AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE

10.1002/2014SW001051

Key Points:
- Solar wind speed prediction up to 5 days
- Probability distribution functions of the solar wind velocity
- Periodicity of the solar wind velocity related to the rotation of the Sun

Correspondence to:
C. D. Bussy-Virat,
cbv@umich.edu

Citation:
Bussy-Virat, C. D., and A. J. Ridley (2014), Predictions of the solar wind speed by the probability distribution function model, *Space Weather*, *12*, 337–353, doi:10.1002/2014SW001051.

Received 7 FEB 2014
Accepted 7 MAY 2014
Accepted article online 11 MAY 2014
Published online 2 JUN 2014

Predictions of the solar wind speed by the probability distribution function model

C. D. Bussy-Virat[1] and A. J. Ridley[1]

[1] Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, Michigan, USA

Abstract The near-Earth space environment is strongly driven by the solar wind and interplanetary magnetic field. This study presents a model for predicting the solar wind speed up to 5 days in advance. Probability distribution functions (PDFs) were created that relate the current solar wind speed and slope to the future solar wind speed, as well as the solar wind speed to the solar wind speed one solar rotation in the future. It was found that a major limitation of this type of technique is that the solar wind periodicity is close to 27 days but can be from about 22 to 32 days. Further, the optimum lag between two solar rotations can change from day to day, making a prediction of the future solar wind speed based solely on the solar wind speed approximately 27 days ago quite difficult. It was found that using a linear combination of the solar wind speed one solar rotation ago and a prediction of the solar wind speed based on the current speed and slope is optimal. The linear weights change as a function of the prediction horizon, with shorter prediction times putting more weight on the prediction based on the current solar wind speed and the longer prediction times based on an even spread between the two. For all prediction horizons from 8 h up to 120 h, the PDF Model is shown to be better than using the current solar wind speed (i.e., persistence), and better than the Wang-Sheeley-Arge Model for prediction horizons of 24 h.

1. Introduction

One of the basic problems with space weather prediction is that the near-Earth space environment is strongly driven, and the drivers can only be measured about 1 h before they affect the environment. In order to allow for adequate planning for some members of the commercial, military, or civilian communities, reliable long-term space weather forecasts are needed [*Wright et al.*, 1995]. The goal of this study is to move the community beyond the current 1 h predictive horizon based on ACE measurements of the solar wind by utilizing intrinsic periodicities of the solar wind speed, statistics, ensemble modeling, and probabilistic forecasting. The solar wind speed is the focus here due to its periodic nature. Future studies will focus on other geospace drivers such as the solar irradiance and interplanetary magnetic field components B_x and B_y, which are also periodic. The component B_z, which is much more challenging to model, will be studied after these.

Many predictive empirical and physics-based solar wind speed models have been created, most of them coronal or heliospheric. One of the most commonly used is the Wang-Sheeley-Arge (WSA) Model, currently used at the Space Weather Prediction Center of the National Oceanic and Atmospheric Administration. *Wang and Sheeley* [1990, 1992] found a correlation between the solar wind speed and the inverse of the divergence rate of the magnetic field in the corona. Based on this relationship, they created a model to predict the solar wind speed by extrapolating observations of the photospheric magnetic field into the corona. *Arge and Pizzo* [2000] improved and tested its performance by comparing the predictions to data collected by the Wind satellite for 3 years (1994–1997), in particular for the 1996 solar minimum. The average fractional deviation and the correlation were found to be equal to 0.15 and 0.4, respectively. *Arge et al.* [2003] proposed a new relationship for the solar wind speed as a function of the magnetic field expansion factor of open coronal field lines and of the minimum angular separation at the photosphere between an open field foot point and its nearest coronal-hole boundary. Another longer-term validation study conducted by *Owens et al.* [2005] showed that the root-mean-square error between the model and the data is better at solar maximum than at solar minimum. However, the large-scale structure is better predicted during solar minimum than during solar maximum. The WSA Model has also been compared in previous studies to the Persistence Model (which keeps the velocity constant at its current value). *MacNeice* [2009a] validated the ability of the WSA Model to forecast specific solar events, while *MacNeice* [2009b] showed that the WSA Model provides better predictions than the Persistence Model clearly only after 2 days.

The Hakamada-Akasofu-Fry (HAF) Model (currently used at the Air Force Weather Agency) is a solar wind speed prediction model. *Fry et al.* [2007] presented results of the solar wind forecast based on this model. It particularly distinguished simulations of the ambient solar wind and simulations of event-driven solar wind (for example, coronal mass ejections, or CMEs). Predictions of shocks following solar events such as CMEs are detailed in *Fry et al.* [2001] and *Smith et al.* [2004]. *Norquist and Meeks* [2010] compared the predictions between the WSA and the HAF Models (5 day forecasts over 6 years of Solar Cycle 23). The WSA gave slightly better predictions than the HAF Model for the speed of the solar wind. For instance, correlations between the model and the data are slightly better for the WSA Model (0.42) than for the HAF Model (0.32). However, both models underrepresent the temporal variability of speed of the solar wind: the standard deviation is smaller in the models than it is in the data, by about 15%.

Toth and Odstrcil [1996] conducted a comparison of methods for simulations of MHD Models, one of which was the ENLIL Model, which enables numerical modeling of solar wind structures and disturbances in 3-D [*Odstrcil*, 2003]. An idea proposed in the past was to couple models. For instance, the ENLIL heliospheric model was also coupled to the Magnetohydrodynamics Around Sphere (MAS) coronal model (also called the CORHEL model) to predict solar wind parameters based on solar and coronal structures [*Odstrcil et al.*, 2004]. *Owens et al.* [2008] also made a comprehensive study of the coupling of kinematic, empirical, and MHD Models. The MAS/ENLIL and WSA/ENLIL Models were compared in *Lee et al.* [2009] to measurements from the ACE satellite during the declining phase of Solar Cycle 23. However, this was a study of a more scientific nature, and it did not test the forecasting ability. In particular, it revealed a good agreement for the general large-scale solar wind structures but not for the CME or shock associated with active regions.

Other methods based on observations of the chromosphere have been created. For instance, a hybrid intelligent system uses magnetic field observations and combines the potential field model and an artificial neural network to give prediction of the daily solar wind speed [*Wintoft and Lundstedt*, 1997]. The correlation between predictions from the model and the data between 1976 and 1994 varies from 0.2 to 0.5, depending on the year. *Robbins et al.* [2006] presented another model to predict solar wind speed related to geomagnetic events. The model was based on the location and the size of coronal holes. It differs from the WSA Model in particular because it does not need a full magnetic synoptic map but only the image of one coronal hole to give predictions. The linear correlation was 0.38 for the 11 years of comparison between the model and data. The Pch method [*Luo et al.*, 2008] was based on a correlation between the speed and the brightness of the solar EUV emission (characterizing the brightness and the area of a coronal hole). *Leamon and McIntosh* [2007] also presented predictions based on the structure of the chromosphere.

Another method, the support-vector-regression algorithm, was applied in *Liu et al.* [2007] to predict the value of solar wind speed. The comparison of the predictions to the data shows very good results, but the predictions are only 1 to 3 h ahead of real time.

Finally, *Owens et al.* [2013] made a comprehensive study of the 27 day periodicity of the solar wind parameters and presented a possible way to make predictions based on this periodicity. In particular, they showed that the correlation between two solar rotations is very good for the speed during solar minimum or the declining phase of a solar cycle. However, this is not the case during solar maximum, where no clear correlation was found. The study also explained how such a model can represent a benchmark for other space weather forecast models.

The goal of the present study is to describe another type of databased empirical model of the solar wind speed, based solely on observations of the solar wind: the Probability Distribution Function (PDF) Model. The ultimate goal is to use the general technique to also predict the interplanetary magnetic field (IMF) and the solar irradiance flux (e.g., $F_{10.7}$).

2. Methodology

There are two concepts that drive the PDF Model. The first is the idea that the solar drivers do not randomly change from hour to hour. For example, it has been shown that persistence is one of the best models of the solar wind speed for short-term predictions [e.g., *Norquist and Meeks*, 2010]. For the two first days of prediction, keeping the speed constant at its current value is one of the best estimators for the value of the

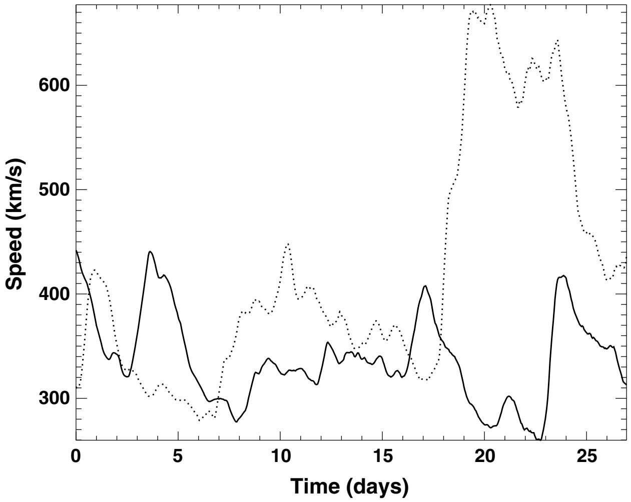

Figure 1. Example of the solar wind speed during two rotations that correspond to a bad correlation.

speed. This implies that the solar wind speed changes relatively slowly, such that, if the speed at the present time is known, then the speed in an hour from now will most likely be quite similar to the value now. This idea can be formalized into a probability-based model, where PDFs can be derived for the solar wind speed, given the current speed and the prediction horizon (each hour from the current time until 5 days into the future). Through exploring the behavior of the solar wind speed over times of many tens of hours, it was noted that the prediction can be separated into two different groups, differentiated by whether the speed had been increasing or decreasing over the previous 12 h.

