scores need to be checked so that they are not simply the result of persistence. We now look into different ways of exploring this.

	BIAS	MAE	RMSE	CORR	MSESS:PERS	MSESS:BMR
LUND	-6.69	10.06	12.91	0.85	-7.20	0.24
SN_1	-0.96	8.15	11.35	0.84	-5.34	0.41
BMR	0.46	9.94	14.83	0.74	-9.82	nan
OM	16.74	19.60	23.07	0.78	-25.20	-1.42
PERS	-0.00	2.75	4.51	0.98	nan	0.91

Table 2: Measures and scores for observed *Dst* and predicted *Dst* using all data.

As *Dst* is dominated by quiet conditions (see Table 1) it is natural to select a subset with non-quiet data. Selecting only those samples where observed $Dst < -50$ nT, about 4% of the data, results in the measures and scores in Table 3. In most cases the performance measures becomes slightly poorer, but persistence still leads. However, the MSESS:BMR improves significantly for all models except for the persistence model.

The selection of data to consider can be further refined by considering storm events. Storms have previously been identified using various criteria and also involving some manual efforts. See for example event list by Echer et al. (2008) for $Dst \leq -100$ nT. Our approach here is to have an automatic procedure to select storms, with the start times, end times, and time of minimum *Dst* (largest negative value). Of course, there will always be some ambiguity to the definition of storm extent. We apply a lowpass wavelet filter using the Maximal Overlap Discrete Wavelet Transform (MODWT) (Percival & Walden 2000) to study variations in *Dst* on typical storm time scales. Some experimenting shows

	BIAS	MAE	RMSE	CORR	MSESS:PERS	MSESS:BMR
LUND	10.40	19.56	27.86	0.72	-3.79	0.64
SN_1	13.80	20.13	28.18	0.76	-3.90	0.63
BMR	8.26	33.59	46.16	0.73	-12.15	nan
OM	16.44	26.15	33.42	0.72	-5.89	0.48
PERS	1.25	7.59	12.73	0.95	nan	0.92

Table 3: Measures and scores for observed *Dst* and predicted *Dst* when observed $Dst < -50$ nT.

that filtering at level $J = 2$, corresponding to lowpass filter of about 8 hours, keeps the main characteristics of the storms without loosing too much detail. The locations of the maxima in the filtered series are then used to mark the start and end times for each event. For a random signal that would produce about $n/8$ events, where n is the length of the series. The *Dst* series results in close to 11 000 events, which is less than the random limit of $149\,000/8 \approx 19\,000$. Most events are uninteresting as they merely corresponds to some random fluctuations. To select an event as a storm event we further require that the minimum *Dst* within an event to be less than -50 nT. Through this procedure 507 storm events are identified and one example is given in Figure 1. We also define the period from start (first dashed line) to minimum (dotted line) as the main phase, and the period from minimum to end (second dashed line) as the recovery phase. It should be noted that our use of "main phase" includes both the original definitions of initial and main phases.

The errors between observed and predicted *Dst* are summarised in Figure 2 for all data (upper left), storm data (upper right), main phase (lower left), and recovery phase (lower right). As before, the persistence model has the smallest errors, with median errors close to zero for all data and storm data. However, the main phase is dominated by positive errors and the recovery phase by negative errors, which highlights the fact that the persistence model is always lagging after the observed *Dst*.

The LUND model predicts slightly more negative *Dst* on average, but for storms, main phases, and recovery phases the median errors are very close to zero, thus implying timely predictions. The SN_1 model has median errors close to zero when all data are considered, and slightly positive medians for the other subsets. The BMR and OM models show positive errors for the storm data, but consistently so also for the main and recovery phases. This could mean that there is a positive bias in the forecasts and after removal they would also have medians close to zero for both phases, thus also indicating timely predictions.

Another way of detecting lags in the predictions is to shift the predictions in time and compute the errors. In Figure 3 the predictions have been shifted -1, 0, 1, and 2 hours and the RMSE computed on the main phase data. The minimum RMSE should appear at the 0 hours shift for timely predictions. The persistence model shows, as expected, minimum RMSE at a lag of one hour. The LUND and BMR models have minima at zero lag, while the SN_1 and OM models have minima at 1 hour.

Finally, another test is to estimate the number of storm main phase events that were timely predicted. There is some ambiguity to define what is meant by timely prediction,

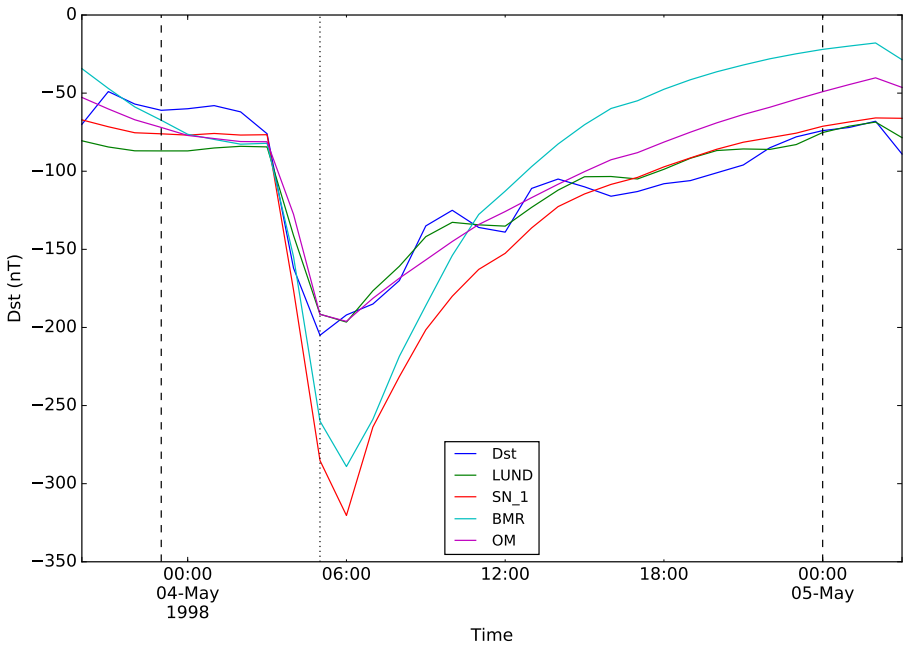

Figure 1: A storm in 1998 with observed and predicted *Dst*. Dashed lines mark start and end of the event, and the dotted line marks observed *Dst* minimum.

but we consider a prediction to be timely if the RMSE using observed *Dst* is smaller than the RMSE on the one hour shifted *Dst*. For all storm main phases the number of correct predictions, in terms of timeliness, are counted. The result is summarised in Table 4 where also the fraction of correct is given. The persistence model is included for completeness, but naturally with this definition no predictions are correct as seen in the table.

	n	p
LUND	273	54
SN_1	174	35
BMR	203	40
OM	123	24
PERS	0	0

Table 4: Number (n) and percent (p) of correctly predicted storm main phases. See text for meaning of correct.

