Time-window approaches to space-weather forecast metrics: a solar wind case study

Article

Accepted Version

Owens, M. J. ORCID: https://orcid.org/0000-0003-2061-2453 (2018) Time-window approaches to space-weather forecast metrics: a solar wind case study. Space Weather, 16 (11). pp. 1847-1861. ISSN 1542-7390 doi: 10.1029/2018sw002059 Available at https://centaur.reading.ac.uk/80649/

Time-Window Approaches to Space-Weather Forecast Metrics: A Solar Wind Case Study

Mathew J. Owens[1]

[1]Space and Atmospheric Electricity Group, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, UK

ORCID: 0000-0003-2061-2453

Correspondence to m.j.owens@reading.ac.uk

Abstract

Metrics are an objective, quantitative assessment of forecast (or model) agreement with observations. They are essential for assessing forecast accuracy and reliability, and consequently act as a diagnostic for forecast development. Partly as a result of limited spatial sampling of observations, much of space-weather forecasting is focused on the time domain, rather than inherent spatial variability. Thus metrics are primarily "point-by-point" approaches, in which observed conditions at time *t* are compared directly (and only) with the forecast conditions at time *t*. Such metrics are undoubtedly useful. But in lacking an explicit consideration of timing uncertainties, they have limitations as diagnostic tools and can, under certain conditions, be misleading. Using a near-Earth solar wind speed forecast as an illustrative example, this study briefly reviews the most commonly-used point-by-point metrics and advocates for complementary "time window" approaches. In particular, a scale-selective approach, originally developed in numerical weather prediction for validation of spatially patchy rainfall forecasts, is adapted to the time domain for space-weather purposes. This simple approach readily determines the time scales over which a forecast is and isn't valuable, allowing the results of point-by-point metrics to be put in greater context.

Point 1: Forecast timing errors can complicate the interpretation of point-by-point metrics

Point 2: Ideally, all forecasts would include an estimate of forecast uncertainty

Point 3: A scale-selective analysis can determine the time scales over which a forecast is valuable

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2018SW002059

1. Introduction

When determining how well a space-weather forecast[1] performs, human assessment can rapidly scrutinise a large number of facets: simply looking over the observations and forecast gives an immediate "feel" for what features are reproduced and missed, how the general structure differs, over what temporal/spatial scales the forecast is applicable, whether the forecast exhibits any obvious bias, performs better within certain parameter regimes, etc. But this is inherently subjective, qualitative, lacking in repeatability and simply infeasible for large volumes of data. Metrics are an automated, objective quantification of forecast performance relative to observations. As such, metrics are vitally important not just for validation[2] of space-weather forecasts [e.g., *Spence et al.*, 2004], but also as a diagnostic tool to inform future forecast development. Different metrics quantify different, specific qualities of a forecast. Thus while there are no right or wrong metrics *per se*, it is nevertheless essential to select a metric which actually measures the features of interest (this, as will be seen in the subsequent examples, is not always as straightforward as it seems). Changes to a forecast scheme made on the basis of a poorly chosen metric can potentially reduce its usefulness for an end-user (though of course the chosen metric will measure an improvement).

