Forecasting the properties of the solar wind using simple pattern recognition

AGU PUBLICATIONS

Space Weather

RESEARCH ARTICLE
10.1002/2016SW001589

Key Points:
- Solar wind parameters can be predicted with lead times of up to several weeks
- IMF B_n, arguably the most important space weather parameter, is also the least predictable
- Dynamic time warping may lead to improvements in the forecasting abilities of these algorithms

Correspondence to:
P. Riley,
pete@predsci.com

Citation:
Riley, P., M. Ben-Nun, J. A. Linker, M. J. Owens, and T. S. Horbury (2017), Forecasting the properties of the solar wind using simple pattern recognition, *Space Weather*, *15*, 526–540, doi:10.1002/2016SW001589.

Received 15 DEC 2016
Accepted 1 MAR 2017
Accepted article online 6 MAR 2017
Published online 28 MAR 2017

Forecasting the properties of the solar wind using simple pattern recognition

Pete Riley[1], Michal Ben-Nun[1], Jon A. Linker[1], M. J. Owens[2], and T. S. Horbury[3]

[1]Predictive Science Inc., San Diego, California, USA, [2]Space and Atmospheric Electricity Group, Department of Meteorology, University of Reading, Reading, UK, [3]Blackett Laboratory, Imperial College London, London, UK

Abstract An accurate forecast of the solar wind plasma and magnetic field properties is a crucial capability for space weather prediction. However, thus far, it has been limited to the large-scale properties of the solar wind plasma or the arrival time of a coronal mass ejection from the Sun. As yet there are no reliable forecasts for the north-south interplanetary magnetic field component, B_n (or, equivalently, B_z). In this study, we develop a technique for predicting the magnetic and plasma state of the solar wind Δt hours into the future (where Δt can range from 6 h to several weeks) based on a simple pattern recognition algorithm. At some time, t, the algorithm takes the previous Δt hours and compares it with a sliding window of Δt hours running back all the way through the data. For each window, a Euclidean distance is computed. These are ranked, and the top 50 are used as starting point realizations from which to make ensemble forecasts of the next Δt hours. We find that this approach works remarkably well for most solar wind parameters such as v, n_p, T_p, and even B_r and B_t, but only modestly better than our baseline model for B_n. We discuss why this is so and suggest how more sophisticated techniques might be applied to improve the prediction scheme.

1. Introduction

Since its prediction in 1958 [*Parker*, 1958] and observational confirmation in 1959 [*Harvey*, 2007], forecasting the future conditions of the solar wind has become ever more important as our society relies increasingly on technology [e.g., *Board et al.*, 2012]. The value and variability of the z component of the interplanetary magnetic field has, arguably, a more significant impact than any other parameter for geoeffective phenomena. More strictly, it is the dawn-dusk component of the solar wind electric field ($E_y = -v_x \times B_z$), as well as the plasma-β, Mach number, and density that modulate the transmission of energy from the heliosphere into the magnetosphere and, potentially, drive magnetic storms [*Dungey*, 1961; *Cassak and Shay*, 2007; *Borovsky et al.*, 2008]. Given its key role in space weather, it may seem surprising that it is noticeably absent from any of the parameters that the National Space Weather Centers forecast. The reason, of course, is that predicting B_z is extremely difficult.

Over the years, a variety of techniques have been proposed to predict the state of the solar wind. These vary from purely statistical approaches to physics-based models and all manner of hybrids in between. For example, *Chen et al.* [1996] proposed a technique based on identifying sinusoidally varying large-scale features in the z component (meridional) of the interplanetary magnetic field (IMF B_z). However, the work remained a "proof of concept" and has, thus far, not been further developed in any rigorous way.

The WSA-Enlil solar wind prediction model, which is the first operational space weather model at NOAA, provides 1–4 day advance warning of large-scale solar wind structure as well as Earth-directed coronal mass ejections (CMEs) [*Farrell*, 2011]. Two models are combined to produce predictions of the ambient solar wind: WSA [*Wang and Sheeley*, 1990; *Arge and Pizzo*, 2000], which is a modified Potential Field Source Surface (PFSS) model, and Enlil, which is a heliospheric MHD model [*Odstrcil*, 1993]. The former computes estimates of the solar wind speed at 30 solar radii (R_S), as well as the radial component of the coronal magnetic field (B_r), which are used to drive Enlil. Together, the ambient solar wind solutions can produce estimates of speed, density, and temperature, as well as IMF field strength and sector boundaries information. However, no meaningful estimates for B_z can be made. A simple CME generator can also produce ICMEs traveling through these solutions for specific time periods [*Pizzo et al.*, 2015]. Primarily, these are used to estimate the dynamic properties of the ejecta at 1 AU as well as the time of transit from the Sun to Earth. While there are no magnetic fields embedded

within the simulated ejecta, estimates of B_z within the sheath can be made for events sufficiently fast to drive a fast-forward shock [e.g., *Mays, et al.*, 2015].

Several other ambient solar wind models have been proposed over the years, only a handful of which have been taken into the "operational" arena. One such model is the Empirical Solar Wind Forecast [*Reiss et al.*, 2016], which is based on the observed correlation between areas of coronal holes and the solar wind speed at 1 AU. The published root-mean-square errors for the forecasts are on the order of 100 km s^{-1}, with uncertainties in the arrival times and sizes of high-speed streams of ~1 day and 100 km s^{-1}, respectively. Importantly, these more recent models are accompanied by quantitative estimates of their accuracy.

Jackson et al. [2015] recently proposed an appealing but as yet tentative model for estimating nonzero B_z contributions to the field at 1 AU based on estimates from the low corona. They used a PFSS model to estimate nonzero B_z in the low corona and propagated it out to 1 AU where they compared with in situ measurements. They provided several Carrington rotations that appeared to show a modest correlation between model results and observations. However, no mechanism was provided for how these fields could be transported out, a process that contradicts all current global models of the extended corona.

In another study, *Savani et al.* [2015, 2017] combined several empirically based models to create a "pipeline" for predicting the magnetic field properties of magnetic clouds (MCs) in the solar wind. Strictly speaking, this approach must be viewed as a proof of concept and not a prediction; however, it sets out an appealing framework for considering how to best approach event-based prediction and which models should or should not be included. The use of interchangeable components allows the user to test different, potentially superior approaches to address a specific piece in the chain. Their specific framework relied on estimating the initial orientation and location of the flux rope (only flux rope CMEs are amenable to this approach) based on statistical estimates and the "solar hemispheric rule" in particular [*Bothmer and Schwenn*, 1998]. Then, using the graduated cylindrical shell model [*Thernisien et al.*, 2009], they estimated the physical extent of the ejecta and thus the likely trajectory along which the Earth would pass through. Finally, using a force-free model for the magnetic structure of the flux rope, they extracted the time series traces that would be measured by Earth-based spacecraft. They were able to demonstrate a basic agreement for a handpicked set of eight events. However, it would be fairer to say that they were able to "reproduce" the observations, not "predict" them.

Our ability to forecast the value of solar wind parameters is extremely sensitive to the specific parameter considered. For example, the sign of the radial component of the interplanetary magnetic field (IMF) is relatively straightforward to predict reasonably accurately. Similarly, the bulk solar wind speed can, in the absence of transient phenomena, be predicted with a basic degree of fidelity [e.g., *Riley et al.*, 2001]. The *z* component of the IMF, B_z, however, has, thus far, remained difficult, if not impossible to predict. At least in part, these differences can be attributed to the fundamental processes that drive the large-scale variations in them. The sign of B_r, for example, is a reflection the Sun's large-scale field. In contrast, B_z, under quiescent conditions, fluctuates about a mean value of zero.

The origins of nonzero values of B_z are quite complicated. In fact, as we will discuss below, as a first approximation, and in Parker's original derivation, we could assume that $B_z = 0$. However, for any practical purposes, and certainly for space weather applications, this approximation is not useful. There are a multitude of processes, some related to one another and some overlapping, that can be branded as "B_z-producing" phenomena. Generalizing slightly from Parker's original simple picture, if we allow the solar wind flow speed to vary, even in an idealized manner, fast flow over the poles, say, and slow flow around the neutral line, this would introduce velocity shear, as fast solar wind caught up to, and overtook slower wind, and this would produce large-scale and recurrent B_z variations, albeit modest. These might be responsible for the apparent ambient solar wind B_z that *Jackson et al.* [2015] and *Ulrich and Tran* [2016] may have detected.

A significantly more important phenomena for producing nonzero B_z are CMEs and, in particular, magnetic clouds (MCs). In addition to the large, smooth, and rotating fields within them, fast MCs drive shocks, which compress the plasma ahead of them and amplify any transverse fields within [*Owens et al.*, 2005]. This creates a distinctive sheath region that is responsible for a significant fraction of all geomagnetic activity [*Lugaz et al.*, 2016]. Field lines also drape over fast-moving ICMEs resulting in non-Parker field lines, slippage, and the creation of meridional components to the magnetic field. Corotating interaction regions (CIRs) too produce intervals of nonzero B_z since they tend to be organized in tilted patterns [*Riley et al.*, 1996].