With the above issues in mind when comparing models without and with past values of *Dst* as input, where persistence is a special case, we compute the measures and scores for the storm data, main phase data, and recovery data. The results are shown in Tables 5 to 7. However, the persistence model still scores best.

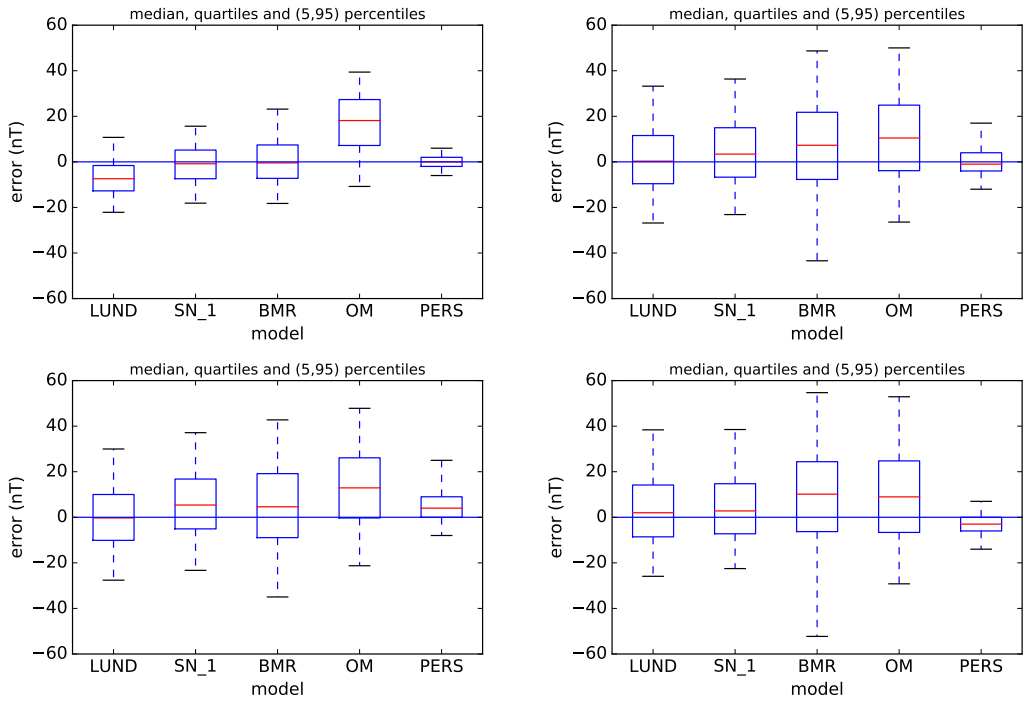

Figure 2: Error box plot of predicted *Dst* vs. observed *Dst*. For each model the median error (red horizontal line), the quartiles (box), and the 5 and 95% percentiles (whiskers) are shown. The errors are computed on all data (top left), storm data (top right), main phase data (bottom left), and recovery phase data (bottom right).

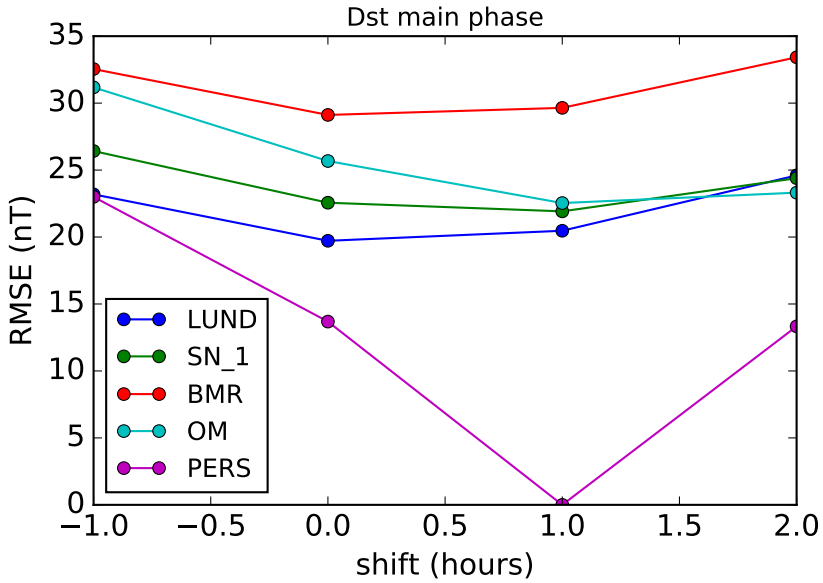

Figure 3: RMS errors of the predictions compared to observed *Dst* shifted by -1, 0, 1, and 2 hours.

	BIAS	MAE	RMSE	CORR	MSESS:PERS	MSESS:BMR
LUND	1.92	14.56	21.48	0.83	-2.94	0.58
SN_1	4.67	14.80	22.02	0.83	-3.14	0.56
BMR	5.15	22.48	33.32	0.79	-8.48	nan
OM	10.92	20.30	26.40	0.81	-4.95	0.37
PERS	0.77	6.62	10.82	0.96	nan	0.89

Table 5: Measures and scores for observed *Dst* and predicted *Dst* for storm events.

	BIAS	MAE	RMSE	CORR	MSESS:PERS	MSESS:BMR
LUND	0.37	13.65	19.72	0.88	-1.08	0.54
SN_1	5.62	15.38	22.56	0.85	-1.72	0.40
BMR	3.69	19.71	29.12	0.83	-3.53	nan
OM	13.08	19.86	25.68	0.85	-2.52	0.22
PERS	5.84	8.38	13.68	0.96	nan	0.78

Table 6: Measures and scores for observed *Dst* and predicted *Dst* for storm main phases.

	BIAS	MAE	RMSE	CORR	MSESS:PERS	MSESS:BMR
LUND	4.18	15.65	23.34	0.79	-7.86	0.60
SN_1	5.01	14.95	22.34	0.81	-7.12	0.64
BMR	6.63	25.19	36.98	0.76	-21.24	nan
OM	9.99	21.15	27.66	0.76	-11.44	0.44
PERS	-2.83	5.34	7.84	0.98	nan	0.96

Table 7: Measures and scores for observed *Dst* and predicted *Dst* for storm recovery phases.

2.2 *Kp*

The global range index *Kp* (Mayaud 1980) is a weighted summary of local *K* indices derived from 13 mainly northern mid-latitude magnetic observatories. The *K* index measures semi-logarithmic the range of variation of the local horizontal magnetic field over a 3-hour interval, with each interval fixed in UT windows 00-03, 03-06, ..., 21-24. Thus, different physical processes with different temporal dynamics and from different domains are merged into one variable although it is mainly controlled by the main geomagnetic storm. This means that within the 3-hour *Kp* interval there will be variations in the horizontal magnetic field with different time lags following the solar wind disturbance, although for the larger *Kp* values the lag will be about an hour or longer.