The space-weather community is in the process of adopting both more sophisticated forecast approaches and metrics with enhanced diagnostic capability [e.g. *Jian et al.*, 2016; *Murray et al.*, 2017; *Murray*, 2018]. Many of these approaches have been adapted from numerical weather prediction (NWP) [*Siscoe*, 2007]. In NWP, there is extensive coverage by the observation network, allowing both spatial and temporal agreement to be explicitly treated. Extremely sparse observational sampling of the Sun-Earth system, however, means space-weather forecast validation is often primarily concerned with the time domain (though errors in the time domain may well result from spatial variations). For example, while forecasts of the solar wind (such as the example of near-Earth solar wind shown in Section 2.1) cover the largest spatial domain within the Sun-Earth system, they are typically validated solely against single-point *in situ* observations made in near-Earth space [e.g. *Owens et al.*, 2008; *MacNeice*, 2009; *MacNeice et al.*, 2018]. Consequently, validation is primarily focussed on a point-by-point analysis: The observed conditions at time *t* are compared directly (and only) with the forecast conditions at time *t*. As is illustrated in Sections 2.2 and 2.3, such approaches inflict a "double penalty" for timing offsets in forecast events, due to both missing the event and generating a false alarm. On the one hand, this is a legitimate assessment of the forecast. On the other hand, it does not always provide a useful diagnostic of the forecast, and many operators will tolerate relatively small errors in event timing if the general outlook is correct. One solution is for forecasts to include a measure of their own uncertainty, as illustrated in Section 2.5. However, this is not always practical. Thus in addition to point-by-point metrics, it may be advantageous to also employ "time window" metrics. One useful approach, outlined in Section 3.1, is to specify criteria for discrete features within forecast and observation time series and to compare feature correspondence, including the timing. However, such feature specification requires *a priori*

[1] Throughout this study, metrics are discussed with regards to "forecasting", though the same issues and principles apply for general model diagnostics. Consequently, anywhere the term "forecast" appears, the term "model" could be directly substituted.

[2] As in the majority of the space-weather literature, the term "validation" is here used to refer to the process of comparing forecasts and observations to establish accuracy and truth of the forecast. This is often referred to as "verification" in meteorology.

knowledge of the properties of interest, as well as repeatable signatures in said features, both from event to event and across forecast and observation data. Thus in Section 3.2 a more feature-agnostic approach is proposed, based upon NWP validation of rain forecasts. It compares forecasts and observations at a range of different spatial scales and is here adapted to the time domain as a space-weather forecast metric. It is shown that this analysis provides a useful assessment of the time scales over which a forecast is and isn't valuable.

2. Point-by-Point Metrics

2.1 Example Forecast

In order to illustrate the strengths and limitations of different metrics, an example forecast is considered. The black line in Figure 1a shows hourly near-Earth solar wind speed (*V*) for Carrington rotation (CR) 2049, spanning mid-October to mid-November 2006. Data are from the *Omni* dataset of near-Earth spacecraft measurements [*King and Papitashvili*, 2005]. CR 2049 was chosen as there are three distinct high-speed enhancements (HSEs) on 20 October, 28 October and 9 November.

An illustrative forecast was produced using the "Magnetohydrodynamics Around a Sphere" [MAS; *Linker et al.*, 1999; *Riley et al.*, 2012] global coronal model. The inner boundary conditions are set by the observed photospheric magnetic field for CR 2049. Model output is available from http://www.predsci.com/mhdweb/. Typically, the MAS solution would be propagated to near-Earth space with a numerical magnetohydrodynamic solar wind model and the forecast *V* extracted from the model grid point closest to Earth. Here, however, for the purposes of demonstration, the solution was perturbed to (retrospectively) produce a closer match to the observations. Specifically, the model solar wind at 30 solar radii was sampled 5-degrees above the sub-Earth point, as this was found to improve the representation of the HSE on 28 October. The solar wind speed was then propagated from 30 solar radii to Earth using a simple "upwind" technique [*Owens and Riley*, 2017] to produce the time series shown in red in Figure 1a.

2.2 Error Functions

Forecasts are commonly assessed using simple error functions (otherwise called cost or loss functions). The results for CR2049 are summarised in Table 1. For solar wind speed, the mean-square error (MSE) is given by:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}[V_F(t) - V(t)]^2$$

where $V_F(t)$ and $V(t)$ are the forecast and observed solar wind speeds at time *t*, respectively, and *T* is the total number of time points considered. Smaller MSE values indicate better agreement, with zero being a perfect forecast. For the forecast shown in Figure 1a, the MSE is 1.30x10^4 km^2 s^{-2}. This is usually converted to root-mean-square (RMS) error:

$$RMS = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[V_F(t) - V(t)]^2}$$

RMS has the advantage of being a linear measure of the magnitude of the errors with the same units as the parameter of interest. The RMS error for the forecast is 114 km s^{-1}.