Within, and surrounding CIRs, we can further distinguish nonzero-producing B_z phenomena: stream interfaces, the heliospheric current sheet, and CIR-associated forward and reverse shocks.

In addition to these large-scale effects, there are many small-scale features that produce substantial power in the fluctuations of B_z, including Alfvén waves and turbulence. The properties of these types of fluctuations—at least statistically—are well described [*Horbury et al.*, 1995, 2005; *Owens et al.*, 2011]. In fact, given a power law relationship for the magnitude of the fluctuations as a function of frequency, we can reconstruct these fluctuations precisely for a variety of types of solar wind. Unfortunately, we cannot reconstruct the phase information for the fluctuations. Thus, we are limited to a statistical forecast of their properties. While not ideal, it is likely still useful for geomagnetic forecasts, where the actual phase information about the fluctuations only becomes critical below some characteristic frequency [*Owens et al.*, 2014].

Finally, we remark that there are a "potpourri" of other phenomena that may produce nonzero B_z to varying but modest degrees. For example, reconnection jets [*Shimojo and Shibata*, 2000], magnetic holes [*Fränz et al.*, 2000], reconnection exhausts [*Gosling*, 2011] as well as "blobs" which may be a source of some of the slow solar wind [*Wang*, 1994]. In principle, these could all produce limited nonzero intervals in B_z.

In this report, we describe a simple technique for predicting the state of the solar wind over the next Δt hours based on recognizing the pattern of the solar wind during the previous Δt hour period and assuming that previous intervals with similar variations might provide some insight into the state during the following interval. We provide a detailed statistical analysis of this technique by applying it to the entire NASA OMNI_M data set [*King and Papitashvili*, 2005] and conclude that it can—under certain conditions—be a powerful tool for forecasters. We focus on demonstrating the potential of pattern recognition (PR) for solar wind forecasting. The sensitivity of forecast skill to the details of the pattern recognition and the subsequent implications for geomagnetic forecasting will be examined in a future study.

In the following section, we describe the data set we analyze as well as the forecasting model we have developed. We then show several case studies emphasizing where the model works and where it does not. Finally, we discuss the implications of this model and suggest several refinements that can be made that we believe will improve the accuracy and robustness of the basic model proposed here.

2. Methods

2.1. Data

For this study, we use data from NASA's OMNI_M data set (obtained through the COHOWeb data server). We chose 1 h resolution data since these were sufficiently resolved to capture large-scale variations in the solar wind (e.g., MCs, CIRs, and long-period Alfvén waves) but coarse enough that the contribution from high-frequency turbulence was reduced. Additionally, these data span a much longer epoch than the higher-resolution 1 and 5 min data sets. Using 1 h averaged data suggested a minimum reasonable prediction window of 6 h, but allowed for much longer prediction windows to be considered (12 h, 24 h, …, 27×24 h, etc.).

Although we have, thus far, avoided a precise definition of what we mean by the *z* component of the magnetic field (B_z), for the remainder of the study, we will instead use the heliospheric-centered RTN coordinate system. The RTN coordinate system is a "spacecraft-centered" system where **R** is a unit vector from the Sun to the spacecraft, **T** is $\mathbf{\Omega} \times \mathbf{R})/|(\mathbf{\Omega} \times \mathbf{R})|$, and **N** completes the right-handed triad. Here $\mathbf{\Omega}$ is the Sun's spin axis. Intuitively, **T** points in the direction of planetary motion, and **N** points northward. In part, our choice to use RTN coordinates is designed to avoid any misunderstanding about whether B_z is in a geomagnetic-centered system (GSM) or ecliptic (GSE). However, more importantly, we believe that it is better to base our forecasts on the fundamental measurements that are made by the spacecraft and published by the forecasting centers, such as NOAA.

We used in situ measurements of the magnetic field vectors and magnitude (B_r, B_t, B_n, and B) as well as standard plasma parameters: speed (v), number density (n_p), and proton temperature (T_p). For this report, we emphasize results for B_n, v, n_p, and T_p, which, together, can be used to construct the parameters most necessary to forecast geoeffective phenomena and, in particular, the dawn-dusk electric field ($E_y = -v_x \times B_z$) and momentum flux ($n_p \times v^2$).

Although the full data set stretches back to the early 1970s, resulting in almost 400,000 one hour data points, we restricted the statistical component of our analysis to all data from 2000 to 2010 (~96,000 points).

Setting the start date to 2000 ensured that there would be sufficient historical data prior to these data points on which to construct the forecast ensembles, and setting the end date to 2010 provided a data set that was proportionately representative of all phases of a solar cycle.

2.2. Models

In this study, we develop a simple pattern recognition technique for identifying previous intervals in the entire solar wind data set that are most likely the interval recently observed and use the data that follow those intervals as a set of forecasts (realizations) for what is likely to occur in the near future. It relies on the assumption that past variability is an indicator of future variability. In a purely stochastic time series, the approach would fail. Thus, we anticipate that the value to this scheme lies in identifying large-scale coherent structure, the leading portions of which are forbearers of what will come later.

The procedure we adopt is as follows. First, the algorithm takes the last Δt hours (say, 24) of observations of some solar wind parameter (say, B_n) at 1 h resolution and slides it backward in time, hour by hour, with a window of Δt hours. For each interval, the Euclidean distance between those earlier observations and the current (last) Δt hour window is calculated. The Euclidean distance is

$$d\left(\mathbf{d}_1, \mathbf{d}_2\right) = \sqrt{\sum_{i=1}^{n} \left(d_1^i - d_2^i\right)^2} \tag{1}$$

which can be thought of as an estimate of the difference between the two time series (d_1 and d_2). (The more familiar "chi-square" distance is a weighted Euclidean distance.)

For each window under consideration within the 2000–2010 period, we compute the Euclidean distance d over the entire data set prior to that window. Since the data set stretches back to the early 1970s, this results in between 300,000 and 400,000 estimates. Each window is then ranked in terms of its Euclidean distance. It is important to stress that only past data points are used by the pattern recognition algorithm in assembling the forecast. We retain the "top" $N = 50$, that is, those N intervals with the lowest values of d. For each of these intervals, the following Δt hour period is then used to form the basis of the forecast. Here we focus on 6, 24 h, and a few multiday (up to 40 days) intervals.

We also introduce a baseline, or reference model. Based on the discussion in section 1, on average, the z component of the magnetic field is zero. Thus, an obvious model against which to compare is a so-called "zero" model, which predicts that our best forecast in the future is that B_n will be and remain exactly zero. It turns out, somewhat disconcertingly, that this is a surprisingly accurate, though not particularly useful, forecast. For other parameters, we use the average value of the parameter during the previous Δt hours as the baseline prediction for the next Δt hours. These are typically referred to as "persistence" (or "baseline") models.

2.3. Results

To introduce the simple pattern recognition model, we apply it to a well-studied magnetic cloud observed in mid-September 2000 [e.g., *Nieves-Chinchilla et al.*, 2002]. Figure 1 shows a 2 day interval of B_n from 17 September 2000 04:00 UT to 19 September 2000 04:00. The data (black line) show a sharp rise to 20 nT, followed by a subsequent swing down to −20 nT and a rise once more. For the purpose of making a forecast, we take the current or "now" time to be 04:00 on 18 September 2000, marked by the thick red vertical line. We assume that only data to the left of this line are available for analysis. These are significant values of the z component of the magnetic field, and they produce notable geomagnetic activity ($Dst \sim -200$ nT). Using the algorithm described in the previous section, we located the 50 intervals in the entire (prior to 17 September 2000 04:00 UT) OMNI_M data set that most closely matched these variations (as estimated using equation (1)). These are shown by the gray traces to the left of the red vertical line. We note that these traces match the variations that occurred during the last 24 h well. We then take the 24 h intervals that followed each of these best matches and plot them to the right of the red vertical line. They show considerably more scatter. However, as suggested by the average ensemble curve (purple), the overall evolution of the forecasted curve matches the actual magnetic field that was subsequently observed (black curve). It is worth noting that in this particular case, the high correlation might be driven by the phenomenon of "regression toward the mean"; that is, whenever you have a significant deviation from some average, say, zero for B_n, there will be a tendency for it to return to that value.