The *Kp* index is discrete and can only be one of the 28 values

$$Kp \in \{0_0, 0_+, 1_-, 1_0, 1_+, \ldots, 8_0, 8_+, 9_-, 9_0\} \tag{9}$$

where the range of variability increases for increasing *Kp*, as shown in Table 8. The *Kp* values have been translated into *ap* according to Mayaud (1980), where *ap* is in units of 2 nT. Thus, the resolution is higher below 100 nT than above, and the scale has an upper limit.

Kp	g = ap	nT
0	0	0
1	4	8
2	7	14
3	15	30
4	27	54
5	48	96
6	80	160
7	132	264
8	207	414
9	400	800

Table 8: Relation between *Kp* and range ap in units of 2 nT and nT.

The analysis for *Kp* is performed in a similar way as for *Dst*. The models, described in D3.1, are:

- LUND_NC: Nowcast of current *Kp* (Boberg et al. 2000)
- LUND_FC: Forecast of next *Kp* (Boberg et al. 2000)

The models are driven by solar wind data. The nowcast model predicts the *Kp* value valid for the latest 3-hour interval, while the forecast model predicts *Kp* value valid for the coming 3-hour interval.

The results are summarised in Table 9 for 2 different *Kp* models together with persistence. *Kp* is dominated by small values with observed median of 2_- (1.7) and similar

	Kp	LUND_NC	LUND_FC	PERS
count	49672	48866	48865	49671
mean	1.8	2.3	2.4	1.8
std	1.4	1.0	0.9	1.4
min	0.0	0.0	0.1	0.0
1%	0.0	0.5	1.0	0.0
50%	1.7	2.1	2.2	1.7
99%	6.0	5.2	4.9	6.0
max	9.0	9.0	9.0	9.0

Table 9: Summary statistics of observed and predicted *Kp* based on the years 1998 to 2014.

model medians. The 99% percentile is at $Kp = 6_0$ and with slightly lower model values. The models also reach the highest values of $Kp = 9_0$.

The scatter plot shows that the models tend to predict higher values than observed for $Kp \lesssim 3$. However, from Table 8 it is seen that those values correspond to variations less than 30 nT. For the nowcast model there is a tendency for slight under-prediction in the mid *Kp* range, but it improves for the highest values. The forecast model performs slightly poorer.

The measures and skill scores based on all data are shown in Table 10. In this case, compared to *Dst*, the persistence model has only marginally better scores due to the weaker autocorrelation in the *Kp* series. But, still, based only on these scores the persistence model would be the winner.

	BIAS	MAE	RMSE	CORR	MSESS:PERS
LUND_NC	0.43	0.72	0.87	0.84	-0.05
LUND_FC	0.54	0.88	1.04	0.78	-0.50
PERS	-0.00	0.63	0.85	0.81	nan

Table 10: Measures and scores for observed *Kp* and predicted *Kp* using all data.

The temporal evolution of *Kp* is different from *Dst* and we choose not to identify individual storms for the continued analysis, but instead simply look at events when there is increase or decrease in activity. Now both models perform better than the persistence model (Table 11). In the case of decreasing *Kp* the persistence model scores best (Table 12).

The error box plots for the three different datasets are shown in Figure 5. Similar to the *Dst* analysis the persistence errors are distributed around zero for the full dataset (top panel), and negatively and positively distributed for the increasing and decreasing sets, respectively. The models show positively distributed errors for the full set due to that the majority of *Kp* values are low and that the models over-predict low values. However, for the increasing events the errors are distributed around zero.

The timeliness plot of the *Kp* predictions (Figure 6) shows a minimum in RMSE at no

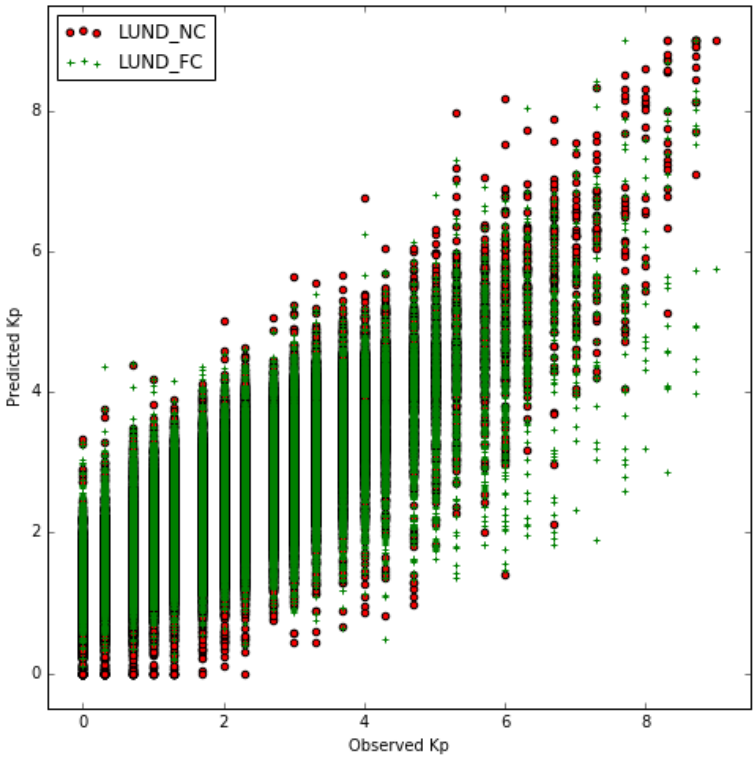

Figure 4: Scatter plot of predicted *Kp* vs. observed *Kp*.

	BIAS	MAE	RMSE	CORR	MSESS:PERS
LUND_NC	0.06	0.58	0.73	0.88	0.44
LUND_FC	-0.02	0.69	0.89	0.79	0.18
PERS	-0.80	0.80	0.98	0.91	nan

Table 11: Measures and scores for observed *Kp* and predicted *Kp* using only data when *Kp* increses.

	BIAS	MAE	RMSE	CORR	MSESS:PERS
LUND_NC	0.67	0.83	0.98	0.78	-0.27
LUND_FC	0.95	1.04	1.19	0.81	-0.86
PERS	0.72	0.72	0.87	0.92	nan

Table 12: Measures and scores for observed *Kp* and predicted *Kp* using only data when *Kp* decreases.