The second idea of the PDF Model is based on the approximately 27 day rotation of the Sun. The rotation depends on the latitude: it is 25 days at the equator and 35 days at the poles. On average, it is about 27.2 days. Therefore, the same active regions of the Sun are directed toward the Earth every 27 days. This implies that a 27 day periodicity should also be observed in the solar wind speed, the IMF, and the solar radiation flux. A comprehensive study of this 27 day periodicity has been detailed by *Ram et al.* [2010]. That study highlighted a strong 27 day periodicity but also signatures of periodicities corresponding to 13.5, 9, and 6.8 days. Given the 27 day periodicity of the Sun, a model theoretically can be created in which the solar wind speed prediction can be based on the solar wind speed 27 days ago. PDFs can therefore be created given the present speed and the speed in approximately 27 days from now.

The biggest problem with this technique is that the solar structures are not perfectly periodic. Two consecutive solar rotations can lead to very different speeds of the solar wind, as shown in Figure 1. The Pearson correlation between these two speeds is 0.2 (all the correlation numbers that we present in this study correspond to Pearson correlations). The Pearson correlation, a number between −1 (anticorrelated) and 1 (correlated), quantifies the similarities between the shapes of two curves. The problem with a cross-correlation comparison is that the two series could be quite steady, with small variations on top of a large baseline, and the cross correlation could return a value of almost zero, even though the baselines are identical. This is because the cross correlation focuses only on the variations, which could be a minor part of the signal. With the solar wind speed, the baseline can sometimes be quite large, compared to the variations. A root-mean-squared (RMS) difference can also be used to quantify the differences between two series. For an RMS, if baselines are similar, even though the (small) variations are different, the result will be close to 0. A normalized root-mean-squared difference divides the RMS by the mean value of one of the series, so that one can judge the relative difference between the series. For example, an N-RMS of 0 still implies perfect agreement, but an N-RMS of 0.8 (which is the value calculated for the two speeds in Figure 1) implies that one series is different from the other by an offset of 80%. The mathematical expression is

$$\mathrm{N-RMS} = \sqrt{\frac{\overline{(u-v)^2}}{\overline{u^2}}} \tag{1}$$

where u and v are two different series and the symbol $\overline{u}$ represents the mean of the series u.

It was often found that the lag between consecutive solar rotations is not exactly 27 days. Indeed, this is often the case. The optimum lag could vary between approximately 22 and 32 days ago. Table 1 shows that when the lag was forced to exactly 27 days, the correlation and N-RMS are relatively poor. However, when the lag

Table 1. The Mean Correlation and N-RMS Between Approximately 6000 Consecutive Solar Rotations With an Absolute 27 Day Lag and With Allowing the Lag to be Adjusted Until an Optimum Is Found

Condition	Correlation	N-RMS
27 day	0.22	0.19
27±5 day	0.89	0.09

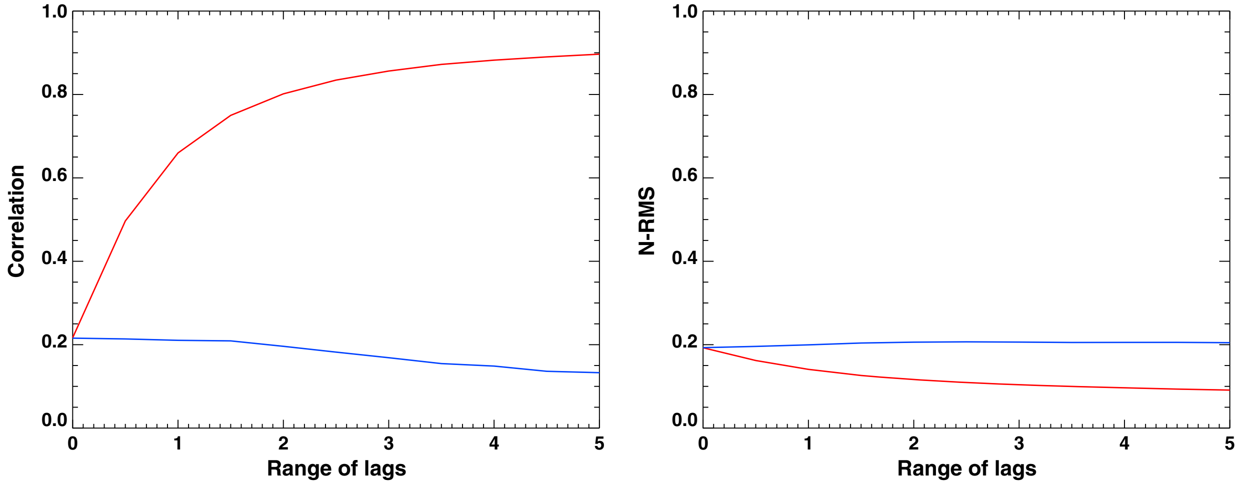

Figure 2. Dropoff in (left) the correlation and (right) the N-RMS of the predictions. With a correlation, values close to 1 are better, while with an N-RMS, values close to 0 are best.

was allowed to adjust to optimize the correlation or the N-RMS, these values were dramatically improved. The correlation and N-RMS were calculated for six thousand 5 day periods of solar wind speed during 1995–2011.

The general idea for forecasting the solar wind speed *i* hours into the future would be to use the data from one solar rotation ago (OSRA), or 27 days plus an optimum lag ago (denoted $t_{\text{OSRA}} = t_{\text{now}} - 27$ days + lag). This optimum lag could be found using the last, for example, 3 days of solar wind data, compared to the solar wind speed approximately 27 to 30 days ago. The expectation was that using the lag t_{OSRA}, the predictions of the solar wind speed would improve. Instead, the opposite was found. Using the lag of exactly 27 days ($t_{27} = t_{\text{now}} - 27$ days) produced the best results. Figure 2 illustrates this. The red lines show the correlation (left) and N-RMS (right) between the 3 day period just before the current time and the period 0–3 days before t_{OSRA} days ago, where the lag was allowed to vary to optimize the result up to the value on the *x* axis. For example, at an *x* axis value of 1 day, the lag was allowed to shift to any value between ±1 day (for a total shift of between 26 and 28 days). For a value of 4 days, the lag was allowed to shift to any value between ±4 days (for a total shift of between 23 and 31 days). With the knowledge that the lag is not exactly 27 days, but a bit more or less than this value, the red lines show the expected behavior: as the window size of comparison is opened more and more, the correlation and N-RMS improve. The values of the red lines at the 0 and 5 day marks are shown in Table 1. The blue lines are the comparisons between the subsequent 3 days, using the optimized lag determined for the past 3 days. At lag = 0, the red and blue lines have the same values, since t_{27} was used for all times compared. When the lag was allowed to increase, the comparisons for the subsequent days become worse. This means that, statistically, the optimum lag for the last 3 days is not the optimum lag for the next 3 days. The optimized lag changes rapidly, which makes prediction of the solar wind speed using the previous solar rotation extremely difficult.

This is the main issue that the PDF Model has to face. It implies that if the optimum lag cannot be predicted, then the method cannot use the past solar rotation in an optimized way for the prediction of the following 5 days. At this time, it is unknown how to predict the optimum lag. Currently, as described below, the lag is calculated by using a temporally weighted N-RMS comparison, where the current time is weighted the strongest and the data from 3 days ago is weighted the least. This allows the delay to be optimized for the time now, instead of an average of the last 3 days. Attempting to determine this lag will be the focus of future research and should greatly improve the forecast ability of this type of model. With the current PDF Model, there is a greater reliance on the PDFs based on the current solar wind speed and trend.

2.1. Construction of the Probability Distribution Functions

The first set of PDFs are simply the distribution of speeds each hour for the next 120 h, given the speed now and the slope of the 12 preceding hours. For example, if the current solar wind speed is 450 ± 10 km/s and if the mean of the speed of the 12 past hours is greater than the speed now (meaning that the speeds are

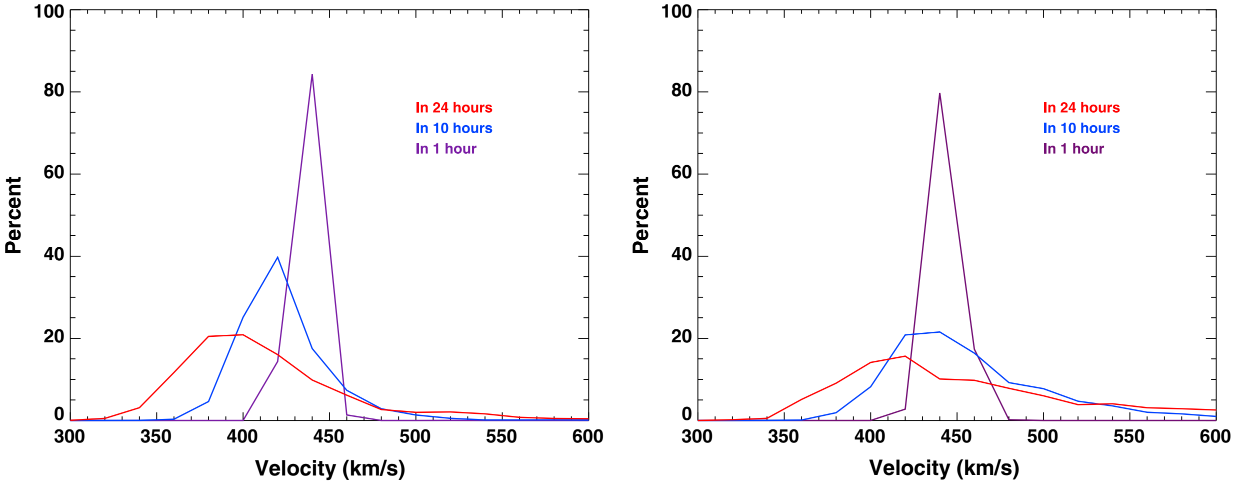

Figure 3. Examples of three probability distribution functions (1, 10, and 24 h from the current time) from P1 based on the current speed being 450 km/s and the speed (left) decreasing and (right) increasing over the last 12 h.