In isolation, these values say relatively little about the quality of the forecast. Metrics are most useful as a comparative tool. Thus it is instructive to also consider a second solar wind speed prediction. The blue line shows the average *V* for CR 2049, 432 km s^{-1}. For validation purposes, this climatological mean would be a poor choice of comparison prediction, as it has zero variability. In practice, it would be preferable to use another simple forecast, such as 27-day recurrence [*Owens et al.*, 2013]. But for the purposes of illustrating certain issues, the climatological mean is useful here. The MSE between the observed *V* for CR 2049 and the climatological mean is 0.98x10^{4} km^{2} s^{-2}, while the RMS 98.9 km s^{-1}, both smaller than the forecast values.

An alternative measure of a similar property is the mean absolute error (MAE):

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |V_F(t) - V(t)|$$

For the *V*, MAE is essentially the same as for the forecast (84.6km s^{-1}) and the climatological mean (85.0 km s^{-1}).

In order to further put error functions in perspective, the “skill” of a forecast is calculated as:

$$Skill = \ 1 - \frac{MSE}{MSE_{REF}}$$

where MSE_{REF} is the MSE of a reference “baseline” model, such as the climatological mean. Skill is negative when the forecast is worse than the baseline, 0 when they are equal, and 1 for a perfect forecast. (Sometimes skill is further multiplied by 100 to express it as a percentage of a perfect forecast.) By comparing directly with a baseline model, skill potentially allows disambiguation between bad forecasts and periods/situations which are inherently difficult to forecast. For the forecast shown in Figure 1a, using the climatological mean as the reference, the forecast skill is -0.32. Thus the forecast is deemed to be “worse” than assuming the solar wind is always a constant 432 km s^{-1}.

Thus the general conclusion from these error functions for this example period is that the climatological mean is at least as “good” as the forecast for CR 2049. This is, of course, an entirely correct and fair assessment. But it is obvious that it does not tell the whole story; the climatological mean lacks sharpness and discrimination, in that it does not reconstruct any of the features of the solar wind structure. It would be useless as a predictive tool for almost all applications and thus could be said to lack value. In contrast, the forecast appears to work quite well for this interval: By eye, it can be seen that the forecast produces three HSEs, as observed, and they are of comparable magnitudes and durations to the observations. By inspection of the time series, it can be seen that the error functions for the forecast are relatively high due to the approximately 1 to 2 days errors in the timings of the HSEs, which result in the “double penalty” of first over-predicting *V,* closely followed by under-predicting *V.* But, depending on the application, the forecast may well still be regarded as valuable in that it enables users to make decisions which lead to beneficial outcomes [*Murphy*, 1993].

In this particular example, other forms of point-by-point comparisons *are* able to discriminate between the predictive value of the forecast and climatological mean (see Section 3.3 for an example where this is not the case). While not strictly an error function, Pearson (or linear) correlation, r_L, is often used in a similar manner to RMS and MAE to quantify forecast and observation agreement, where:

$$r_L = \frac{\sum_{t=1}^{T}[V_F(t) - \overline{V_F}]\,[V(t) - \bar{V}]}{\sqrt{\sum_{t=1}^{T}[V_F(t) - \overline{V_F}]^2}\,\sqrt{\sum_{t=1}^{T}\,[V(t) - \bar{V}]^2}}$$

It is weakly positive for the forecast (r_L = 0.28). Spearman correlation, r_S, replaces the observed and forecast values at time *t* with their ranks within their respective distributions. As a result, r_S is less susceptible to outliers then r_L. It is effectively zero (r_S = 0.06) for the forecast. The zero variance of the climatological mean results in both $r_L = 0$ and $r_S = 0$. Figure 2 summarises these results in the form of a Taylor diagram [*Taylor*, 2001; *Riley et al.*, 2013a]. It displays the RMS (centred by the mean values to remove forecast bias) and linear correlation between forecast and observation, along with the standard deviation of the time series under consideration. In short, the closer the forecast (red point) to the observation (black circle), the better. Thus while the Taylor diagram does not strictly conclude that the forecast is superior to the climatological mean (blue point), the issues with the latter as predictive tool are immediately obvious. For a more realistic "baseline" forecast, this may not always be the case.