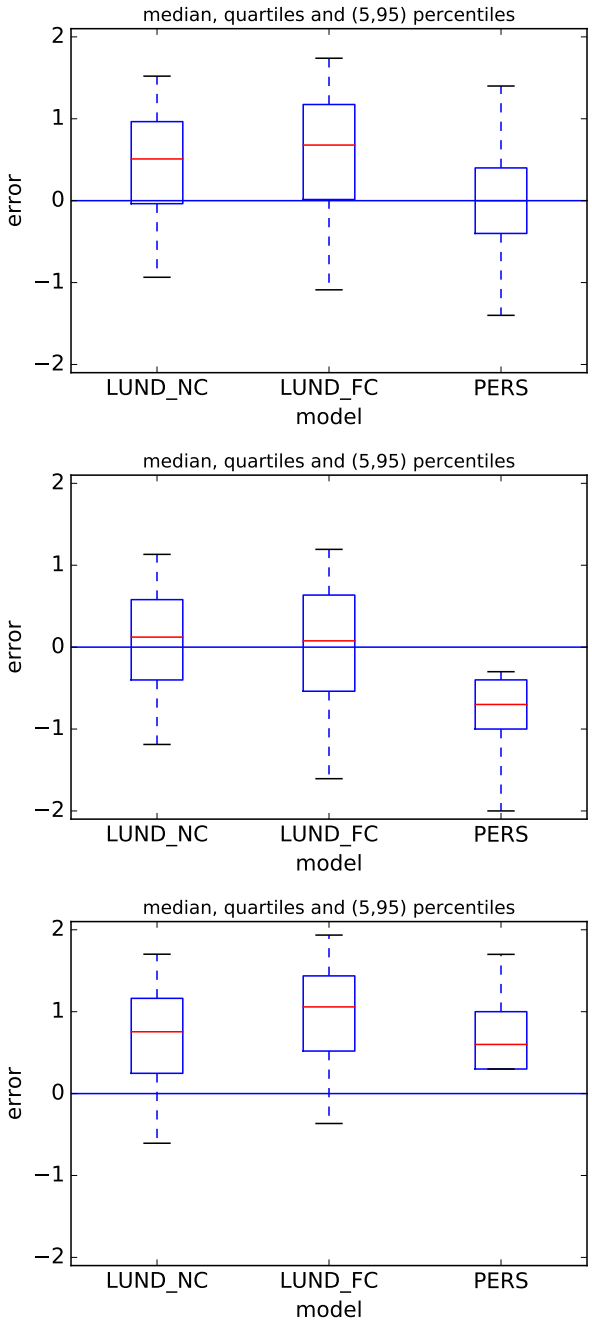

Figure 5: Error box plot of predicted Kp vs. observed Kp. For each model the median error (red horizontal line), the quartiles (box), and the 5 and 95% percentiles (whiskers) are shown. The errors are computed on all data (top), increasing events (middle), and decreasing events (bottom).

shift for the nowcast model, while the forecast model shows a minimum at 3 hours shift, indicating some lag in the predictions.

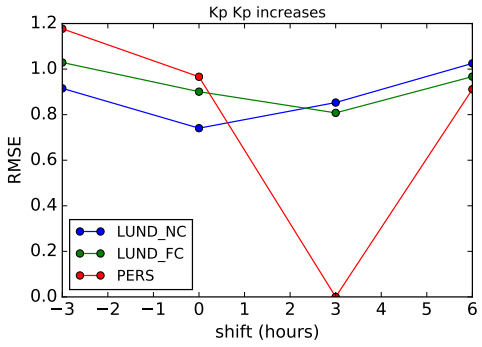

Figure 6: RMS errors of the predictions compared to observed *Kp* shifted by -3, 0, 3, and 6 hours.

The number and fraction of correctly predicted *Kp* increase events are summarised in Table 13. See section on *Dst* for the definition of correct. The fraction correct is high, above 70%, which means that most of the events have smallest RMS errors at the correct time.

	n	p
LUND_NC	10554	78.731816
LUND_FC	9886	73.748601
PERS	0	0.000000

Table 13: Number (n) and percent (p) of correctly predicted *Kp* increase events. See text for meaning of correct.

2.3 *AE*

Currently, we do not have any operational *AE* prediction models, therefore we have not carried out any tests. However, the same strategies that we have outlined for *Kp* and *Dst* will be useful for the verification of the coming *AE* prediction models.

3 Verification approaches for models driven by L1 predictions

The models developed in WP 2 will provide forecast of solar wind plasma and magnetic field vector at L1 for ambient solar wind and high speed streams from coronal holes with a temporal resolution of 1 hour. This will provide a prediction lead time of several days.

However, the resolution and accuracy at L1 will not be as high as the measurements provided by ACE and DSCOVR spacecraft.

The verification of the models predicting indices from L1 using predicted solar wind data as input is different from the verification when using measured solar wind data. Models relying on solar wind L1 measurements can only make predictions with 1–2 hours lead time and therefore the timeliness is important. Those models may also used past values of the geomagnetic index as inputs when they are available in real time. Performance measures of index-prediction models that are driven by the predicted solar wind are not affected by persistence as the autocorrelation is small past a few hours lag. However, there is some recurrence in geomagnetic activity due to the persistence of coronal holes. An example with recurrent high-speed streams from coronal holes is shown in Figure 7. There is also an CME causing the highest Kp of about 8 and $Dst \approx -150$ nT.

The ≈ 13.5 days recurrence is most clearly visible in the solar wind velocity and Kp, and somewhat less visible inDst due to their linear scales and that coronal holes do not produce as strong geomagnetic activity as CMEs do. For example, the strongest Dst storms caused by corotating interaction regions (CIR) reach about -150 nT (Echer et al. 2008, Ji et al. 2012).

3.1 *Dst*

As the prediction lead time is comparable to the duration of individual storms it is interesting to study event based metrics, similar to that of Ji et al. (2012) where they identified 63 storms in Dst and related them to different types of solar wind drivers. In addition to the linear correlation (CORR) and RMS error, they also studied the difference between observed and predicted Dst minima $\Delta Dst_{\min}$, and the difference in the times of the observed and predicted minima $\Delta t_{\min}$. However, the time of minimum is not always a well defined quantity. There may be several Dst values close to the minimum but spread out over several hours. Figure 8 shows the differences in Dst and timings of the two lowest Dst values for each event in the Echer et al. (2008) storms list. In some cases the two lowest Dst are well separated in magnitude indicating a clear minimum, while many more events have differences less than 10 nT (top panel). Simultaneously, several of the minima with small differences in magnitude have large differences in timings (bottom panel).

Figure 9 shows one of the events with ambiguous time of minimum. The storm event duration used by Ji et al. (2012) is showed with the red bar. The automatic wavelet method classifies this as two events illustrated with green and blue bars.

The measures applied in Section 2.1 can also be applied here but now on an event-by-event basis. In addition the error between the observed and predicted minimum Dst is calculated. The measures for all storms with $Dst < -50$ nT were given in Table 3 and the errors for storm events in Figure 2 (top right).