decreasing), then the distribution of speeds for 1, 10, and 24 h after the current time is given by Figure 3 (left). Figure 3 (right) shows the same thing but for increasing solar wind speed. The most important thing to notice about the progression of PDFs is that the peak decreases in intensity and that the width of the distribution increases with time. This means that, as the forecast horizon becomes longer, there is a larger number of speeds that could happen and that the percentage of likeliness that the most probable speed will happen decreases. Furthermore, the peaks of the PDFs in Figure 3 (left) move toward slower speeds. This is consistent with the fact that the speed is decreasing. In Figure 3 (right), the peaks in the PDFs also decrease in speed, but the tails at high solar wind speeds become larger, which shows that there is a population of speeds that can be significantly larger than the current 450 km/s speed. These PDFs can be compared to a PDF of all of the solar wind speeds over the entire model time period, shown in Figure 4 (the analysis includes the CMEs that occurred during the time period). While these plots are on different scales, it can be seen that the global solar wind distribution peaks at about 9.5%, while the 24 h PDF peaks at approximately 20% for the decreasing-speed case (Figure 3, left) and 15% for the increasing-speed case (Figure 3, right). This indicates that the 24 h PDF is still more useful for predictions than simply using a PDF of all of the solar wind.

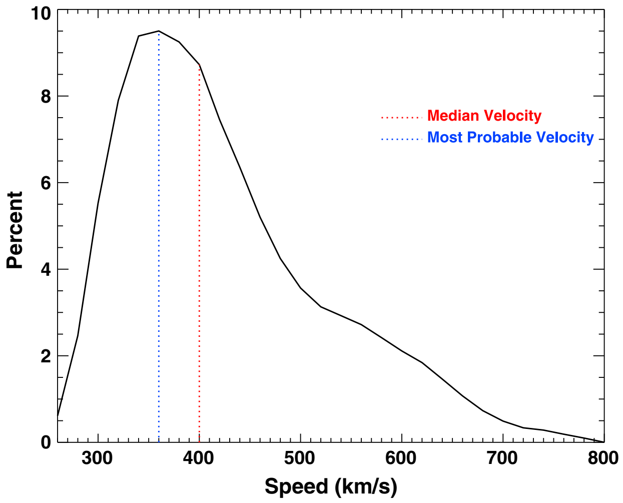

Figure 4. The probability distribution function of the solar wind over 1995–2011 from 260 to 800 km/s, with the median and most probable values indicated.

The PDFs were created for bins from 260 to 800 km/s with bin sizes of 20 km/s (i.e., bin center ±10 km/s). The prediction horizon stretches from 1 to 120 h, and the data are separated into increasing or decreasing solar wind speeds. This set of PDFs will be called P1. Figure 5 gives examples of predictions made using only the P1 PDFs, given the current solar wind speed (Time = 0) and the direction of the slope of the solar wind speed over the previous 12 h. The top row and the left plot in the middle row show examples in which the solar wind speed was decreasing, while the right plot in the middle row and the

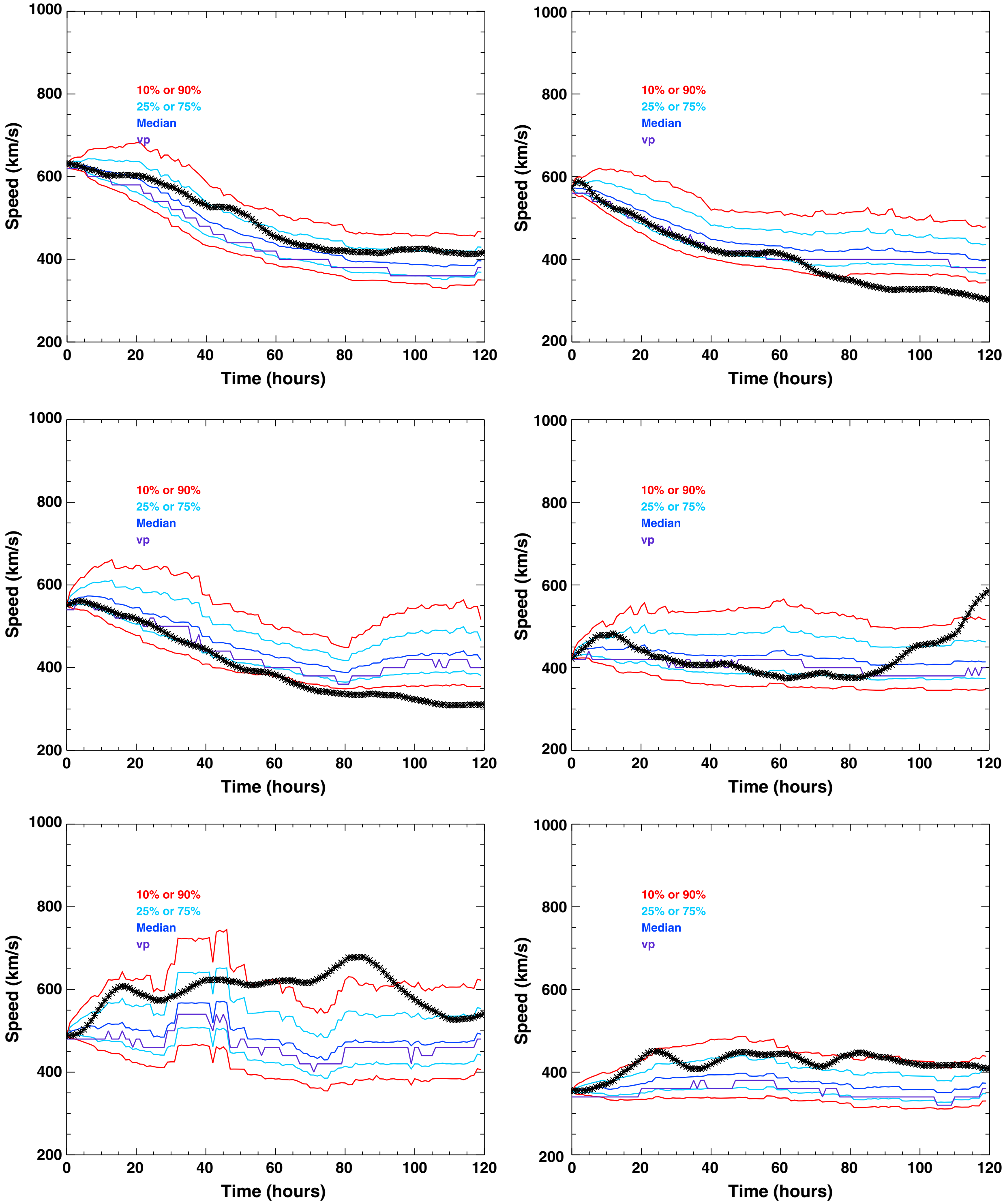

Figure 5. (top to bottom) Six examples of solar wind speed predictions out to 120 h using the P1 probability distribution functions. In these cases, the actual solar wind data are shown as black stars. The predictions are indicated by the colored lines, with the most probable and median values indicated by the purple and dark blue lines. The light blue lines indicate the 25th and 75th percentiles, while the red lines indicate the 10th and 90th percentiles.

bottom plots show times in which the speed was increasing. The majority of the time, the predicted speed is within the red curves, which indicate the 10% and 90% levels of the PDF. The solar wind-decreasing cases appear to be better predicted than the solar wind-increasing cases, which is to be expected, given that the solar wind-decreasing cases have narrower and taller PDFs.

These plots illustrate one of the advantages of using PDFs to determine the predictions—a range (or uncertainties) of the predictions can be determined. Furthermore, an ensemble of different solar wind speed prediction scenarios could be generated to allow for ensemble forecasting of the near-Earth space environment. Using either of these allows the forecaster to have more information about the prediction than simply the value and the past performance.

From the P1 PDFs, the median solar wind speed (50% of the speeds are below/above this speed) and the most probable speed (speed that corresponds to the peak in the PDF, which is typically a bit lower than the median speed), as demonstrated in Figure 4, can be determined. The median speed (M1) and the most probable speed (VP1), as determined by the set of PDFs P1, can be used to generate a typical single-value predictive model of the solar wind speed.

The second set of PDFs will be called P2. They are based on the 27 day periodicity of the solar wind speed. For each hour of solar wind data from 1995 to 2011, the optimum lag was found to the following solar rotation. An N-RMS comparison with the previous 5 days was used to determine the optimum lag. The solar wind speed at t_{OSRA} was then noted. PDFs from 260 to 800 km/s with bin sizes of 20 km/s (i.e., bin center ±10 km/s) were created. The change of speed was not included in this case nor were the predictions beyond the value at the optimum lag time. Therefore, there are significantly fewer PDFs in this set. As with PDF P1, the median speeds (M2) and the most probable speeds (VP2) were determined from PDFs in P2.

To summarize, if a prediction of the solar wind speed in i hours is desired, two PDFs are available. P1 gives directly the distribution of speeds in i hours, given the current speed and the trend in speed. In order to use P2, a lag needs to be determined from "exactly" one solar rotation ago (OSRA), or approximately 27 days $- i$ hours ago. Mathematically, this can be thought of as $t_{\text{OSRA},i} = t_{\text{now}} - 27$ days + lag + i hours. The solar wind speed at $t_{\text{OSRA},i}$ determines which PDF P2 should be used to predict the solar wind speed in i hours from now. A slightly different set of PDFs (P2_{27}) can be created simply based on a 27 day delay, instead of finding the optimized delay. This time can be referred to mathematically as $t_{27,i} = t_{\text{now}} - 27$ days + i hours.