In addition to potentially misleading forecast assessment, error functions can also have unintended consequences for model development. *Riley et al.* [2013a] note that changes to their coronal model which wipe out all solar wind speed variability (and thus value of the resulting forecast) are not reflected in RMS, which is essentially unchanged. Similarly, any forecast scheme trained to minimise RMS or MAE may tend preferentially towards a conservative, climatological-mean-like prediction, rather than a valuable forecast.

2.3 Binary Metrics

As error functions quantify the magnitude of forecast deviation from observations at every time step, they can have limitations as diagnostic tools. Firstly, by considering every time step equally, rather than focussing on specific times or parameter ranges of interest, these metrics can be skewed towards measuring properties that are inconsequential to an operator. E.g., whether the forecast correctly reproduces the details of the slow-speed wind may be unimportant, but is given equal weighting to the times of high speeds, which are important. Secondly, large outliers can have a relatively strong influence on error functions and especially on linear correlation. In some circumstances, this will be appropriate, as the magnitude of the extremes is of interest. In other circumstances, this may be less critical, as what matters is whether or not a given threshold is exceeded, not by how much. To address these issues, an alternative approach is to consider each time step as a binary "yes/no" state and compare observations and forecasts on this basis. For probabilistic forecasts, discussed further in Section 2.5, this also involves setting a probability threshold, in addition to an event-definition threshold.

The black dashed line in Figure 1a shows a threshold of $V > 500$ km s^{-1} used to define hourly "events" in the forecast and observed time series. Figure 1b displays the timing of the subsequent forecast and observed events, sorting them into one of four categories; true positives (*TP*, or "hits"; hours for which both observed and forecast events are present), false positives (FP or "false alarms"; hours for which an event is forecast but not observed), false negatives (FN, or "misses"; hours for which an event is observed but not forecast) and true negatives (TN; hours for which both observation and forecast have no event). The occurrence of these classifications is summarised in a contingency table [e.g., *Finley*, 1884; *Murphy*, 1996], shown as Table 2 for the forecast and Table 3 for the climatological mean. The forecast produces approximately the correct number of events ($P_F = 177$ versus $P = 192$ observed) and non-events ($N_F = 478$, versus $N = 463$ observed), meaning it has little bias, whereas the climatological mean produces zero events and over-estimates the non-events ($N_F = 655$). The "double penalty" effect on the forecast is apparent: Because of the timing offset in the HSEs, the forecast produces both FN and FP, whereas the null prediction of the climatological mean only produces FN. For the forecast, the total number of false predictions, $FP + FN$, is 233, while for the climatological mean it is only 192.

From the contingency tables alone, it is not immediately clear whether the forecast is "better" than the climatological mean. It will depend on how FP and FN are weighed relative both to each other and to TP (and to a lesser extent, TN). There are a variety of ways to combine these four numbers, to emphasise different forecast aspects. The full range of combinations is not discussed here (see *Thornes and Stephenson* [2001] and *Reiss et al.* [2016], as well as the World Meteorological Organisation guide: http://www.cawcr.gov.au/projects/verification/). Two of the most useful combinations are the true positive rate ($TPR = TP/P$) and the false positive rate ($FPR = FP/N$), as together they provide a reasonable overview of a forecast. A perfect forecast would have $TPR = 1$ and $FPR = 0$. For the forecast of CR 2049, $TPR = 0.35$ and $FPR = 0.24$.

For events defined by $V > 500$ km s^{-1}, the climatological mean results in no true or false positives and so $TPR = 0$ and $FPR = 0$. If events were defined using a V threshold lower than the climatological mean (e.g. $V > 400$ km s^{-1}), it would produce a prediction of events at all times, giving $TPR = 1$ and $FPR = 1$. Thus for any event threshold, the climatological mean over the period under consideration gives $TPR = FPR$. When a forecast results in $TPR > FPR$, it is superior to the climatological mean in being able to predict the occurrence of events and non-events.