Figure 10 shows the predicted vs. observed minima Dst for all events. Above -200 nT the predictions line up quite well with some scatter around the observed Dst. The six strongest storms are marked with blue vertical lines. It is seen that the predicted Dst have a quite large spread. The LUND model saturates at approximate -230 nT. The BMR model often strongly overestimates the Dst magnitudes for the larger storms. The SN_1 model shows quite a large variability from observed $Dst \lesssim -200$ nT, although for some

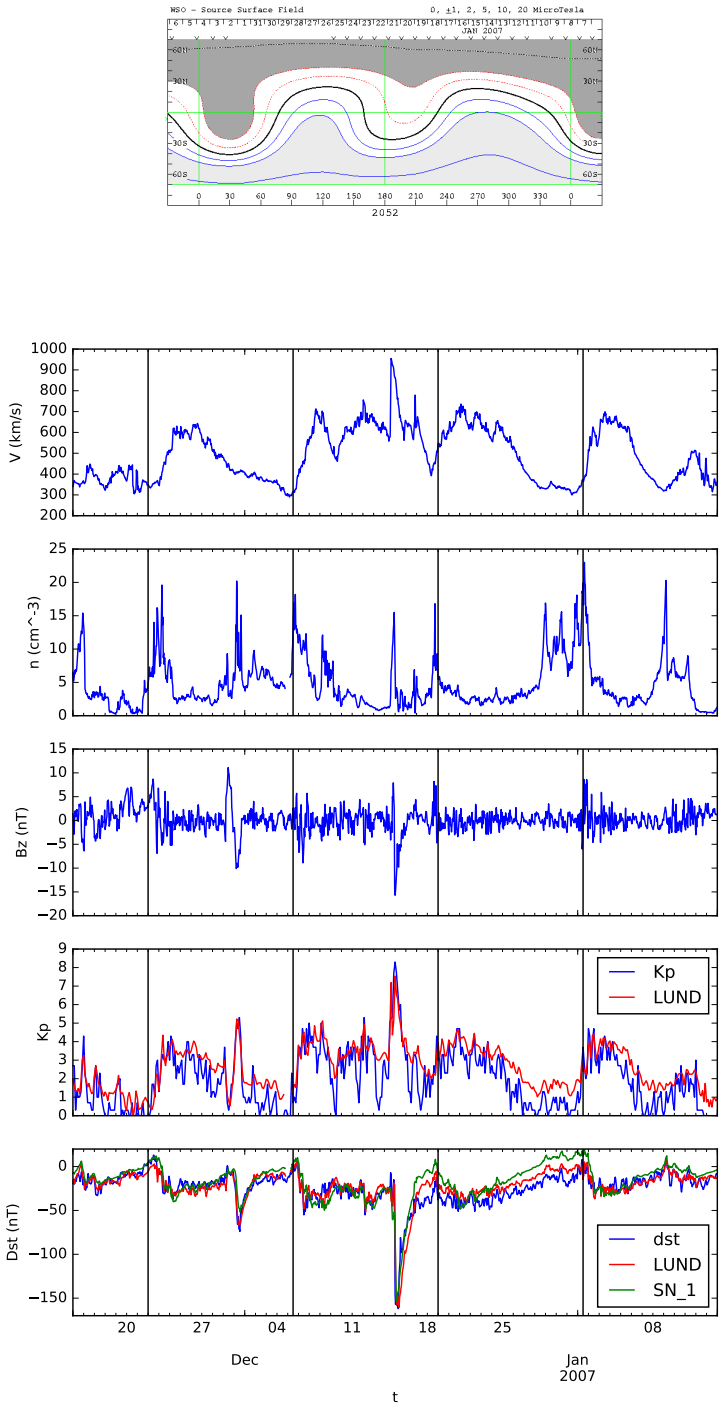

Figure 7: The figure shows from top to bottom: computed source surface magnetic field from Wilcox Solar Observatory, solar wind velocity, solar wind B_z, observed and nowcast Kp, and observed and forecast Dst. The vertical black lines are separated by 13.5 days.

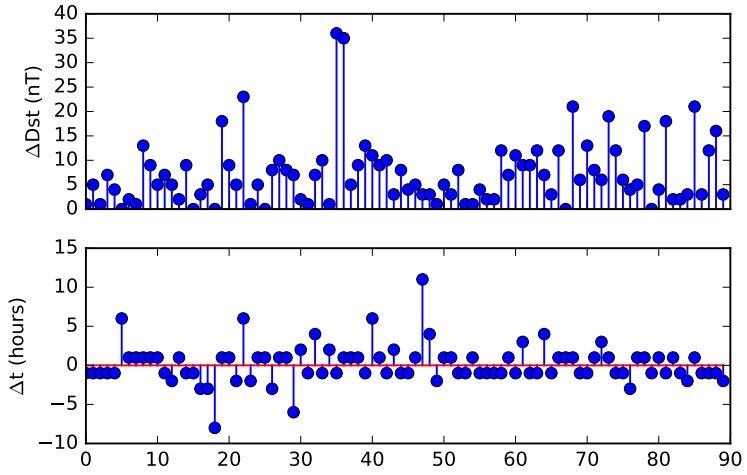

Figure 8: Top: The figure shows the difference between lowest and second-lowest *Dst* values for each event in the list by Echer et al. (2008). Bottom: The difference in hours between the times of the two lowest *Dst* values.

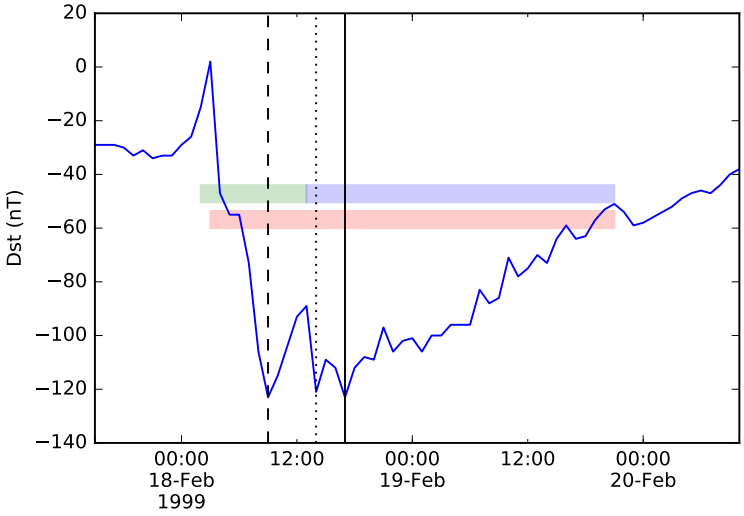

Figure 9: One event illustrating the ambiguity of the time of the *Dst* minimum. Vertical lines show the times of the lowest (solid), second lowest (dashed), and third lowest (dotted) *Dst* values. Coloured bars show extension of events. See text for description.

of the larger events it comes closest. The OM model generally underestimates the large events but with a small variability. Events 310 and 181 have datagaps in the solar wind data and therefore the predictions are wrong with *Dst* clustered around similar values. Both events are associated with proton events (Oct 30, 2003 and Nov 6, 2001) temporarily knocking out the ACE plasma instrument. The other events have solar wind data and the two strongest are shown in Figure 11 together with the predictions.