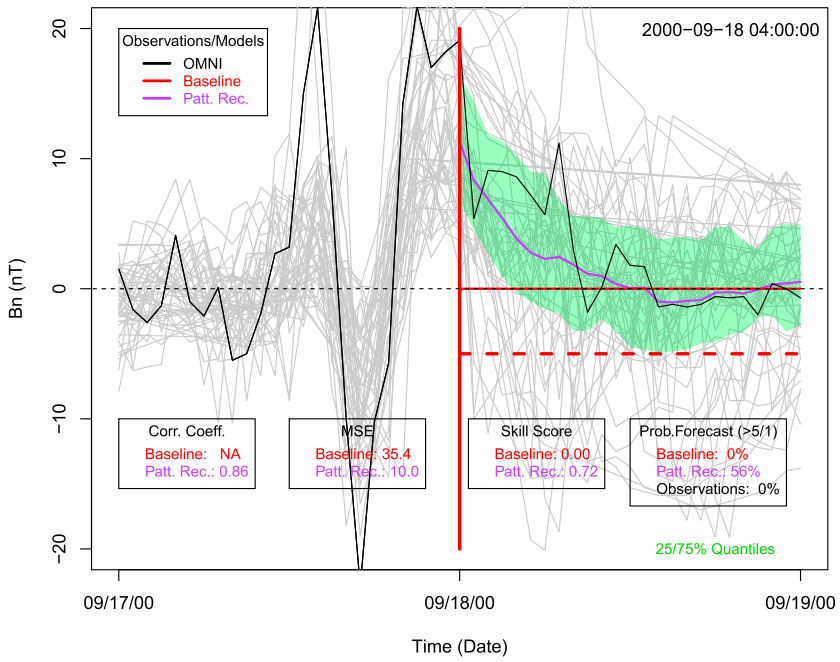

Figure 1. Time series of solar wind B_n for 17–19 September 2000. The now time is marked by the thick red vertical line (and indicated on the top right corner). Only data prior to 17 September 2000 are used in the analysis. The solid black curve shows the 24 h of data preceding now and the data following it that were actually observed. Each of the gray curves represents realizations obtained from the PR analysis. The data that were observed after each matched interval are shown to the right of the red line, providing an ensemble of forecasts. The average of these predictions is shown by the purple curve, and the 25% and 75% quantiles are marked by the green area. Our baseline model, the zero model, is shown by the horizontal red line (at $B_n = 0$). A variety of metrics are given in the boxes across the lower half of the plot: the Pearson correlation coefficient, the mean square error (MSE), the associated skill score, and a probabilistic forecast based on a heuristic forecaster's rule. See text for more information.

We can calculate several metrics to assess the quality of the forecast. First, the correlation coefficient between the observed and PR-forecasted profiles was 0.87. Unfortunately, we cannot compare this to a correlation coefficient calculated from the zero model, which, since its value remains unchanged, cannot be defined. Second, we can compute the mean square error (MSE) between the observed and forecasted measurements. In this case, for the PR model, this was estimated to be 10.0 nT^2, which is significantly less than for the baseline (zero) model (35.4 nT^2), showing that the PR model substantially outperformed the baseline model. We can also define and estimate a skill score. For simplicity, we use the associated skill score, which is defined as

$$SS = 1 - \frac{MSE_{forecast}}{MSE_{baseline}} \tag{2}$$

where $MSE_{baseline}$ is the reference model's mean square error and $MSE_{forecast}$ is the PR model's mean square error. A positive number demonstrates a capable model, while a negative number suggests that the model is worse than the reference or baseline model. Since the zero model is our baseline model, that fraction is one for the zero model, and the skill score reduces to zero. A value of 0.73 for the PR model is promising. Finally, the last box in the bottom right of Figure 1 is a probability forecast. This is an idealized metric based on talks with NOAA/SWPC staff. In particular, they would find it useful to know if the z component of the magnetic field is going to drop below −5 nT for a period of an hour or more. This is, of course, analogous to terrestrial weather forecasts, which might report that there is a 60% chance of rain over the next 3 h say, within a certain region. Here we are predicting whether B_n will remain southward for 1 h or more during the next 24 h. The −5 nT threshold is shown by the dotted red line in the prediction window. For this interval, using the ensemble of realizations, we can estimate that there is a 54% probability that B_n will drop below −5 nT during the following 24 h period.

The quantiles produced by the ensembles serve to bracket the forecast, and it is interesting to note that they encompass the fluctuations in the actual observed measurements. This is a desirable feature of any model

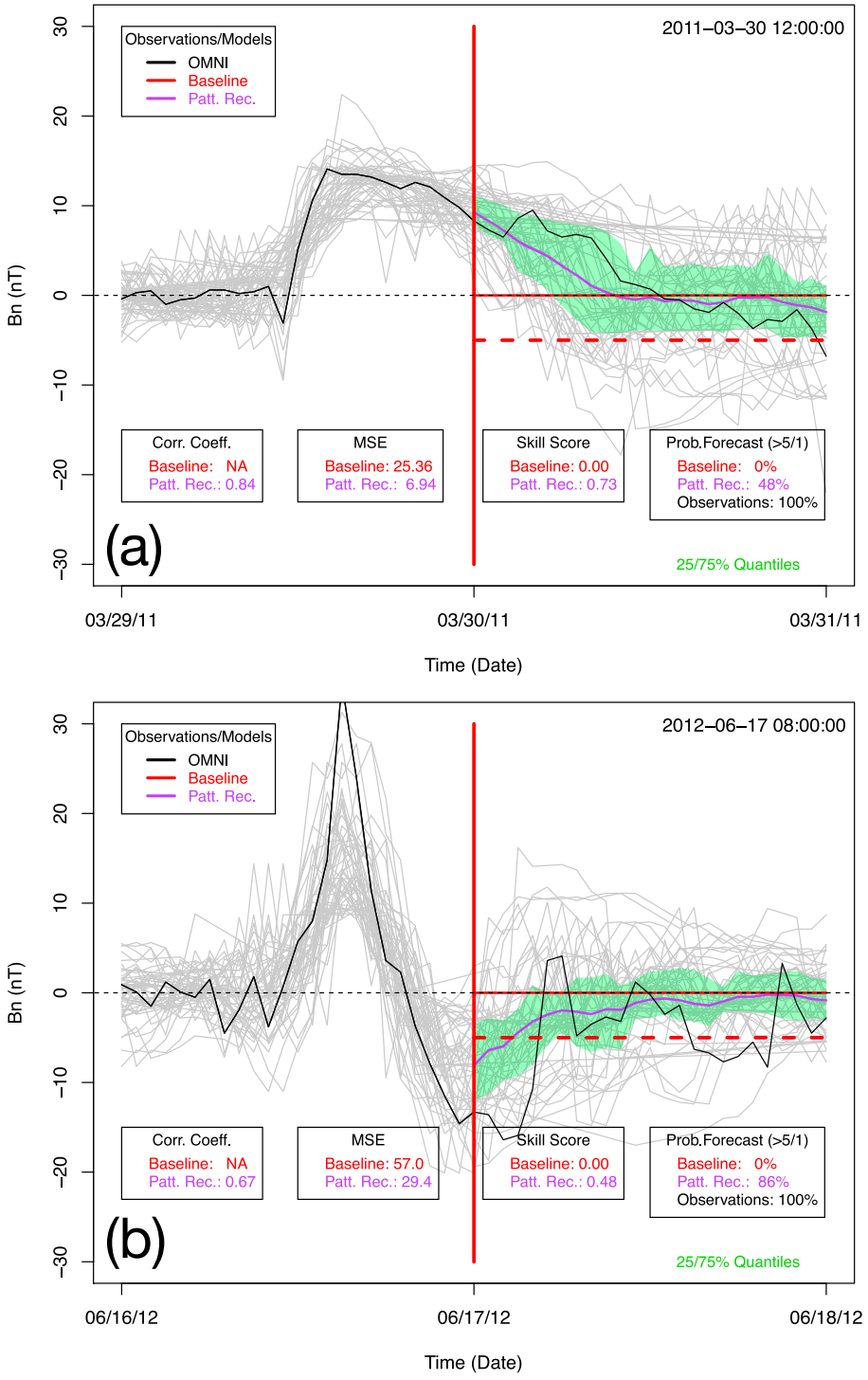

Figure 2. As Figure 1 but for a 1 day forecasting window in (a) March 2011 and (b) June 2012.

prediction that provides confidence bounds. As the quality of the forecast improves (worsens), the confidence intervals should decrease (increase) commensurately.

We also investigated the sensitivity of the results on the number of members in the ensemble. Specifically, we repeated selected intervals using 10, 20, 50, and 100 members. In all cases, the measures of accuracy of the forecast (correlation coefficient, MSE, skill score, and probability that $B_n < -5$ nT) were virtually the same.