2.4 Forecast Summaries

Binary metrics depend on the choice of both event and probability thresholds, and thus ways to summarise parameter space are necessary. The (often complex) relation between FPR and TPR for a range of event thresholds is captured by the receiver operator characteristic [ROC; *Peterson et al.*, 1954; *Mason*, 1982] curve in Figure 1c. This technique is commonly used for validation of probabilistic forecasts at a range of probability thresholds (see Section 2.5), including solar flare forecasts [*Murray et al.*, 2017; *McCloskey et al.*, 2018]. However, it can also be used to summarise the deterministic V forecast. In this example , all event thresholds result in $TPR > FPR$ (i.e., are above the $y=x$ line in Figure 1c) except $V > 600$ km s^{-1}, where the double penalty is strongest. The ROC can be further distilled down to the area under the curve, integrated along the horizontal axis [AUC; *Mason and Graham*, 2002]. AUC represents a forecast's ability to correctly anticipate events and non-events (1 being a perfect forecast, 0.5 being equal to the climatological mean). For the V forecast, the AUC is 0.68.

An alternative summary can be provided by the Cost-Loss analysis [*Murphy*, 1977; *Richardson*, 2000], which determines the benefit an operator would gain from acting on a forecast. The real strength of

Cost-Loss analysis is in the evaluation of probabilistic forecasts (see Section 2.5), as it explicitly accounts for the fact that different operational uses will act on the same forecast in a different manner. E.g., if a forecast gives a low probability of a space-weather event, an operator may still choose to take mitigating action if the cost of doing so (e.g. from lost revenue), *C*, is small relative to *L*, the cost of being caught unprepared by a damaging event. In such situations, forecasts which minimise missed events, even if this means increased false alarms, are more desirable. Conversely, if *C* is a significant fraction of *L*, an operator is unlikely to act on the basis of a low forecast probability. In such circumstances, forecasts which minimise false alarms are more desirable. This analysis has recently been applied to validation of probabilistic solar wind forecasts [*Owens et al.*, 2014; *Owens et al.*, 2017].

Figure 1d shows how the potential economic benefit of acting on the determinsitic forecast of *V* for a range of *C/L* values and for events defined by a range of *V* thresholds. Potential economic benefit is measured relative to the climatological probability of an event, so that values below 0% indicate the forecast is less useful than climatology and 100% indicates a perfect (deterministic) forecast. As shown by the ROC curve, most benefit is gained at intermediate solar wind speeds (400 to 500 km s^{-1}) and for low *C/L* scenarios. When false alarms become costly, the forecast ceases to add value, as the double penalty effect comes into play. Despite the insight gained from binary metrics such as ROC and Cost/Loss analysis, they nevertheless operate on a strictly point-by-point comparison basis and do not account for timing errors/uncertainty. As illustrated in Section 3.3, the resulting double penalty issue is even stronger for B_Z forecasts, which are critical for space weather [*Dungey*, 1961], as large-scale B_Z variations tend to be bipolar in nature.

2.5 Validating Probablistic Forecasts

Ideally, a forecast would include an assessment of forecast uncertainty. Figure 3a shows an example of a probabilistic forecast of solar wind speed for CR2049. It was generated using a perturbed initial condition ensemble [*Owens and Riley*, 2017]. The RMS and MAE of the forecast ensemble median are comparable to the deterministic *V* forecast shown in Figure 1a. But what is of most interest here is the uncertainty estimate. Figure 3b shows the probability of $V > 500$ km s^{-1} as a function of time. For the observations, this is either 0 or 1; for the climatological mean, it is always 0; for the forecast ensemble the probability is the fraction of ensemble members for which $V > 500$ km s^{-1} at each time step [e.g., *Slingo and Palmer*, 2011 and references therein]. For the 21 October HSE, the onset timing uncertainty is reasonable, but the forecast is too confident of no event after 22 October. For the 29 October HSE, the forecast clearly underestimates the uncertainty in the HSE arrival time and duration, as the probability peaks more than a day early and remains high (≈ 0.75) for around a day too long. For the 10 November HSE, there is a 3-day spread in the HSE arrival time in the probabilistic forecast, with the peak probability on the 11 November, approximately the time of the observed peak.