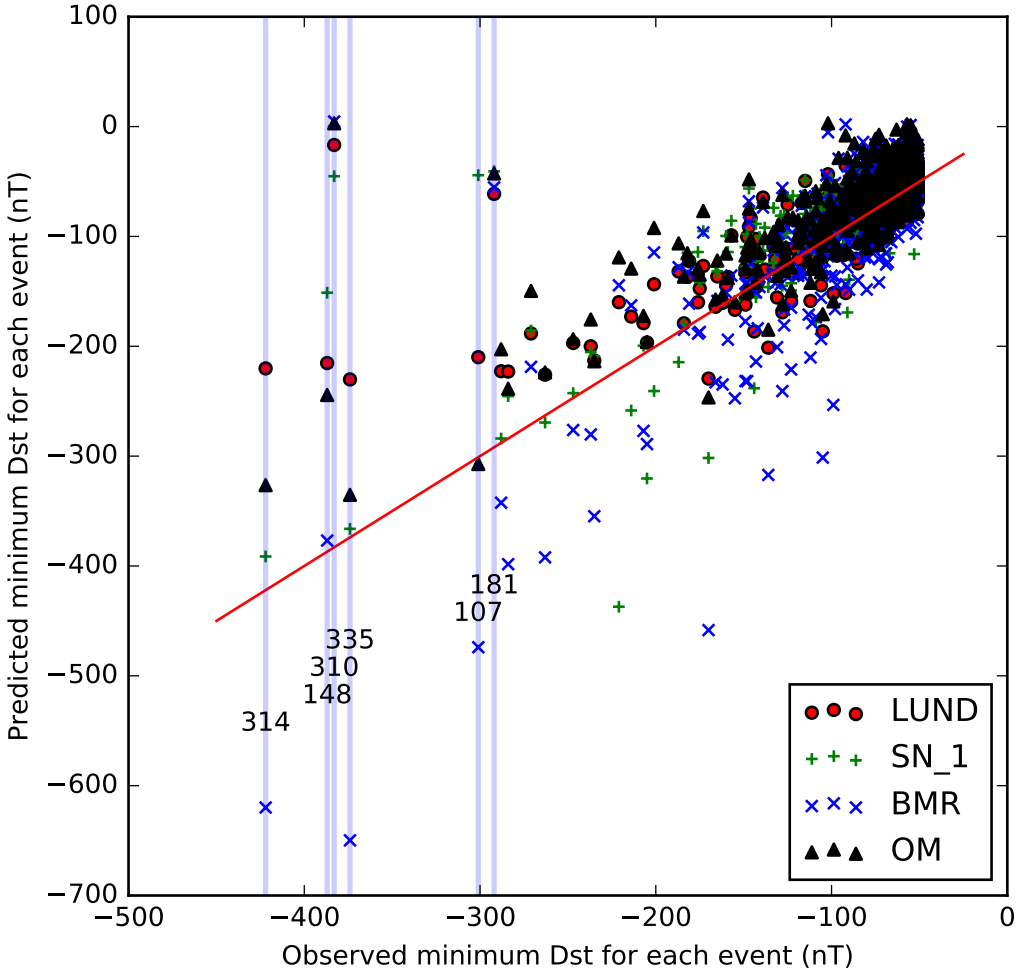

Figure 10: Scatter plot of predicted *Dst* vs. observed *Dst*.

The errors between observed and predicted *Dst* minima for each event are shown in Figure 12. The left plot highlights the errors that are distributed between the 5 and 95 percentiles, while the right plot also shows the largest individual event errors. Events 181 and 310 are excluded due to the lacking solar wind data. The LUND and SN_1 models have very similar ranges for the percentiles, both smaller than for the BMR and OM

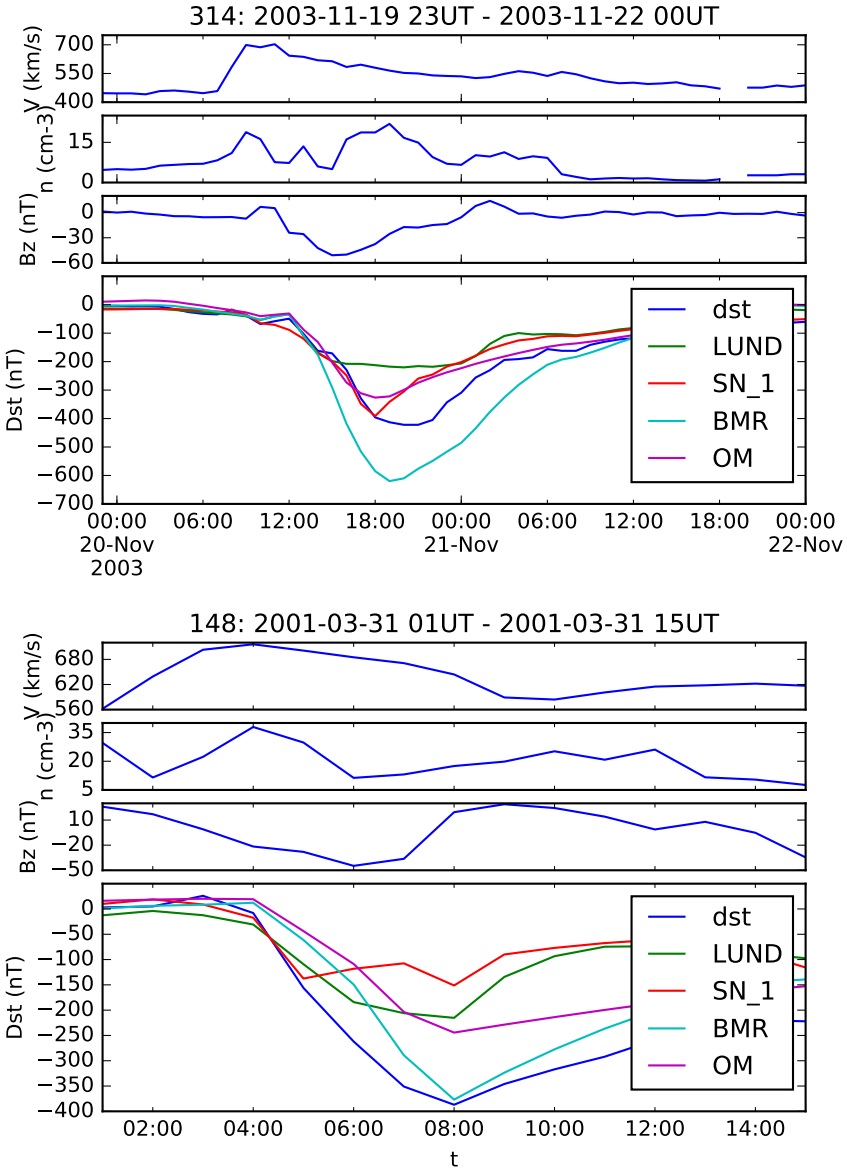

Figure 11: Observed and predicted *Dst* for the two strongest events.

models. The RMSE for all events are: LUND, 32 nT; SN_1, 37 nT; BMR, 46 nT; OM, 37 nT.