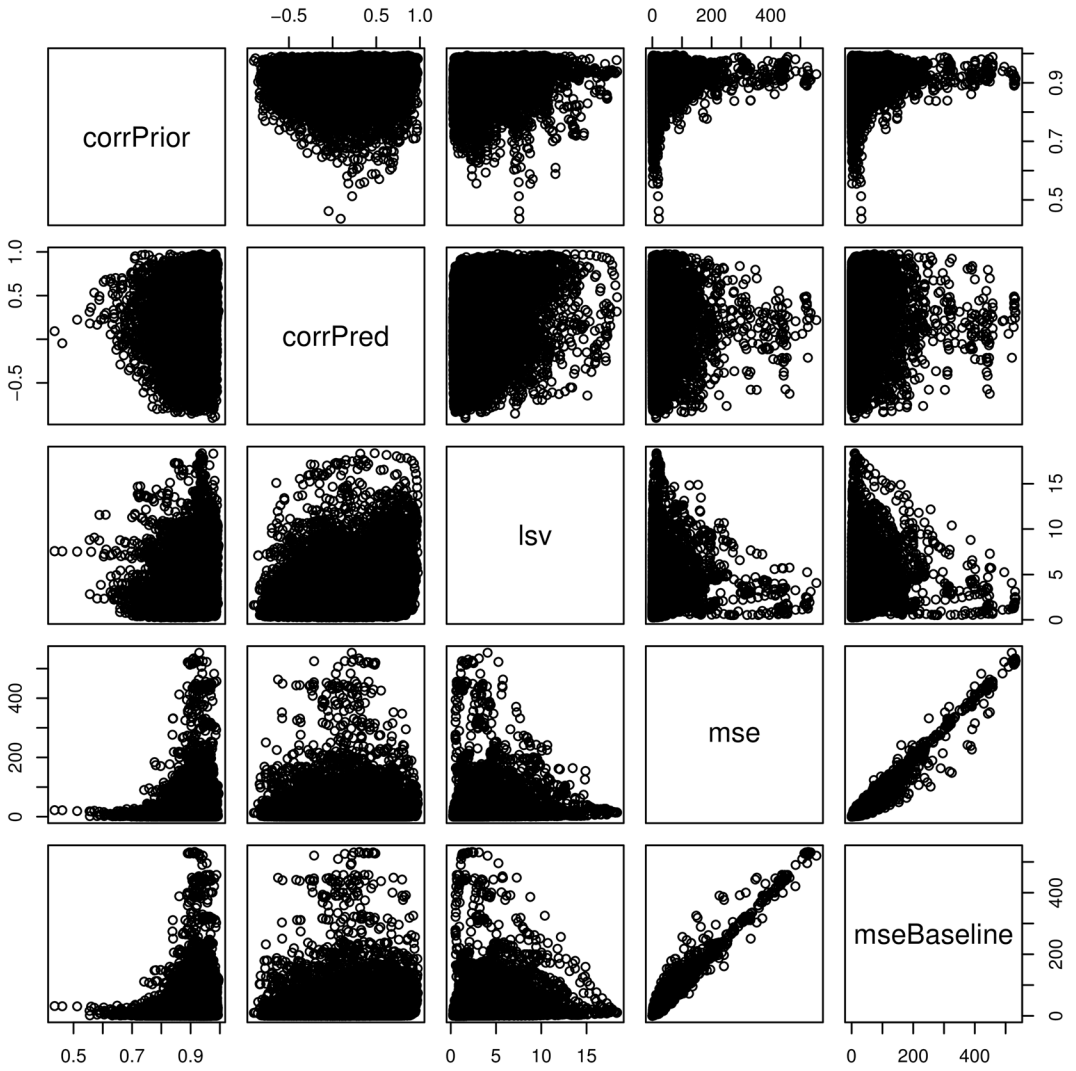

Figure 3. A scatterplot matrix of parameters computed from B_n comparing (a) the prior correlation coefficient (corrPrior), that is, the average of the best pattern-matched intervals with the interval being predicted; (b) the predicted correlation coefficient (corrPred), i.e., the correlation between the predicted and observed future intervals; (c) LSV, a measure of large-scale variations during the observed window; (d) the mean square error between the observed and predicted interval; and (e) the mean square error of the baseline (zero) model.

However, when $N = 10$ or 20, the 25th/75th percentile green band became either very ($N = 10$) or moderately ($N = 20$) jagged. Only for $N \geq 50$ did the envelopes become smooth. Thus, factoring in computational time, we arrived at the value: $N = 50$.

In Figure 2 we provide two additional examples where a clear large-scale signal was observed in the magnetic field during the "previous" 24 h. In Figure 2a, the correlation of the forecasted profile was 0.85, and the MSE was substantially less than that of the zero model. In Figure 2b, the correlation of the predicted time series was lower (0.67), but still significant. For both cases, the correlation of observed and best matched intervals was exceedingly good (>0.96), suggesting that the algorithm can identify a sufficiently large number of prior intervals that closely match the recently observed data. Note also that in each of the three cases (Figures 1 and 2), several consecutive intervals were found to be the best match (that is, a window shifted by one or two more hours) suggesting that the realizations are not completely independent.

Moving away from case studies, we can generalize this analysis by looking at every data point in the data set during the 11 year period from 2000 to 2010 and assuming that this is now. For each now (and there are 97,00 of them) we look at the previous 24 h stretching back through the entire data set (to the early 1970s) then

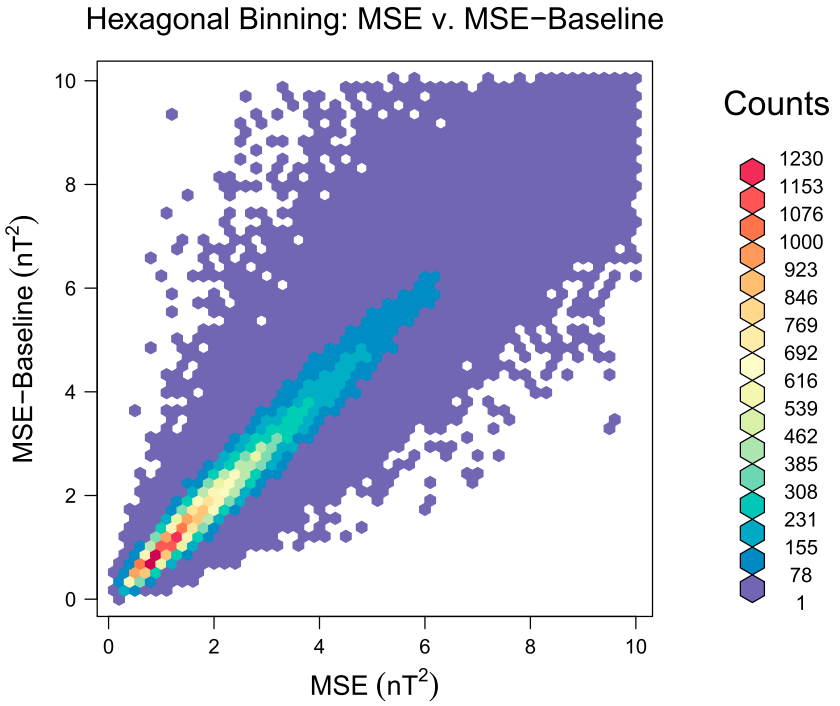

Figure 4. A comparison between the mean square error (MSE) of the pattern PR and the baseline (zero) model for B_n forecasts. The number of data points within each hexagon is shown via the color coding.

compute the Euclidean distance for this interval with every other overlapping interval in the data set, rank each of them, compute the MSEs for the best model forecast, and compare this with the baseline model.

Figure 3 summarizes this analysis using a scatterplot matrix, which visually shows whether any linear correlations exist amongst multiple variables. In it, we plot (1) the correlation coefficient of the observed data (that is during the 24 window prior to now) with the ensemble average of the most closely matching intervals (corrPrior); (2) the correlation coefficient of the PR model forecast with the data that was then observed (corrPred); (3) LSV, defined here as the mean of the absolute value of the parameter (see below for more details); (4) the MSE of the PR model forecast with the observations (mse); and (5) the MSE of the zero model (mseBaseline). Several points are worth noting. First, corrPrior is consistently high, typically around >0.9. Second, corrPred shows more spread and can be negative as well as positive. It is important to note, however, that these data, as displayed, are somewhat misleading: with 97,000 points, many data are superimposed upon one another. Third, as evidenced from either of the two panels in the bottom right, there is an asymmetry in the mse versus mse-Baseline values, with a small tendency for mse to be lower than mseBaseline, typically when both parameters are small. These are the intervals we would like to be able to forecast well. Again, because many points are superimposed upon one another, this asymmetry may be overemphasizing a very modest or infrequent effect.

To mitigate this effect, in Figure 4 we show MSE versus mseBaseline values using a hexagonal binning technique. Now instead of points being overplotted, they are grouped into a density map. Thus, the color of each hexagonal point represents the number of data points falling into that area. From this, we infer that, in general, the PR model and zero model result in very similar forecast accuracy, at least based on MSE as an estimate of performance. The "spur" of points in the rightmost panel of the fourth row in Figure 3 is visible as the asymmetry in the purple points, with more of them tending to be in the upper left, than lower right, again suggesting a small tendency for the PR model to outperform the baseline (zero) model, at least under limited conditions.