2.2. Construction of the Premodels

We introduce the following notations:

$v_{\text{pred},i}$: the speed to be predicted in i hours from the current time;
$v_{\text{pdf},i}$: the speed determined by one of the PDFs for i hours from the current time (or, more specifically, $v_{P1,i}$ and $v_{P2,i}$);
$v_{\text{OSRA},i}$: the actual speed one solar rotation ago (optimized), plus i hours; and
$v_{27,i}$: the actual speed 27 days ago, plus i hours.

An important point is that $v_{\text{OSRA},i}$ and $v_{27,i}$ are actual solar wind speeds and are not derived from PDFs, with the difference being the use of an optimized lag ($v_{\text{OSRA},i}$) or an exact lag of 27 days ($v_{27,i}$).

A simple way to take into account the two speeds $v_{\text{pdf},i}$ ($v_{P1,i}$ or $v_{P2,i}$) and $v_{\text{OSRA},i}$ (or $v_{27,i}$) in the model is to use the expression

$$v_{\text{pred},i} = a \times v_{\text{pdf},i} + b \times v_{\text{OSRA},i} \tag{2}$$

where a and b are parameters that can varied (with $a + b = 1$) to optimize the prediction ability. This model would select a speed from a PDF and an actual speed from approximately 27 days ago in an optimized way.

To summarize, there are four different decisions that can be made to build the model: (1) whether the PDF that is used is based on the current solar wind speed (P1) or the solar wind speed approximately 27 days ago (P2); (2) whether the median or most probable value from the PDFs is used; (3) whether the solar-wind speed from the previous rotation is determined using an optimized lag or not; and (4) the value of a and b in equation (2), which determines the reliance on either the predicted speed based on the PDFs ($a = 1$) or the actual speed approximately 27 days ago ($b = 1$).

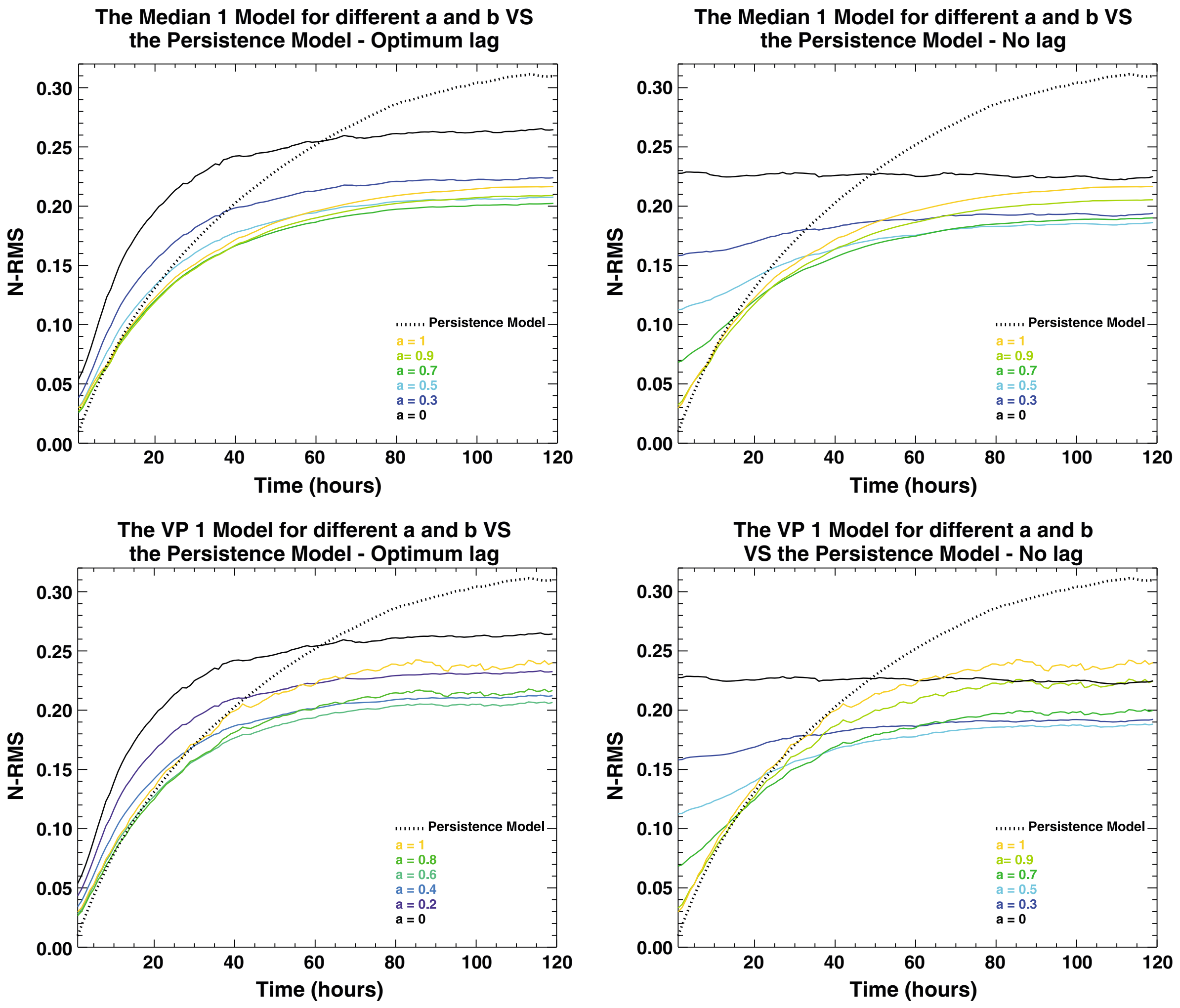

Figure 6. The different premodels based on a linear combination of the P1 PDFs and the solar wind speed one solar rotation ago compared to the Persistence Model.

For example, to predict $v_{\text{pred},i}$, the following parameters could be used:

1. $v_{\text{PDF},i}$ is $v_{P1,i}$;
2. The median of $v_{P1,i}$ is used;
3. $v_{\text{OSRA},i}$ uses an optimized lag; and
4. $a = 0.7$ and $b = 0.3$.

The Median 1 Model is the premodel that uses the median speed determined by P1; the VP 1 Model uses the most probable speed determined by P1; the Median 2 Model uses the median speed determined by P2; and the VP 2 Model uses the most probable speed determined by P2.

Finally, two simple models are presented for comparison: the Persistence Model, which predicts that the solar wind speed will be constant over the next 5 days at the current value; and the OSRA Model, which simply predicts the solar wind speed for the next 5 days will be exactly as it was 27–22 days ago. The OSRA Model is equivalent to $b = 1$ and not using an optimized lag above.

3. Results and Discussion

3.1. Results of the Different Premodels

For each of the premodels and models, the difference was calculated between the actual speed and predicted speed as a function of the prediction horizon (i) and model: $\Delta V_{t,i}^2 = \frac{(v_{\text{pred},i}-v_{\text{data},i})^2}{v_{\text{data},i}^2}$, where $v_{\text{data},i}$ was the real solar wind speed and $v_{\text{pred},i}$ was the prediction, both of which were i hours from the "current" time. The $\Delta V_{t,i}$ was calculated for every hour between the current time and 120 h into the future (i.e., i=1 to 120), and