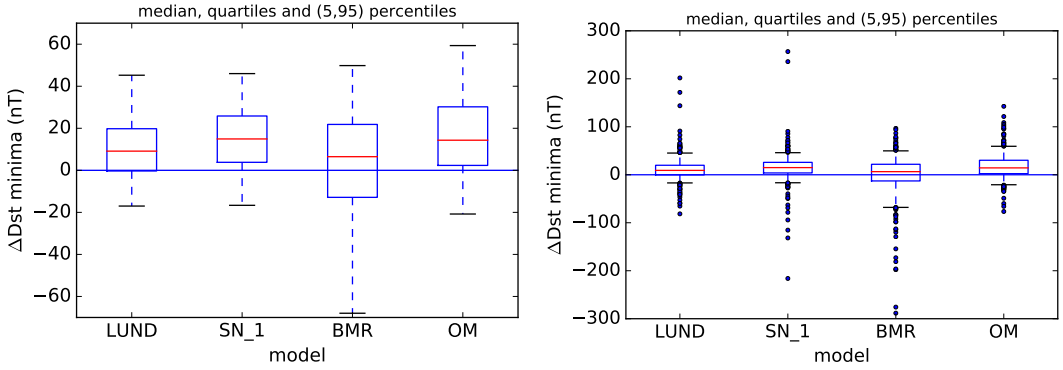

Figure 12: The errors between observed and predicted minima *Dst* for each event. Right plot shows top 10% largest errors as individual points. Events 181 and 310 have been excluded as solar wind data are missing.

3.2 *Kp*

The same measures used in Section 2.2 can be used here. We also add the event based verification by looking at the maxima *Kp* for individual events. Similar to that applied to *Dst* we do a wavelet filtering to find minima in the *Kp* series, where the times of minima mark event boundaries. After some experimenting a filter at level 4 was used, corresponding to a lowpass filter of $3 \cdot 2^{4+1} = 96$ hours $= 4$ days. An example with identified minima is shown in Figure 13. With this approach we identify 1076 events with maximum *Kp* distributed according to Table 14.

Kp	n
0	1
1	7
2	80
3	206
4	323
5	249
6	122
7	51
8	27
9	10

Table 14: The number (n) events with maximum *Kp* from 0 to 9.

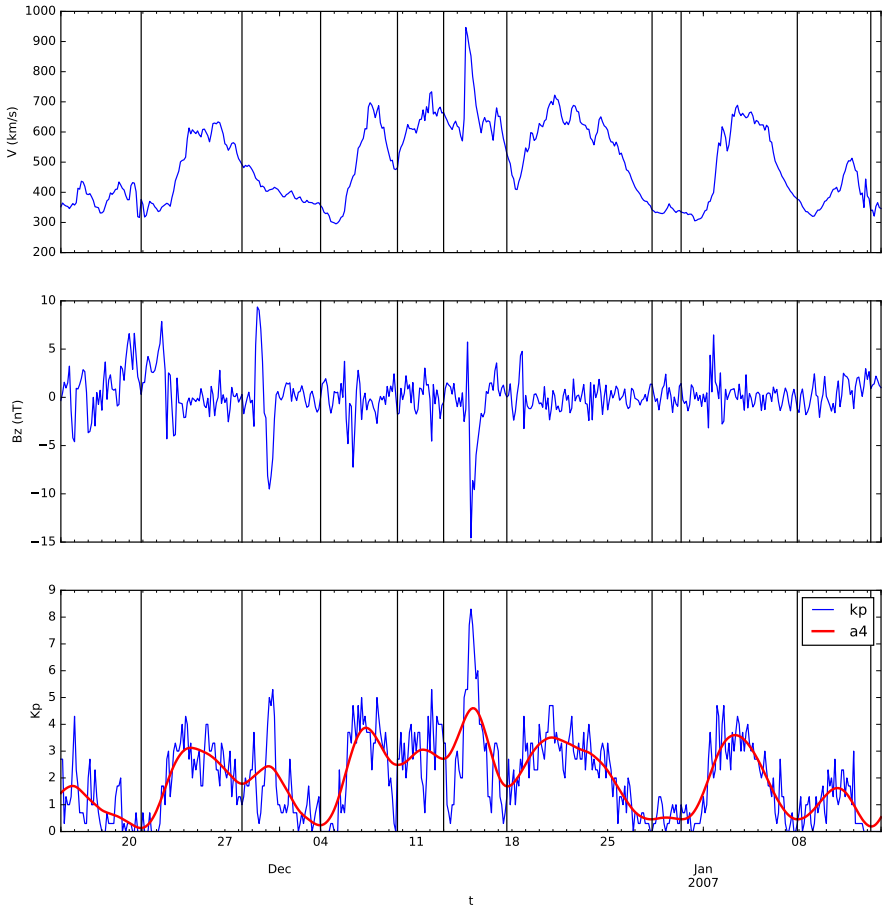

Figure 13: Example of identified times of minima (vertical lines) for filtered Kp (red curve).

For each event the maximum observed and predicted Kp are shown in Figure 14 for the LUND model and a simple 27-day recurrence relation. The predicted Kp maxima follows the observed well over the complete Kp range. There are three outliers (encircled) for large Kp where two of them are due to solar wind datagaps. The third might be due to that the solar wind B_z component turned strongly negative but only for a very short time.

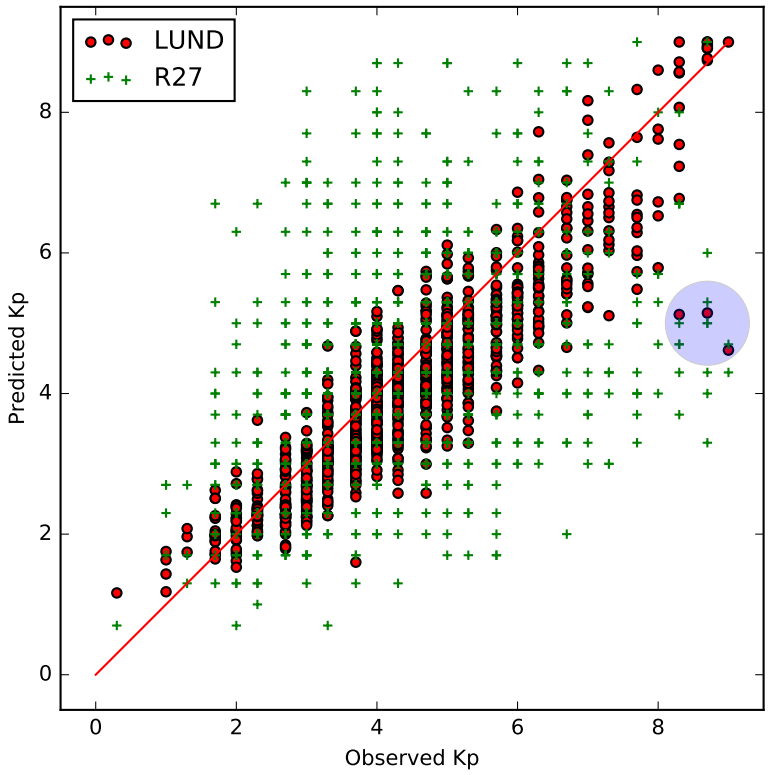

Figure 14: Scatter plot of the observed and predicted maxima Kp for each event.