We can look for potential asymmetries in another way, by plotting the ratio of MSE for the PR model to MSE-Baseline, that is, the reference model. This is shown in Figure 5 for a 24 h window (Figure 5a) and a 6 h window (Figure 5b). We conclude that, in general, there is no significant, systematic difference between the PR and zero models for 24 h windows. On the other hand, Figure 5b illustrates how this changes modestly

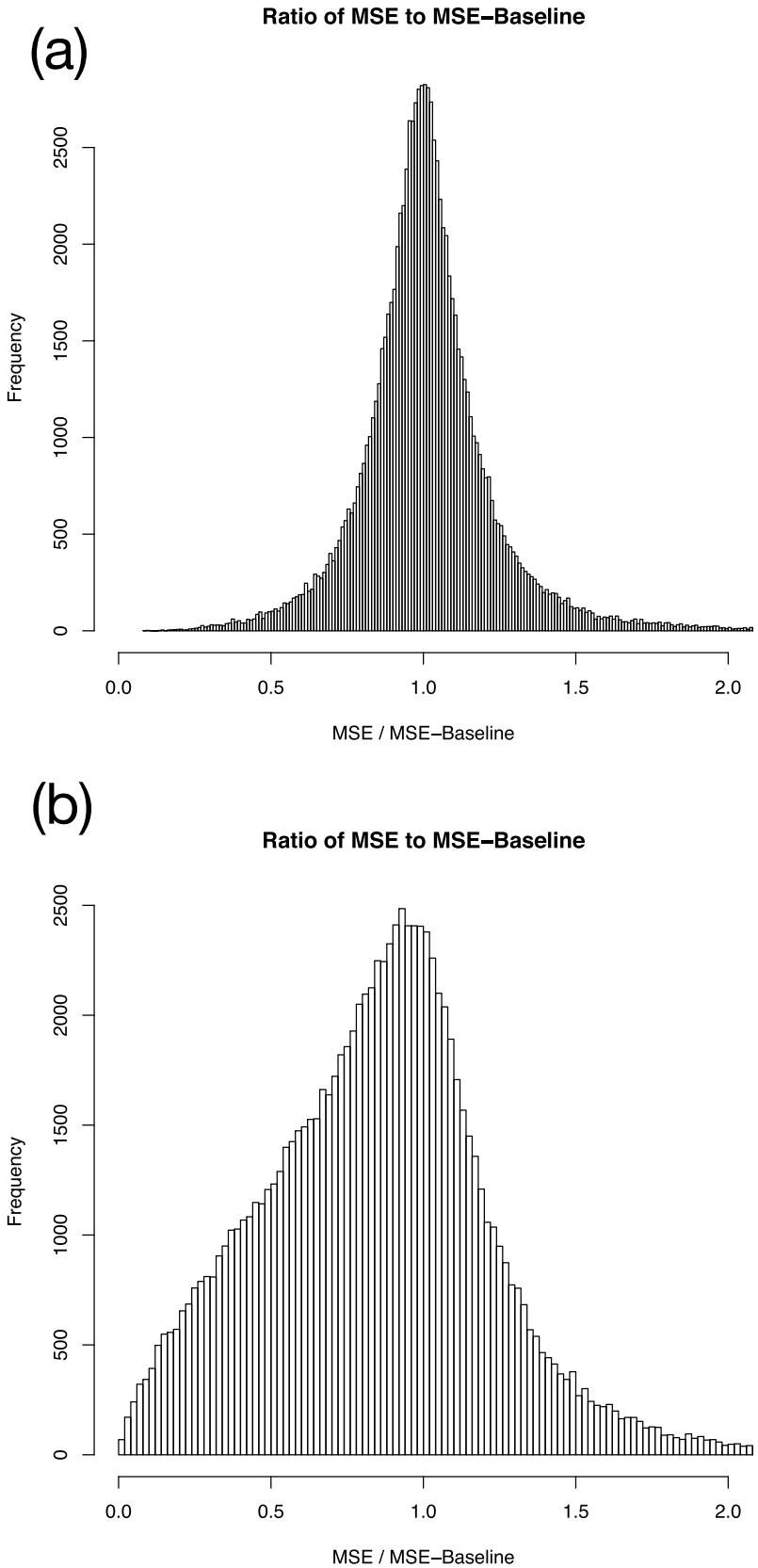

Figure 5. (a) The ratio of the mean square error (MSE) for the pattern recognition model to the baseline (zero) model for B_n forecasts with a window of 24 h. (b) As Figure 5a but for a 6 h window.

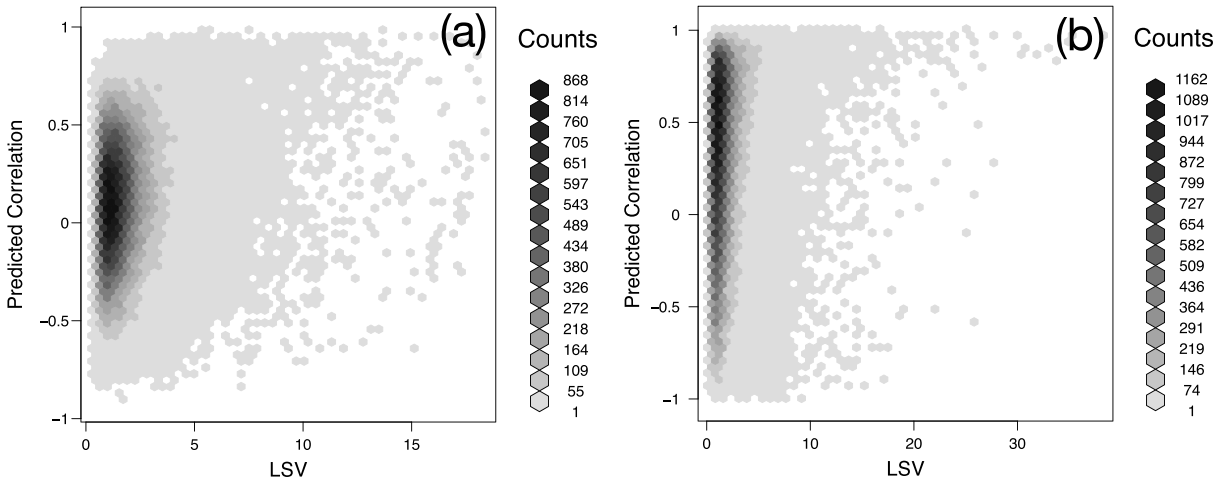

Figure 6. A comparison between the predicted correlation and LSV for B_n for a (a) 24 h and (b) 6 h window. The number of data points within each hexagon is shown via the gray scale coding.

when the prediction window is reduced to 6 h. While statistically significant, it is not yet clear whether this asymmetry, favoring the PR model, is sufficiently large to produce actionable information for the operational community.

In an effort to identify which intervals might be the most amenable to accurate forecasting we created several measures of a parameter, aimed at capturing large-scale variations (LSV) in the magnetic field, during the observed Δt hour window. In essence, it seeks to measure the "predictability" of the currently observed window. One method was to simply compute LSV $=< |B_n| >$ for the interval. A large value of this would suggest the presence of a sustained interval of nonzero B_n. However, intervals of large-amplitude Alfvén waves would also produce somewhat large values of this. Figure 6 compares LSV with the correlation of the forecast for 24 h (Figure 6a) and 6 h (Figure 6b) windows. Generally, and unsurprisingly, for the 24 h window (Figure 6a), most of the solar wind is in a state of low-LSV and there is no obvious trend with how well the predictions correlate with observations. Such cases could be identified in recently observed data by their high LSV value and predicted to have a good forecasting accuracy. Instead, only a few events are seen and it is not clear that they represent a unique set of cases. At 6 h (Figure 6b), the predicted correlation generally increases with a substantial number of cases having a predicted correlation in excess of 0.5. What we had hoped to find was a cluster of intervals in the upper right, which have a high LSV value and high correlation, and thus amenable to prediction, since we would know a priori that the recently observed window had a high LSV value.

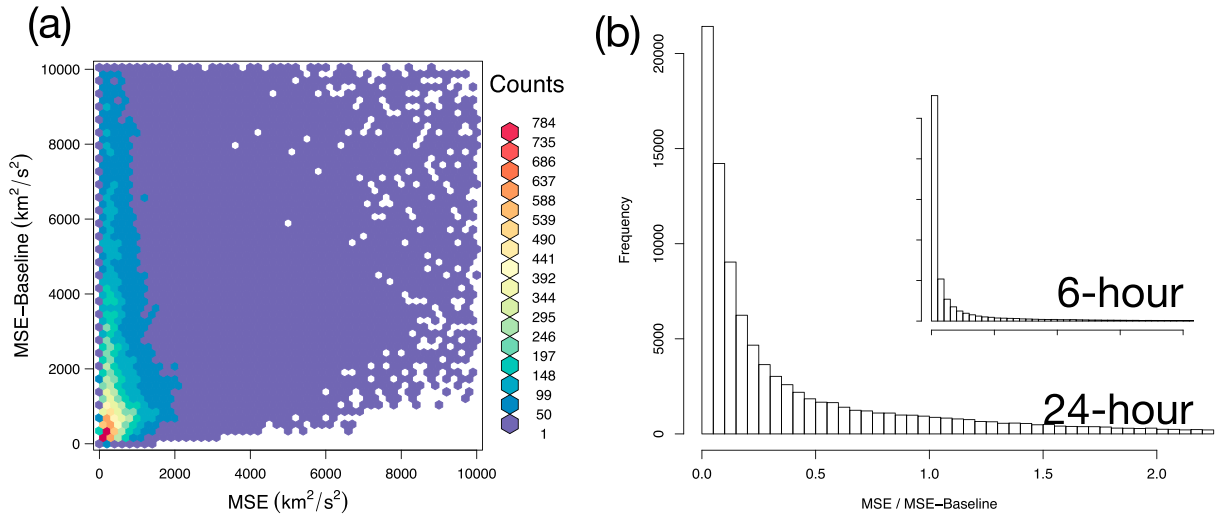

Figure 7. As Figures 4 and 5 but for solar wind speed (v). The inset in Figure 7b shows the histogram for a 6 h window.