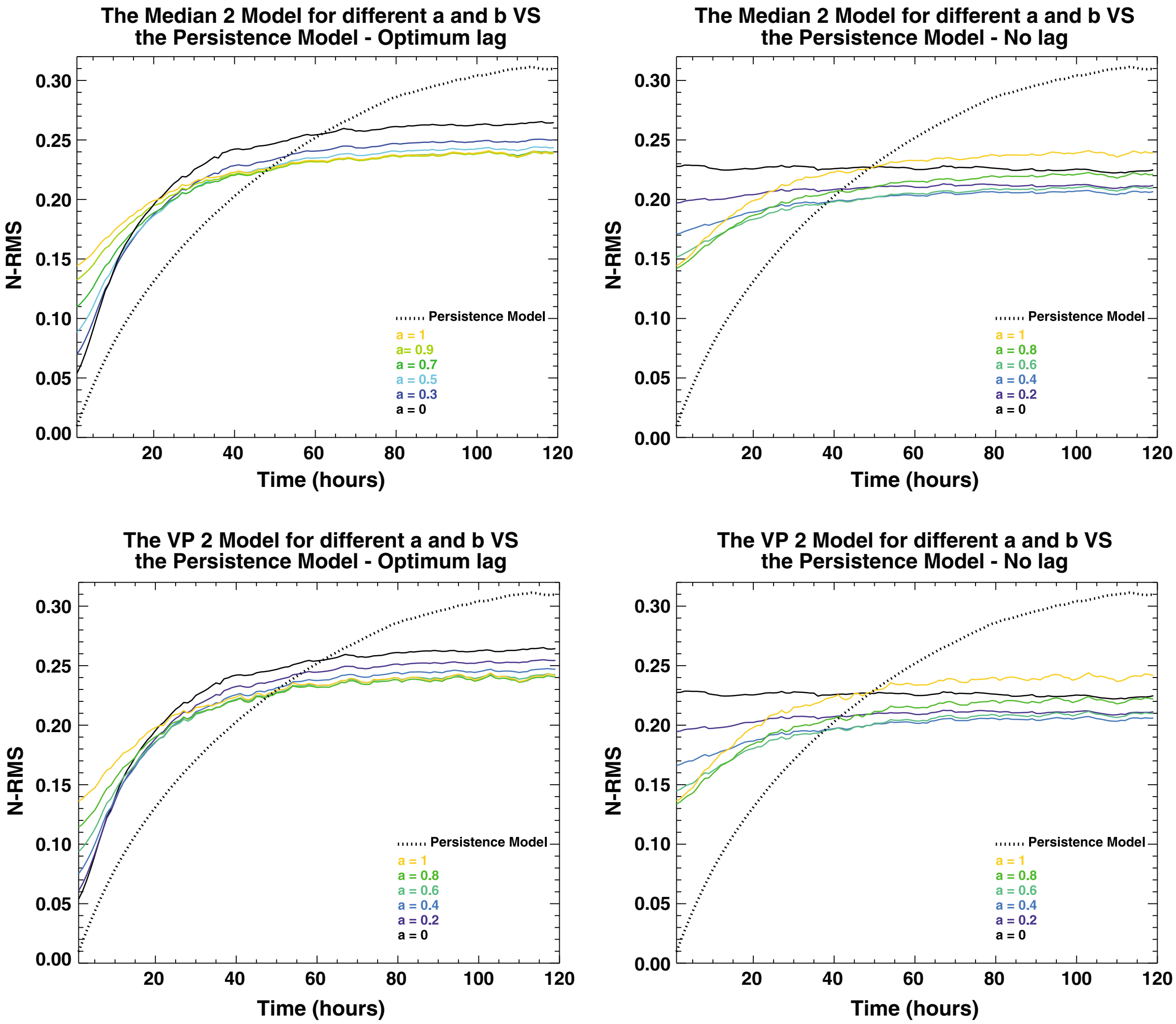

Figure 7. The different premodels based on a linear combination of the P2 PDFs and the solar wind speed one solar rotation ago compared to the Persistence Model.

was done for 6000 different times between 1995 and 2010 (each time moving 1 day forward, t = 1995 to 2010). The square root of the mean over t (the 6000 times) of $\Delta V_{t,i}$ was calculated, to get

$$\mathrm{N-RMS}_{\mathrm{model},i} = \sqrt{\overline{\Delta V_{t,i}}}. \tag{3}$$

This is not exactly the same expression as the N-RMS as described by equation (1), since the normalization is done before the mean as opposed to after the mean.

As described above, the model can be either one of the premodels (Median 1, VP 1, Median 2, or VP 2), the Persistence Model, or the OSRA Model. The premodels can be combined with the solar wind speed from one solar rotation ago through the use of the parameters a and b, which can vary between 0 and 1 (with $a + b = 1$). When combining the premodels with the solar wind speed from the last solar rotation, an optimum lag can be included or a lag of exactly 27 days can be used. When $a = 1$ and $b = 0$, the prediction is based solely on the PDFs—either the PDFs given the solar wind speed now and the trend (P1) or the PDFs based on the solar wind one solar rotation ago (P2). When $a = 0$ and $b = 1$, the prediction is based solely on the actual solar wind speed from (approximately, depending on the method) 27 days ago. Note that if no optimized lag is used and $b = 1$, the OSRA Model is derived. When the values of a and b are between 0 and 1, there is a blending of the techniques.

Figures 6 and 7 show the $\mathrm{N-RMS}_{\mathrm{model},i}$ for many different models with different a and b values. The $\mathrm{N-RMS}_{\mathrm{model},i}$ for the Persistence Model is also indicated as a dotted line on each plot. On the left side are the

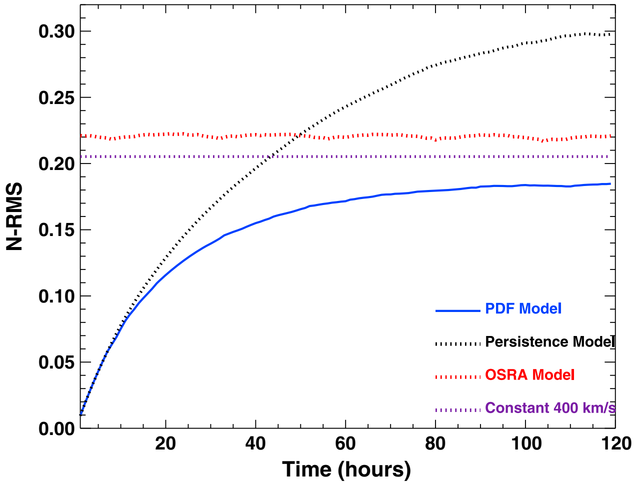

Figure 8. A normalized root-mean-squared comparison between the actual solar wind speed and the PDF Model, the Persistence Model, the OSRA Model, and a constant solar wind value of 400 km/s over the 16 years of this study.

models with an optimum lag included in $v_{\mathrm{OSRA},i}$, and on the right side are the models without an optimum lag included in $v_{27,i}$.

The first observation that can be made is that all of the models tend to have a lower N-RMS$_{\mathrm{model},i}$ value than the Persistence Model after approximately 1 to 3 days. The cases on the right (exactly 27 day lag) have models (a close to 0, or black and blue lines) that have quite poor results for low prediction horizons. This is because the lag might often be far from a perfect 27 days, such that taking the value exactly 27 days ago is often quite a poor choice. The models that use an optimized lag tend to perform better for shorter prediction horizons, but the models with an exactly 27 day lag tend to perform better for longer prediction horizons, as evidenced by the black line being lower in the right plots than in the left plots. This is because the optimized lag calculation is good for the present time but rapidly becomes useless, as described above and in Figure 2.

Additionally, independent of the lag or which speed is chosen (median versus most probable), neither the prediction based on the PDFs alone ($a = 1$) nor the prediction based on the previous solar rotation alone ($b = 1$) is optimum. It is a linear combination of the two that provides the best performance. Further, the optimum linear combination changes with time, such that low prediction horizons tend to be predicted best with a closer to 1, while later times tend to be predicted best with a and b close to 0.5.

A subtle feature to note is that the median values are better than the most probable values over the entire range of prediction horizons. There is almost always a consistent bias between the most probable value and the median value, with the median being larger, as indicated by Figure 4.

Finally, none of the models give better predictions than the Persistence Model for the first seven hours. The models that are almost as good as the Persistence Model are the Median 1 and the VP 1 Models (with an optimized lag) with a close to one. All these results mean that the PDF Model has to be a combination of the different premodels and the Persistence Model in order to have the best performance.

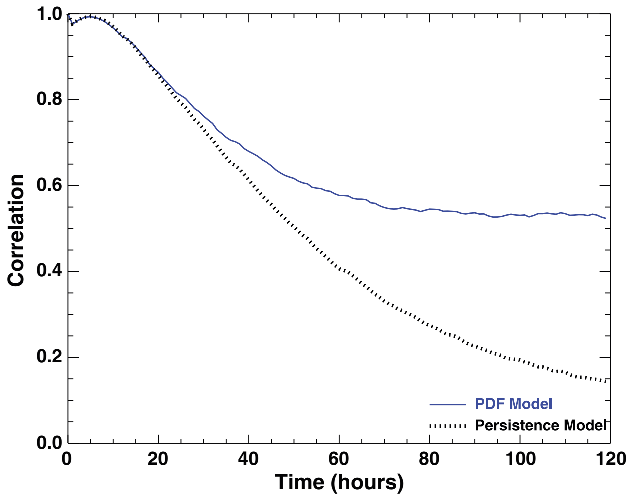

Figure 9. Correlation between the PDF Model and the observations (blue) and between the Persistence Model and the observations (black dash line) over the 16 years of this study.

3.2. The Completed PDF Model

As described above, the PDF Model has to be a combination of the different premodels and the Persistence Model. More precisely, the PDF Model predicts the speed in i hours using the premodels and the Persistence Model as follows: (a) for $1 \leq i \leq 7$ h, Persistence Model; (b) for $i = 8$ h, Median 1 with an optimized lag and $a = 0.9$; (c) for $9 \leq i \leq 12$, Median 1 with an optimized lag and $a = 0.8$; (d) for $13 \leq i \leq 17$, Median 1 with 27 day lag and $a = 0.9$; (e) for $18 \leq i \leq 32$, Median 1 with 27 day lag and $a = 0.8$; (f) for $33 \leq i \leq 51$, Median 1 with 27 day lag

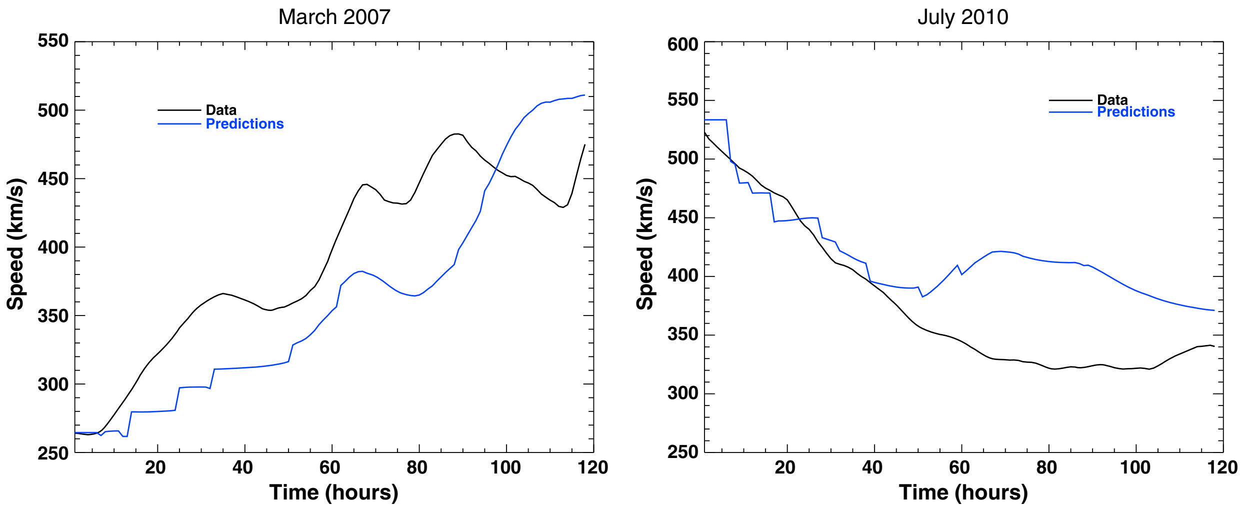

Figure 10. Two examples of solar wind predictions based on the PDF Model with the prediction in blue and the data in black.

and $a = 0.7$; (g) for $52 \leq i \leq 89$, Median 1 with 27 day lag and $a = 0.6$; and (h) for $90 \leq i \leq 120$, Median 1 with 27 day lag and $a = 0.5$.