The error box plots are shown in Figure 15 for the LUND model and the 13- and 27-day recurrence relations for comparison. The 5 and 95 percentile marks show that 90% of the predictions are within ± 1 units from observed Kp. All predictions are within ± 2 units from observed when the three events mentioned above are excluded.

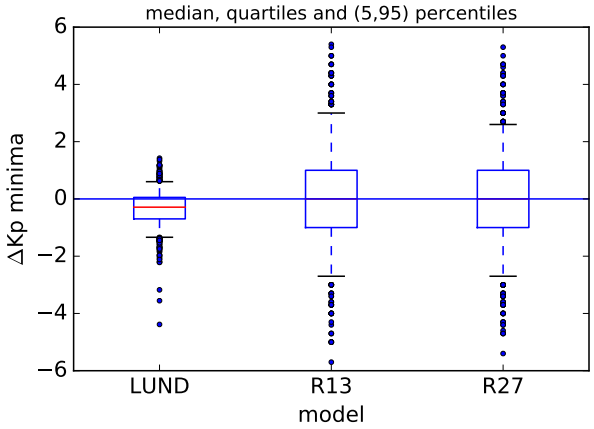

Figure 15: Box plot of errors between observed Kp maxima and predicted Kp maxima.

4 Conclusions

We have applied standard verification measures and visualisations to predicted *Dst* and *Kp*. We suggest that the verification approach should be slightly different for models driven by solar wind measurements compared to models driven by predicted solar wind data. In the first case, timeliness is very important as the prediction lead time is very short. We therefore compute the measures on different phases of the storms. For *Dst* we used a wavelet filter to automatically determine storm onset and duration, while for *Kp* we simply looked at *Kp* increases and decreases. As most measures will give good scores for persistence models for the short lead time forecasts it is important to detect it when comparing measures and trying to rank models.

When the solar wind-index models are driven by predicted solar wind it is more interesting to study the performance on an event basis. The same measures can be applied but now persistence is not applicable. The event selection algorithm for *Dst* is also used and the algorithm is also applied to *Kp*. In addition to the mentioned measures the minimum *Dst* and maximim *Kp* for each event is analysed.

As the models verified here are computationally very lightweight it is possible to run the models for very long sequences, in this work for all data since 1998. Both the Dst-LUND (IRF-Lund) and the Dst-SN_1 (Univ. Sheffield) *Dst* models perform similarly with the largest differences for the larger events. The Dst-LUND model saturates at -230 nT. It should be noted that the model was develop on data from before 2001 and has not been changed since then. During that period there are very few *Dst* events below -250 nT with simultaneous solar wind data which may explain why it saturates at -230 nT. The Dst-SN_1 model shows a bigger range of variability for the larger events, some predictions come close to the observed while other both underestimates and overestimates. However, having a set of models that perform similarly it is quite difficult to determine a best model, it instead becomes a "beauty competition" (Déqué 2012).

Currently we only have verified the Lund-*Kp* model, but the Sheffield *Kp* model will be included in the analysis in the future. The Lund-*Kp* models are implemented both for nowcast (NC) and 3-hour forecast (FC). However, from physical considerations it is difficult to imagine how a three-hour forecast would be possible, and this also partly shows up in the analysis. From the event based analysis it is seen that the LUND model captures the full range of *Kp* maxima.

The verification approach described here will be implemented on the future model development within this project. We expect that there will be some improvements in the future *Dst* and *Kp* models, and this improvement can be checked against current models using the approaches described here. Also the future *AE* models will be analysed similarly.

References

Boberg, F., Wintoft, P. & Lundstedt, H. (2000), 'Real time Kp predictions from solar wind data using neural networks', *Physics and Chemistry of the Earth, Part C: Solar, Terrestrial & Planetary Science* **25**, 275–280.

Boynton, R. J., Balikhin, M. A., Billings, S. A., Sharma, A. S. & Amariutei, O. A. (2011), 'Data derived narmax dst model', *Annales Geophysicae* **29**, 965–971.

Burton, R. K., McPherron, R. L. & Russell, C. T. (1975), 'An empirical relationship between interplanetary conditions and dst', *Journal of Geophysical Research* **80**(31), 4204–4214.

Déqué (2012), *Forecast verification: A practioner's guide in atmospheric science*, 2nd edn, John Wiley and Sons Ltd, chapter Deterministic forecasts of continuous variables, pp. 77–94.

Echer, E., Gonzalez, W. D., Tsurutani, B. T. & Gonzalez, A. L. C. (2008), 'Interplanetary conditions causing intense geomagnetic storms (dst $\leq$ -100 nt) during solar cycle 23 (1996–2006)', *Journal of Geophysical Research* **113**, A05221.

Ji, E. Y., Moon, Y. J., Gopalswamy, N. & Lee, D. H. (2012), 'Comparison of dst forecast models for intense geomagnetic storms', *Journal of Geophysical Research* **117**, A03209.

Lundstedt, H., Gleisner, H. & Wintoft, P. (2001), 'Operational forecasts of the geomagnetic *Dst* index', *Geophysical Research Letters* **106**(A6), 1–4.

Mayaud, P. N. (1980), *Derivation, meaning, and use of geomagnetic indices*, Vol. 22 of *Geophysical monograph*, American Geophysical Union.

Murphy, A. H. (1993), 'What is a good forecast? An essay on the nature of goodness in weather forecasting', *American Meteorological Society* **8**, 281–293.

O'Brien, T. P. & McPherron, R. L. (2000), 'Forecasting the ring current dst in real time', *Journal of Atmospheric and Solar-Terrestrial Physics* **62**, 1295–1299.

Percival, D. B. & Walden, A. T. (2000), *Wavelet methods for time series analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK.

Rastätter, L., Kuznetsova, M., Glocer, A., Welling, D., Meng, X., Raeder, J., Wiltberger, M., Jordanova, V. K., Yu, Y., Zaharia, S., Weigel, R. S., Sazykin, S., Boynton, R., Wei, H., Eccles, V., Horton, W., Mays, M. L. & Gannon, J. (2013), 'Geospace environment modeling 2008–2009 challenge: Dst index', *Space Weather* **11**, 187–205.