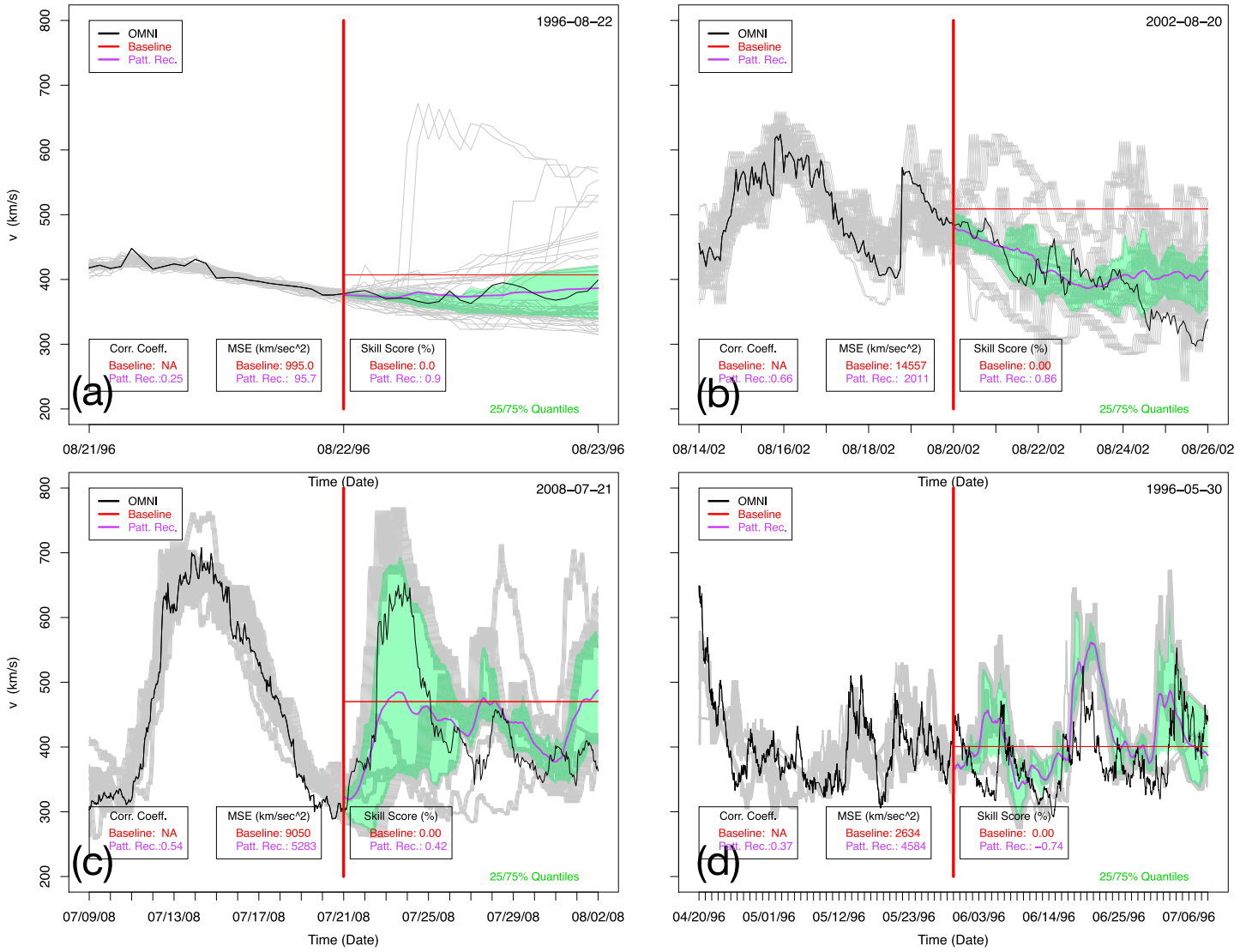

Figure 8. A selection of progressively longer forecasting windows. (a) A 1 day window, (b) a 6 day window, (c) a 12 day window, and (d) a 40 day window. In each panel the black curve leading up to the red vertical line are the data that were observed prior to the forecast. The gray curves are realizations based on pattern recognition of the top 50 intervals most closely matching the observed data. The data following these intervals are used to create the future realizations. The ensemble average of these forecasts is shown in purple, and the actual data observed is shown in black. The red horizontal line is the baseline (persistence) model prediction based on the average observed speed during the previous time window. The green areas mark the 25% and 75% quantiles of the predictions.

Several measures of LSV were investigated, but none were found to be useful. In spite of this, the main message from Figure 6b is that the PR technique, when applied to 6 h windows, usually provides predictions that result in positive correlations (strictly, however, only those correlations above 0.81 are statistically significant for six points, at a significance level of 0.05).

Turning our attention now to the speed of the solar wind, the results and implications are substantially different. Figure 7 summarizes the same parameters as in Figures 4 and 5 but for solar wind velocity, v. The PR model consistently outperforms the baseline (persistence) model as suggested by the fact that MSE values are consistently less than MSE-Baseline values (Figure 7a). The majority of points are found to be less than 1000 km^2 s^{-2}, whereas the persistence model MSE values range up to >4000 km^2 s^{-2}. Similarly, looking at the ratio of the two errors (Figure 7b), the vast majority lie in the region: MSE/MSE-Baseline <0.3. The asymmetry is even more pronounced when the window is limited to 6 h (inset in Figure 7b).

The differences between the results of B_n and v translate into measurable improvements in forecast ability. Figure 8 summarizes four intervals of increasing window size, each of which was chosen primarily so as to

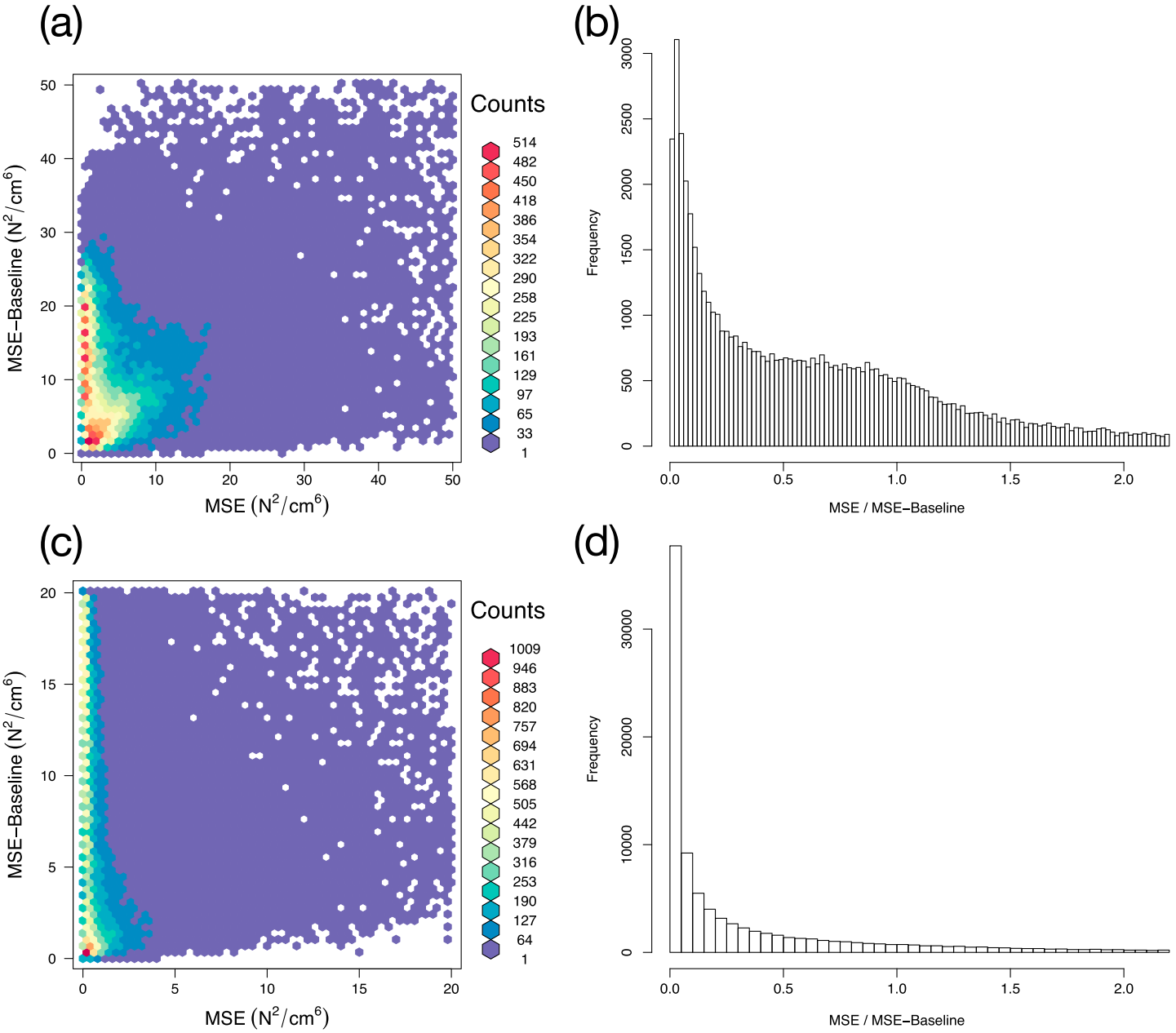

Figure 9. (a and b) As in Figures 4 and 5 but for proton density, n_p. (c and d) As in Figures 8a and 8b but for a 6 h forecasting window.