Using this formulation, the N-RMS$_{\text{model},i}$ of the prediction is minimized. Figure 8 illustrates the N-RMS$_{\text{model},i}$ of the final PDF Model using this combination of the premodels and the Persistence Model, as well as comparing the PDF Model to the Persistence Model and the OSRA Model. After 7 h, the PDF Model gives better predictions than the Persistence Model. Additionally, the difference between the N-RMS of the PDF Model and the Persistence Model increases as the time goes by. After 2 days, the PDF Model gives considerably better predictions, and after 5 days the N-RMS$_{\text{model},i}$ is equal to 0.19 for the PDF Model and is equal to 0.30 for the Persistence Model. Further, the PDF Model is better than the OSRA Model all of the time. The Persistence and PDF Models are both better than the OSRA Model until about 45 h, at which time the OSRA Model becomes better than the Persistence Model. The OSRA and PDF Models converge to be within approximately 0.03 of each other by the 120 h prediction horizon, with the PDF Model being slightly better.

Figure 9 shows that the same trend is true for the cross correlation also. The correlation for the PDF Model is higher than for the Persistence Model after about 12 h all the way out to 120 h. The study by *Arge and Pizzo* [2000] showed that the correlation for the WSA Model for a 5 day prediction horizon was about 0.4, while Figure 9 shows that the PDF Model has a correlation of about 0.52 for the same prediction horizon, which is quite comparable.

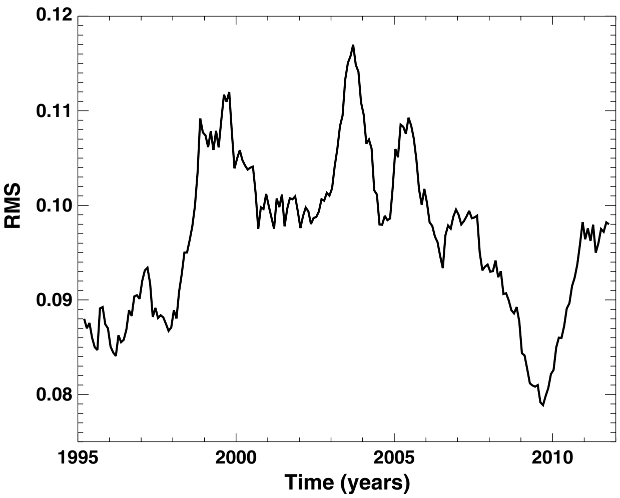

Figure 11. The N-RMS over the two last solar cycles for 24 h ahead predictions.

One fact that might be surprising in this study is that the set of PDFs P2 (i.e., PDFs based on the solar wind approximately 27 days ago) are not used in the final model. The main reason for this is illustrated in Figure 2: the determination of the lag is not optimal and is quite difficult to determine accurately. This means that we do not know exactly where to look back 27 days ago to find the time that will best match the solar wind speed in a few days from current time. Therefore, the speed that is used to determine which P2 PDF is not optimal. In other words, the inaccuracy in the lag results in using the wrong P2. If the time delay

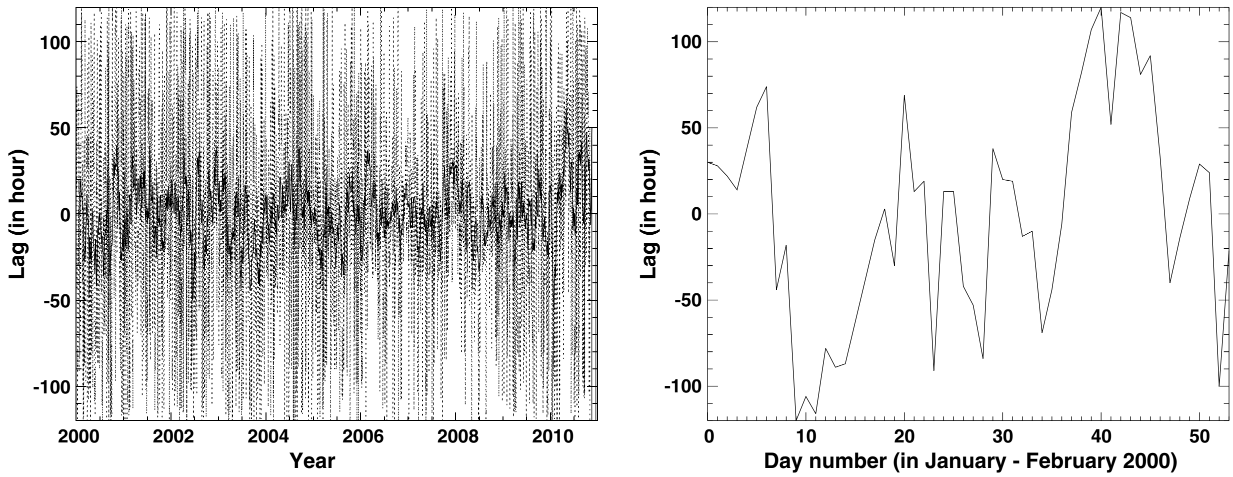

Figure 12. The optimum lag between solar rotations over 11 years (left) and over two solar rotations (right).

could be determined accurately, then the technique could go back in time 27 days plus the optimum lag, then move forward *i* hours and determine the right PDF P2 to use, considering the speed at that time. This bias, not present in the prediction made by the set of P1 PDFs, makes the predictions by P2 worse than the ones by P1.

Two examples of predictions using the PDF Model are provided in Figure 10. These two examples were chosen to show that even during a strongly increasing (left) or decreasing (right) solar wind speed, the predictions follow the data quite well. In the first case (left), the solar wind speed increases from about 260 km/s to 450 km/s over the 5 days. The PDF Model roughly captures the increase, although not the exact details of the smaller-scale variability in the speed. In the second case (right), the predictions are very close to the data for the two first days. However, important differences appear around 3 days and then decrease around 5 days. The solar wind speed observed by ACE has been averaged using a running average over 11 h to reflect the trend of the variation and filter the high-frequency variability of the solar wind speed.

Predicting the solar wind speed during solar maximum is more challenging than during solar minimum. Figure 11 shows the variation of the normalized difference in speeds between the PDF Model and the measurements as a function of time. As illustrated in this figure, the error in the 24 h ahead prediction was higher during the solar maximum in 2002 than it was during the two last solar minima, in 1996 and 2008.

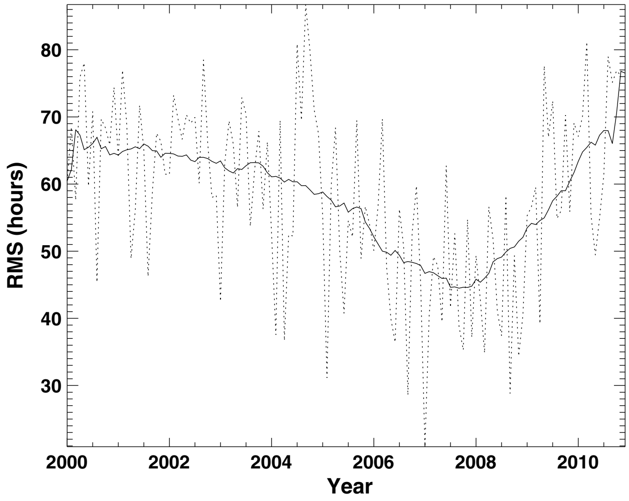

Figure 13. The dotted line shows the standard deviation of the lag during each month. The solid line shows a 13 month running average of the standard deviation, so the trend can be determined.

The fact that the errors in the predictions made by the PDF Model are higher during a solar maximum may be a consequence of two different phenomena: (1) as detailed in *Owens et al.* [2013], no clear correlation in speed between two solar rotations was found during solar maximum, whereas during solar minimum, the variability of the solar wind speed is much more periodic. This implies that using data from one solar rotation ago during solar maximum may not be the best idea. (2) During solar maximum, there are more impulsive events, which are not really accounted for in the PDF Model. These impulsive events, such as CMEs, make it so the current speed (preceding the CME) is not a good indicator of the future speed (during the CME). In addition, unless two CMEs occurred

27 days apart, the past solar rotation would not be a good indicator of the solar wind speed. This fact shows a weakness in the PDF Model: it does not have the ability to predict impulsive events, which is an active area of research.