avoid any transient (CME)-related activity as well as to avoid any intervals with data gaps. Thus, these attempt to forecast primarily ambient solar wind conditions. In Figure 8a, a 1 day window of roughly constant but low speed is forecast to continue for the next day. Note how the 25/75% ranges for the realizations, indicated by the green shading, bracket the actual observations. In Figure 8b, a 6 day window containing two modest streams is correctly predicted to decay during the following 6 days. In Figure 8c a 12 day window with another single stream is predicted to show two modest streams over the next 12 days. And finally, in Figure 8d an apparently more complex stream structure over 40 days is predicted to have three distinct, but modestly high-speed streams during the next 40 days. Overall, each window's prediction is relatively good and, with the exception of Figure 8d, better than the baseline model (persistence), as indicated by the significantly lower values of MSE and positive values for the associated skill scores for Figures 8a–8c. Interval (Figure 8d) is interesting in that while the associated skill score for the PR model is worse than the baseline model, its prediction

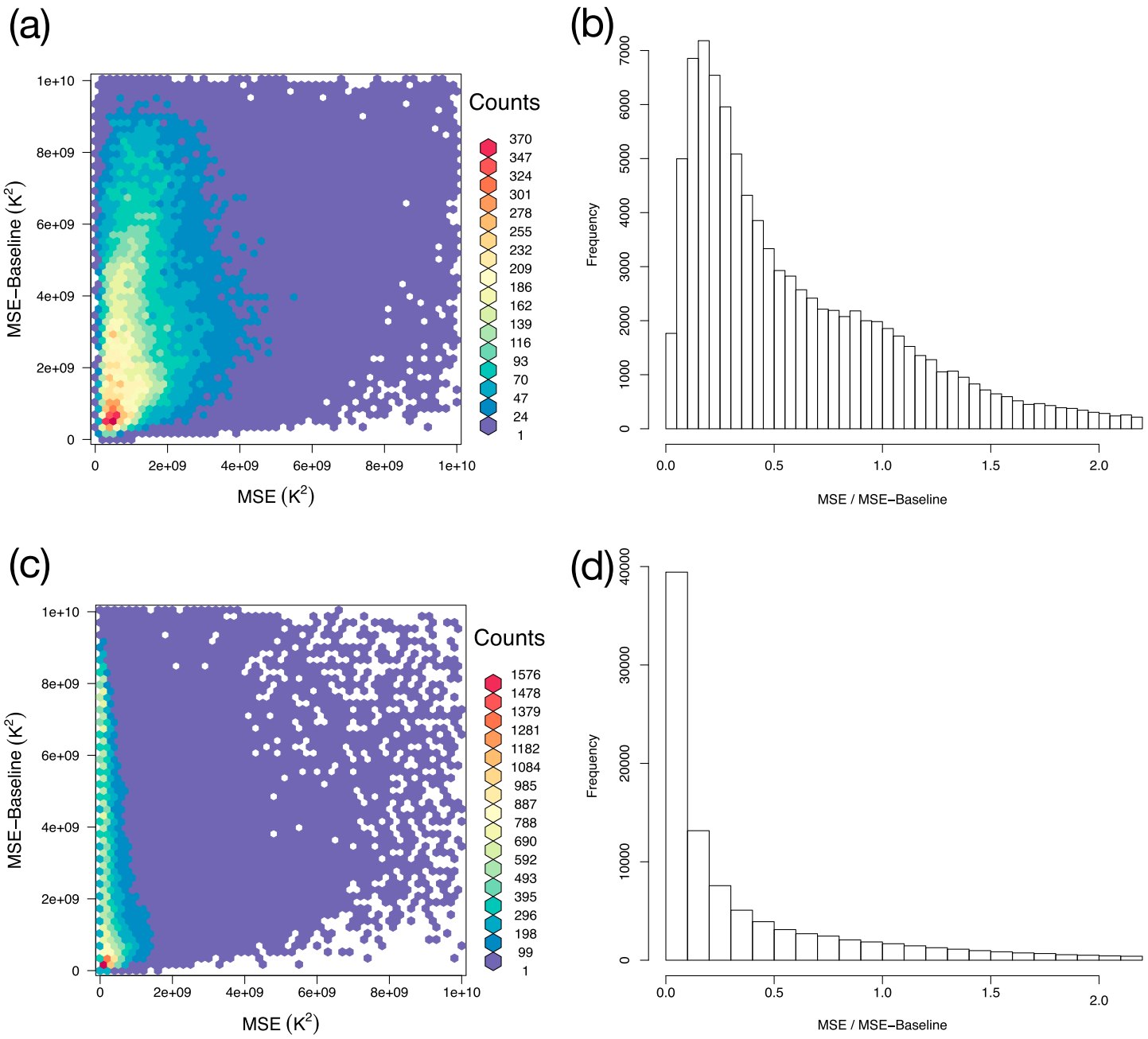

Figure 10. As in Figure 9 but for proton temperature, T_p.

is undoubtedly useful: It correctly predicts a sequence of three high-speed streams, although the exact phasing does not match with observations. It is worth noting that the algorithm does not require that the best realizations are distinct or unique. Thus, one interval, shifted by one or more hours, could serve as the source for several of the realizations. This can be seen in some of the gray traces where the same profile has been slid left or right of another.

We complete our investigation of the PR model by considering proton density and temperature. Figure 9 summarizes the MSE and MSE-Baseline values for proton density for 24 h intervals (Figures 9a and 9b) and 6 h intervals (Figures 9c and 9d). In comparison with the results for B_n and v, we note that the PR model is substantially better than the reference model for both window durations, although there is not as much improvement as was the case for v. Similarly, the distribution of error ratios for T_p (Figure 10) lies between

those of B_n and n_p. In summary then, the order of improvement that the PR model provides over the reference (zero/persistence) mode is v, n_p, T_p, and B_n. This mimics what we have found using MHD models to predict stream structure [*Riley et al.*, 2001, 2012a].

3. Summary and Discussion

In this study, we have outlined a simple pattern recognition technique that may prove useful as a tool for forecasting the properties of the solar wind on the time scale of hours to days. Our results suggest that this approach is a potentially powerful predictor for the bulk solar wind flow velocity, density, and perhaps temperature. However, its use in predicting B_n may be limited to intervals with large-scale variations in the magnetic field, which preferentially occur during the passage of magnetic clouds over the spacecraft. It is not yet clear whether the limited improvement over the baseline (zero) model will yield useful or actionable predictions for B_n.

Our analysis has relied on several assumptions and approximations that deserve consideration. First, we have assumed that past variations are indicative of future variations. To demonstrate that this is the case, we compared the model forecasts with a persistence model, essentially asking whether the predicted forecast was better than assuming some constant value based on historical data. For most of the data sets, this was true. However, for the key B_n data, this was only marginally demonstrated and only for forecast windows of 6 h. Second, we implicitly assumed time stationarity of the data. That is, that the variations in the data 30 or 40 years ago were comparable to variations that we observe today. The unusually quiescent conditions over the last decade or so suggest that this may not be the case [e.g., *Riley et al.*, 2012b]. Moreover, some studies suggest that we may be entering into a grand minimum interval, lasting 40 years [*Lockwood et al.*, 2009]. Third, we did not distinguish between temporal and stationary structures in these data. While most scientific studies would endeavor to make such a classification, here it is benefit of the PR technique that it does not require knowledge of what processes are driving the variations used to make the forecasts.

One potentially significant improvement to the technique outlined here is to employ dynamic time warping (DTW) to the data. Essentially, for each interval that is being compared to the observed interval, a nonlinear stretch is applied to the time axis. This may make sense for solar wind measurements, at least during periods of CMEs, where we envisage a simple flux rope structure becoming increasingly deformed through interactions with the ambient solar wind, as it propagates away from the Sun. Thus, a symmetric CME that is coasting along with the background solar wind could, in principle, be matched with a highly deformed fast CME that is also driving a shock and creating a sheath region. It may be possible for DTW to unravel this deformation by stretching the compressed region or shortening the rarefaction region within such events. We have begun a study that seeks to investigate this.

Although the examples we have presented show considerable promise, it is not yet clear how robust they are, nor how actionable the information they provide might be. We are currently developing a real-time version of the algorithm that will run at PSI's website. From this, we will be able to assess its possible value as an operational tool. We are planning a more extensive study that will, hopefully, find independent criteria for identifying those intervals that are likely to produce predictable future intervals.

Thus far, our limited attempts to identify a parameter capable of capturing the predictability of the interval just observed have not been obviously successful. In this study, we focused on attempts to capture a measure of large-scale variations (LSVs) during the window preceding the prediction. This was intended as a way to identify current conditions (say in the last 24 h) that would be amenable to forecasting. Without it, the best metric for estimating the accuracy of the current forecast lies in the breadth of the confidence intervals. If large, the prediction is highly uncertain, but, if well constrained, the prediction is likely to be more accurate. However, the degree to which this holds remains to be tested.