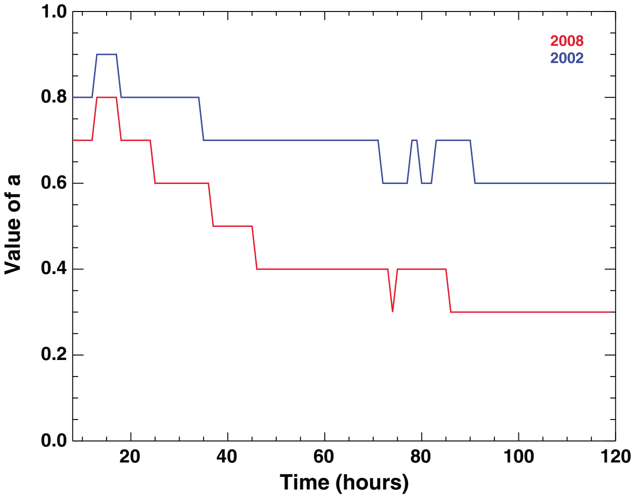

Figure 14. Coefficient *a* in the PDF Model during a solar maximum (blue) and during a solar minimum (red). Coefficient *b* is equal to 1 − *a*.

Figure 12 illustrates the variability of the lag (calculated with a 1 day step) over both a long (solar cycle) and a short (two solar rotations) time scales. There is a great deal of variation in the lag, even on a day-to-day basis. However, it is interesting to note that this variability increases at the solar maximum and decreases at the solar minimum, making the predictions based on the current lag easier during solar minimum (discussed below). Figure 13 quantifies this by showing the monthly standard deviation of the lag as a function of time. The standard deviation is shown to have a minimum in 2007–2008 and a maximum in 2000–2002.

Because the variability of the lag during a solar minimum seems to be less than during a solar maximum (Figure 13), the predictions based on the previous solar rotation should be more relevant during a quiet period of solar activity. This is verified in Figure 14 where *a* and *b* were calculated for a period corresponding to a solar maximum and a period corresponding to a solar minimum (only *a* is plotted, and *b* is simply 1 − *a*). Recall that *a* is the weight of the current solar wind speed, while *b* is the weight of the last solar rotation. One can notice that in 2008, the value of *a* is decreased for all prediction horizons (after 7 h since the seven first hours correspond to the Persistence Model), showing that using the last solar rotation provides a more accurate solution. However, *a* is increased in 2002, reflecting the high variability of the solar speed during such a period, and therefore predictions that have to rely mostly on the predictions based on the current solar wind speed.

While it is impossible to estimate the optimum lag at this time, if the optimum lag was able to be forecast, the predictions would dramatically improve. Figure 15 compares the N-RMS of predictions that use an optimum lag to the N-RMS of predictions made by the OSRA Model, with only a 27 day lag. The error is more than 4 times lower with an optimum lag. Moreover, the corresponding N-RMS is much lower than any other model: the difference between the prediction and the data is about 5%, which means that for a typical solar wind speed of 400 km/s, the error would be around 20 km/s only. This shows how important finding an optimum lag is and points to a need to have studies explore how to determine this optimum lag.

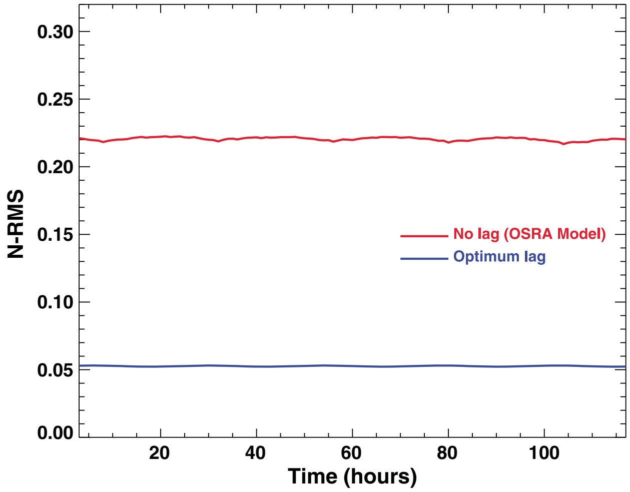

Figure 15. The N-RMS if the lag was optimum compared to the N-RMS without any lag.

One idea that is being explored for the next generation of the PDF Model will be to look at structures on the Sun, such as sunspots, and compare them to the same structures one solar rotation ago. This will enable us to find the lag so that the two sets of structures are superimposed on each other the best. This lag may be

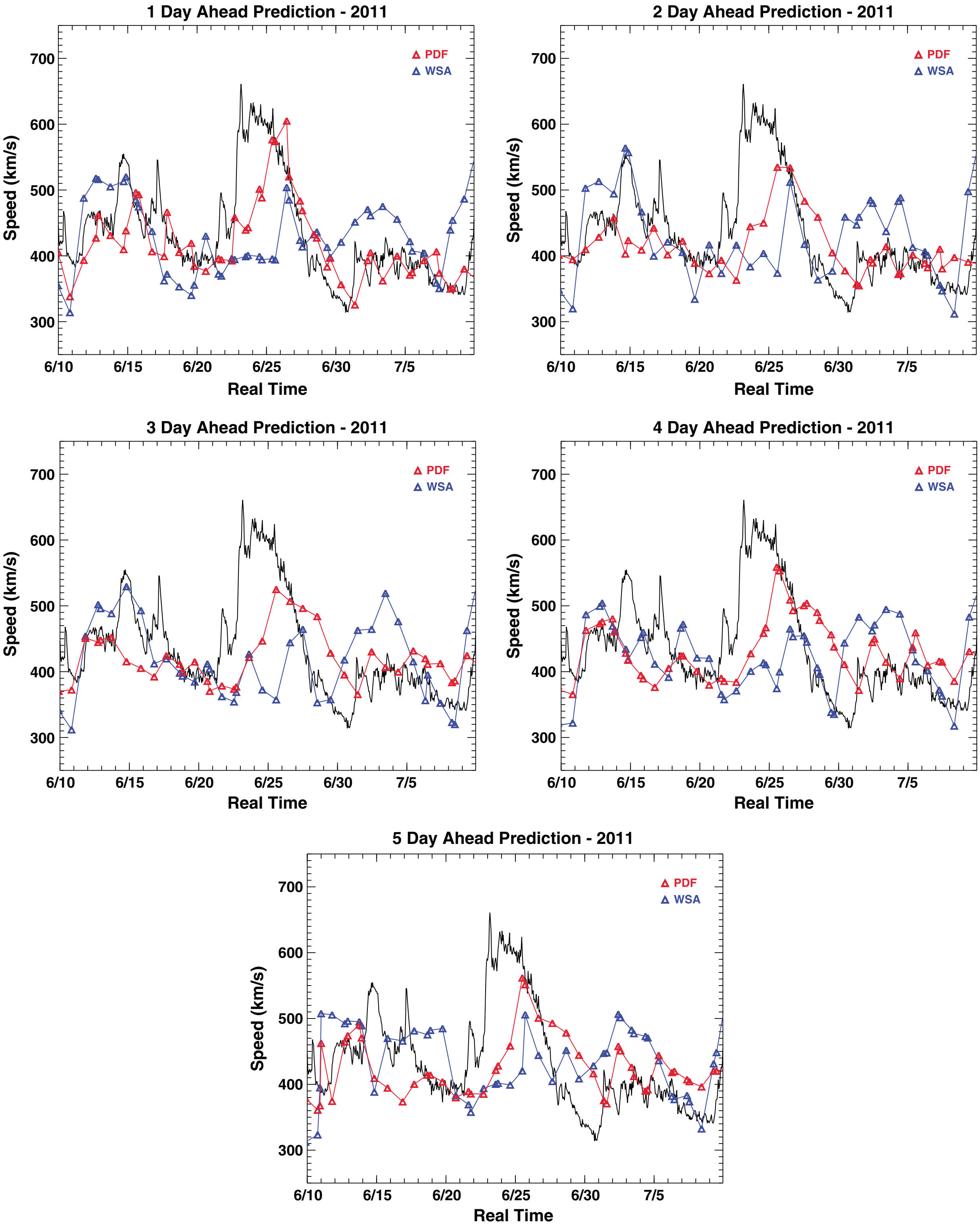

Figure 16. Example of predictions made by the PDF and the WSA Models for 10 June to 10 July 2011. The actual data are shown as a solid black line, while predictions (one or two per day) by the PDF and WSA Models are plotted in red and blue, respectively. One through 5 day ahead predictions are shown.

Table 2. The RMS Between the PDF, WSA, and Persistence Models and the Observations by ACE for 2008 and 2011

	RMS PDF (km/s)		RMS WSA (km/s)		RMS Persistence (km/s)	
Prediction Horizons	2008	2011	2008	2011	2008	2011
1 day ahead	83	66	107	88	79	68
2 day ahead	93	83	105	84	119	100
3 day ahead	90	89	103	88	147	118
4 day ahead	88	88	105	91	163	128
5 day ahead	90	88	107	90	173	130

the one to take into account for the predictions in 3 days, time for the solar wind to travel from the Sun to the Earth.

3.3. Direct Comparison to the WSA Model

The PDF Model has been directly compared to the WSA Model. The WSA Model has been run for two whole years, 2008 and 2011, to make 1 to 5 days ahead predictions that have been compared to the same predictions made using the PDF Model.

Examples of predictions using the PDF and WSA Models are shown in Figure 16. They correspond to 1 to 5 days ahead predictions for 10 June to 10 July 2011. Two main observations can be made. First, the 1 day ahead predictions using the PDF Model get the sudden increase in speeds around 25 June (even if they underestimate the increase for the first 2 days), while the WSA Model does not. Then, both the PDF and WSA Models miss the variability of the solar wind speed under short time scales. The average structure of the speed is conserved but the high-frequency variability is lost.