In closing, we have outlined here a simple technique for providing forecasts that, at least for some parameters, are measurably better than our baseline model. It is unlikely that even sophisticated refinements to this approach will take us beyond incremental improvements, given the underlying complexity in the solar wind data. On the other hand, this model, even as currently implemented, may provide limited forecasts of value. And, even if not, it serves as a new, higher baseline against which future models should be compared.

Acknowledgments
P.R. and M.B.N. would like to acknowledge NASA's Living with a Star program for support in undertaking this study (grant NNX15AF39G). M.O. was funded by STFC grant ST/M000885/1. We acknowledge the use of NASA/GSFC's Space Physics Data Facility's COHOWeb service (http://omniweb.sci.gsfc.nasa.gov/coho/) in retrieving the OMNI_M data. The algorithm used to generate the results of this study is running in real time at PSIs website (http://www.predsci.com/project-zed). Additionally, all data and code used to create the figures contained herein can be obtained from GitHub: https://github.com/predsci/projectzed.

References

Arge, C. N., and V. J. 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, doi:10.1029/1999JA900262.

Board, S. S., et al. (2012), *Earth Science and Applications From Space: A Midterm Assessment of NASA's Implementation of the Decadal Survey*, Natl., Acad. Press, Washington, D. C.

Borovsky, J. E., M. Hesse, J. Birn, and M. M. Kuznetsova (2008), What determines the reconnection rate at the dayside magnetosphere?, *J. Geophys. Res.*, *113*, A07210, doi:10.1029/2007JA012645.

Bothmer, V., and R. Schwenn (1998), The structure and origin of magnetic clouds in the solar wind, *Ann. Geophys.*, *16*, 1–24.

Cassak, P., and M. Shay (2007), Scaling of asymmetric magnetic reconnection: General theory and collisional simulations, *Phys. Plasmas*, *14*(10), 102,114.

Chen, J., P. J. Cargill, and P. J. Palmadesso (1996), Real-time identification and prediction of geoeffective solar wind structures, *Geophys. Res. Lett.*, *23*(6), 625–628.

Dungey, J. W. (1961), Interplanetary magnetic field and the auroral zones, *Phys. Rev. Lett.*, *6*(2), 47–48.

Farrell, P. (2011), New space weather forecasting model going operational with National Weather Service. [Availabe at http://www.bu.edu/cas/news/press-releases/cism/.]

Fränz, M., D. Burgess, and T. Horbury (2000), Magnetic field depressions in the solar wind, *J. Geophys. Res.*, *105*(A6), 12,725–12,732, doi:10.1029/2000JA900026.

Harvey, B. (2007), *Russian Planetary Exploration: History, Development, Legacy and Prospects*, Springer Science & Business Media, New York, and Philadelphia, Pa.

Horbury, T., A. Balogh, R. Forsyth, and E. Smith (1995), Observations of evolving turbulence in the polar solar wind, *Geophys. Res. Lett.*, *22*(23), 3401–3404.

Horbury, T., M. Forman, and S. Oughton (2005), Spacecraft observations of solar wind turbulence: An overview, *Plasma Phys. Controlled Fusion*, *47*(12B), B703–B717.

Jackson, B., P. Hick, A. Buffington, H.-S. Yu, M. Bisi, M. Tokumaru, and X. Zhao (2015), A determination of the north–south heliospheric magnetic field component from inner corona closed-loop propagation, *Astrophys. J. Lett.*, *803*, L1.

King, J. H., and N. E. Papitashvili (2005), Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data, *J. Geophys. Res.*, *110*, A02104, doi:10.1029/2004JA010649.

Lockwood, M., A. P. Rouillard, and I. D. Finch (2009), The rise and fall of open solar flux during the current grand solar maximum, *Ap. J.*, *700*, 937–944, doi:10.1088/0004-637X/700/2/937.

Lugaz, N., C. J. Farrugia, R. M. Winslow, N. Al-Haddad, E. K. J. Kilpua, and P. Riley (2016), Factors affecting the geoeffectiveness of shocks and sheaths at 1 AU, *J. Geophys. Res. Space Physics*, *121*, 10,861–10,879, doi:10.1002/2016JA023100.

Mays, M. L., et al. (2015), Ensemble modeling of SMEs using the WAS–ENLIL+Cone model, *Sol. Phys.*, *290*(6), 1775–1814.

Nieves-Chinchilla, T., M. Espinosa, C. Cid, M. A. Hidalgo, and J. Sequeiros (2002), A new model for the magnetic topology of magnetic clouds, in *Solar Variability: From Core to Outer Frontiers*, vol. 506, edited by A. Wilson, pp. 25–28, ESA Spec. Publ., Prague.

Odstrcil, D. (1993), Improved FCT algorithm for shock hydrodynamics, *J. Comp. Phys.*, *108*(2), 218–225.

Owens, M., R. Wicks, and T. Horbury (2011), Magnetic discontinuities in the near-Earth solar wind: Evidence of in-transit turbulence or remnants of coronal structure?, *Sol. Phys.*, *269*(2), 411–420.

Owens, M. J., P. J. Cargill, C. Pagel, G. L. Siscoe, and N. U. Crooker (2005), Characteristic magnetic field and speed properties of interplanetary coronal mass ejections and their sheath regions, *J. Geophys. Res.*, *110*, A01105, doi:10.1029/2004JA010814.

Owens, M. J., T. Horbury, R. Wicks, S. McGregor, N. Savani, and M. Xiong (2014), Ensemble downscaling in coupled solar wind-magnetosphere modeling for space weather forecasting, *Space Weather*, *12*(6), 395–405, doi:10.1002/2014SW001064.

Parker, E. N. (1958), Dynamics of the interplanetary gas and magnetic fields, *Astrophys. J.*, *128*, 664.

Pizzo, V., C. Koning, M. Cash, G. Millward, D. Biesecker, L. Puga, M. Codrescu, and D. Odstrcil (2015), Theoretical basis for operational ensemble forecasting of coronal mass ejections, *Space Weather*, *13*(10), 676–697, doi:10.1002/2015SW001221.

Reiss, M. A., M. Temmer, A. M. Veronig, L. Nikolic, S. Vennerstrom, F. Schöngassner, and S. J. Hofmeister (2016), Verification of high-speed solar wind stream forecasts using operational solar wind models, *Space Weather*, *14*(7), 495–510, doi:10.1002/2016SW001390.

Riley, P., J. T. Gosling, L. A. Weiss, and V. J. Pizzo (1996), The tilts of corotating interaction regions at midheliographic latitudes, *J. Geophys. Res.*, *101*(A11), 24349–24357.

Riley, P., J. A. Linker, and Z. Mikić (2001), An empirically-driven global MHD model of the corona and inner heliosphere, *J. Geophys. Res.*, *106*, 15,889, doi:10.1029/2000JA000121.

Riley, P., J. A. Linker, R. Lionello, and Z. Mikic (2012a), Corotating interaction regions during the recent solar minimum: The power and limitations of global MHD modeling, *J. Atmos. Sol. Terr. Phys.*, *83*, 1–10, doi:10.1016/j.jastp.2011.12.013.

Riley, P., R. Lionello, J. A. Linker, Z. Mikic, J. Luhmann, and J. Wijaya (2012b), Global MHD modeling of the solar corona and inner heliosphere for the whole heliosphere interval, *Solar Phys.*, *274*, 361–3775, doi:10.1007/s11207-010-9698-x.

Savani, N., A. Vourlidas, A. Szabo, M. Mays, I. Richardson, B. Thompson, A. Pulkkinen, R. Evans, and T. Nieves-Chinchilla (2015), Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 1. Initial architecture, *Space Weather*, *13*(6), 374–385, doi:10.1002/2015SW001171.

Savani, N., A. Vourlidas, I. Richardson, A. Szabo, B. Thompson, A. Pulkkinen, M. Mays, T. Nieves-Chinchilla, and V. Bothmer (2017), Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 2. Geomagnetic response, *Space Weather*, *15*, 441–461, doi:10.1002/2016SW001458.

Shimojo, M., and K. Shibata (2000), Observational evidence of magnetic reconnection in solar X-ray jets, *Adv. Space Res.*, *26*(3), 449–452.

Thernisien, A., A. Vourlidas, and R. Howard (2009), Forward modeling of coronal mass ejections using STEREO/SECCHI data, *Sol. Phys.*, *256*(1–2), 111–130.

Ulrich, R. K., and T. Tran (2016), Generation of a north/south magnetic field component from variations in the photospheric magnetic field, *Sol. Phys.*, *291*(4), 1059–1076.

Wang, Y.-M. (1994), Two types of slow solar wind, *Astrophys. J. Lett.*, *437*, L67–L70, doi:10.1086/187684.

Wang, Y. M., and N. R. Sheeley (1990), Solar wind speed and coronal flux-tube expansion, *Astrophys. J.*, *355*, 726–732, doi:10.1086/168805.