The RMS difference between each model and the observations by the ACE satellite were calculated. The results are presented in Table 2. The year of 2011 is right between a solar minimum (2008) and a solar maximum (~ 2013). Looking at the PDF Model column of this table, the RMS starts off small on the first day, then grows to a maximum value on day 3, at which time it asymptotes. This is consistent with Figure 8. Additionally, the Persistence Model has a similar trend but asymptotes to a higher level, also consistent with Figure 8. However, the WSA Model has RMS errors that are approximately constant for all prediction horizons. These values are close to the value that the PDF Model asymptotes to, indicating that the PDF Model is most likely slightly better than the WSA for the first (approximately) 2 days, then is quite similar to the WSA Model. This is consistent with the study made by *MacNeice* [2009b], which showed that persistence is better than the WSA Model for the first 2 days of prediction. Table 2 also shows that for the last solar minimum that occurred in 2008, the PDF Model performs better than the WSA Model: the RMS is less by about 15 km/s for every prediction horizon. These results are consistent with the study by *Owens et al.* [2005], which found that the root-mean-square error between the WSA Model and the data is better at solar maximum than at solar minimum.

4. Conclusion

From this study a few different conclusions can be drawn:

1. The solar wind speed is quite periodic, but the period is not always exactly 27 days; it changes on a day-to-day basis and can vary between approximately ±5 days of the mean 27 day period. This makes it quite difficult to predict the next 5 days using data from the previous solar rotation. Indeed, using an optimized lag is best for the first 12 h of prediction; afterward, using a straight 27 days for the lag is best. Moreover, keeping the speed constantly equal to the median speed of the solar wind, 400 km/s, gives a constant N-RMS better than the N-RMS calculated from the last solar rotation without any lag.
2. The solar wind speed typically changes quite slowly, such that using the current solar wind speed is the optimum prediction for the first 7 h.
3. The current solar wind speed, as well as the trend in the solar wind speed (speeding up or slowing down) allows creation of probability distribution functions for the solar wind speed *i* hours into the future. The width of the PDF increases as time goes on, while the peak in the PDF decreases. The PDFs are narrower and taller than the distribution function of the solar wind in general, which means that these PDFs are more useful for predictions, than just assuming the solar wind distribution.

4. These PDFs can be used to generate ensemble prediction scenarios for the solar wind speed or to assign an uncertainty on the prediction based solely on the median speed.
5. Using a linear combination of the medians of the PDFs based on the current solar wind speed and trend as well as the actual solar wind speed from approximately 27 days ago allows the best predictions. This linear combination is highly weighted toward using the current value of the solar wind for low prediction horizons and using both roughly evenly for larger prediction horizons. It is shown that this linear weighting can change over the solar cycle, with more weighting on the previous rotation during solar minimum conditions.
6. The final PDF Model performs equal to or better than the Persistence Model for all times up to a 5 day prediction. The further out the prediction, the better the PDF Model does compared to the Persistence Model. After 5 days, the difference in the N-RMS of 0.11 between the two models reflects an improvement of the accuracy of the prediction of 40 km/s for a typical solar wind velocity of 400 km/s. The model also performs better than simply taking data from 27 days ago, although after approximately 3 days, the differences between the two levels off, with the PDF Model then being slightly better than simply using the data from 27 days ago.
7. The comparison of the PDF Model to the WSA Model predictions made for 2011 showed that the final PDF Model performs better than the WSA Model for 1 day ahead predictions. For longer prediction horizons, both models perform about the same. For 2008, the last solar minimum, the PDF Model performs better than the WSA Model for all prediction horizons, with a 15 km/s difference in the accuracy of the predictions.
8. The predictions with the PDF Model give better results during the last two solar minima (in 1996 and 2008) than during the solar maximum in 2002.
9. If the lag between solar rotations was predicted more accurately, then it would improve the predictions of the PDF Model. The current errors vary from 10% (predictions in one day) to 20% (predictions in 5 days), but with a perfect knowledge of the lag, they could decrease to values around 5%. It is unclear how to determine the perfect lag, though.

Acknowledgments
This research was funded by Air Force Office of Scientific Research grant FA9550-12-1-0265. We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb service and OMNI data. We are extremely grateful to Peter MacNeice for providing WSA simulation results for 2008 and 2011.

References

Arge, C. N., and V. Pizzo (2000), Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates, *J. Geophys. Res.*, *105*, 10,465–10,479.

Arge, C. N., D. Odstrcil, V. J. Pizzo, and L. R. Mayer (2003), Improved method for specifying solar wind speed near the Sun, *AIP Conf. Proc.*, *679*, 190–193.

Fry, C. N., W. Sun, C. Deehr, M. Dryer, Z. Smith, S.-I. Akasofu, M. Tokumaru, and M. Kojima (2001), Improvements to the HAF solar wind model for space weather predictions, *J. Geophys. Res.*, *106*, 20,985–21,001.

Fry, C. D., T. Detman, M. Dryer, Z. Smith, W. Sun, C. Deehr, S.-I. Akasofu, C.-C. Wu, and S. McKenna-Lawlor (2007), Real-time solar wind forecasting: Capabilities and challenges, *J. Atmos. Sol. Terr. Phys.*, *69*, 109–115.

Leamon, R. J., and S. McIntosh (2007), Empirical solar wind forecasting from the chromosphere, *Astrophys. J.*, *659*, 738–742.

Lee, C., J. Luhmann, D. Odstrcil, P. MacNeice, I. de Pater, P. Riley, and C. N. Arge (2009), The solar wind at 1 AU during the declining phase of solar cycle 23: Comparison of 3D numerical model results with observations, *Sol. Phys.*, *254*, 155–183.

Liu, D. D., C. Huang, J. Y. Lu, and J. S. Wang (2007), The hourly average solar wind velocity prediction based on support vector regression method, *Mon. Not. R. Astron. Soc.*, *413*, 2877–2882.

Luo, B., Q. Zhong, S. Liu, and J. Gong (2008), A new forecasting index for solar wind velocity based on EIT 284 A observations, *Sol. Phys.*, *250*, 159–170.

MacNeice, P. (2009a), Validation of community models: Identifying events in space weather model timelines, *Space Weather*, *7*, S06004, doi:10.1029/2009SW000463.

MacNeice, P. (2009b), Validation of community models: 2. Development of a baseline using the Wang-Sheeley-Arge Model, *Space Weather*, *7*, S12002, doi:10.1029/2009SW000489.

Mikic, Z., J. A. Linker, D. D. Schnack, R. Lionello, and A. Tarditi (1999), Magnetohydrodynamic modeling of the global solar corona, *Phys. Plasma*, *6*, 2217–2224.

Norquist, D., and W. Meeks (2010), A comparative verification of forecasts from two operational solar wind models, *Space Weather*, *8*, S12005, doi:10.1029/2010SW000598.

Odstrcil, D. (2003), Modeling 3-D solar wind structure, *Adv. Space Res.*, *32*, 497–506.

Odstrcil, D., V. J. Pizzo, J. A. Linker, P. Riley, R. Lionello, and Z. Mikic (2004), Initial coupling of coronal and heliospheric numerical magnetohydrodynamic codes, *J. Atmos. Sol. Terr. Phys.*, *66*, 1311–1320.

Owens, M., C. N. Arge, H. Spence, and A. Prembroke (2005), An event-based approach to validating solar wind speed predictions: High-speed enhancements in the Wang-Sheeley-Arge Model, *J. Geophys. Res.*, *110*, A12105, doi:10.1029/2005JA011343.

Owens, M. J., H. E. Spence, S. McGregor, W. J. Hughes, J. M. Quinn, C. N. Arge, P. Riley, J. Linker, and D. Odstrcil4e (2008), Metrics for solar wind prediction models: Comparison of empirical, hybrid, and physics-based schemes with 8 years of L1 observations, *Space Weather*, *6*, S08001, doi:10.1029/2007SW000380.

Owens, M. J., R. Challen, J. Methven, E. Henley, and D. R. Jackson (2013), A 27 day persistence model of near-Earth solar wind conditions: A long lead-time forecast and a benchmark for dynamical models, *Space Weather*, *11*, 225–236, doi:10.1002/swe.20040.

Ram, S. T., C. H. Liu, and S. Su (2010), Periodic solar wind forcing due to recurrent coronal holes during 1996-2009 and its impact on Earth's geomagnetic and ionospheric properties during the extreme solar minimum, *J. Geophys. Res.*, *115*, A12340, doi:10.1029/2010JA015800.

Robbins, S., C. Henney, and J. Harvey (2006), Solar wind forecasting with coronal holes, *Sol. Phys.*, *233*, 265–276.

Smith, Z., T. Detman, M. Dryer, C. D. Fry, C.-C. Wu, W. Sun, and C. Deehr (2004), A verification method for space weather forecasting models using solar data to predict arrivals of interplanetary shocks at Earth, *IEEE Trans. Plasma Sci.*, *32*, 1498–1505.

Toth, G., and D. Odstrcil (1996), Comparison of some flux corrected transport and total variation diminishing numerical schemes for hydrodynamic and magnetohydrodynamic problems, *J. Comput. Phys.*, *128*, 82–100.

Wang, Y.-M., and J. N. R. Sheeley (1990), Solar wind speed and coronal flux-tube expansion, *Astrophys. J.*, *355*, 726–732.

Wang, Y.-M., and J. N. R. Sheeley (1992), On potential field models of the solar corona, *Astrophys. J.*, *392*, 310–319.

Wintoft, P., and H. Lundstedt (1997), Prediction of daily average solar wind velocity from solar magnetic field observations using hybrid intelligent systems, *Phys. Chem. Earth*, *22*, 617–622.

Wright, J. M., T. J. Lennon, R. W. Corell, N. A. Ostenso, W. T. Huntress, J. F. Devine, P. Crowley, and J. B. Harrison (1995), *National Space Weather Program: The Strategic Plan*, National Space Weather Program Council, FCM-P30, Washington, D. C.