In order to produce the ROC curve (Figure 3c), a probability threshold is required to define events at each *V* threshold. In general for CR 2049, higher probability thresholds produce better forecasts as given by AUC (though it is not a simple linear relation). On this basis alone, it may be tempting to conclude that the probabilistic forecast is most beneficial in operational situations where few false alarms are present (i.e., high *C/L* ratios). However, that is not generally the case (as shown in Figure

3c and discussed below). What the ROC is actually revealing is simply that higher probability thresholds reduce the total number of forecast events and, in the presence of timing errors, minimise the double penalties described in the previous section. Thus again, even with probabilistic forecasts, point-by-point metrics can favour overly conservative forecasts.

From Figure 3b, it can be seen that for $V > 500$ km s^{-1}, there are no periods where the forecast probability of an event exceeds around 0.75. This means that operational settings in which forecast certainty is critical (i.e., where false alarms are costly), the forecast will not be useful. This demonstrated in the cost-loss analysis in Figure 3d, where for $V > 500$ km s^{-1}, there is no economic benefit to acting on the forecast when $C/L > 0.5$. At lower speed thresholds, e.g., 400 km s^{-1}, there are times when the forecast correctly predicts 0 probability of an event (6 to 8 November) and 1 probability of an event (22 October). This results in a valuable forecast for higher C/L values, unlike the similar deterministic forecast.

Clearly, forecasts should intrinsically account for uncertainty, including the timing of features. However, forecasts often do a poor job in this respect (as shown in the example above), and uncertainty can be costly to estimate. E.g., estimating the timing uncertainty in a CME forecast through a numerical model ensemble [*Riley et al.*, 2013a; *Riley et al.*, 2013b; *Mays et al.*, 2015] will require a minimum of an order-of-magnitude more computing resources. Additionally, an operator may tolerate a greater timing error than the estimated forecast timing uncertainty. Thus it is also desirable to use metrics which explicitly allow for timing uncertainty.

3 Time-Window Metrics

3.1 Feature-Based Metrics

One approach to dealing with timing uncertainty is to define discrete features (also known as objects or events) on the basis of extended spatial information or time history (rather just using a simple threshold on a point-by point basis, as in the case of binary metrics), and compare their properties, including timing [e.g., *Ebert and Gallus Jr*, 2009]. For example, *Owens et al.* [2005] defined a HSE as a net 100 km s^{-1} increase in V over a 2 day interval in 8-hour smoothed data (computed as the mean in a rolling 8-hour window). The smoothing allows the analysis to be readily applied to both observations and numerical solar wind model output. The HSE lasts as long as these criteria are met, with the characteristic time of the HSE being the time of maximum V gradient. *Reiss et al.* [2016] and *MacNeice* [2009] used similar definitions. Figure 4 shows the analysis applied to the CR2049 observations and forecast. In practice, when applying the analysis to years of data, observed and forecast HSEs are paired up algorithmically. In this instance, there are 3 observed and forecast HSEs, with forecast/observed pairs overlapping in time, so the pairing is trivial. Results are summarised in Table 4.

During this short interval of comparison, the forecast produces approximately the correct magnitude of HSE (in 8-hour smoothed data), but the timing of HSEs is systematically biased early. Clearly, this approach provides quantitative diagnostic information about *why* the RMS and MAE are high for this forecast relative to the climatological mean. The limitation in this kind of analysis is that features of interest have to be rigorously defined *a priori*. For solar wind speed, this is reasonable, but for B_Z, it may be more difficult, particularly regarding time scale and magnitude, as further discussed in Section 3.3.

An alternate approach to timing uncertainties is to consider the peak value within a fixed time window (e.g., maximum *V* in a 24-hour window of 1-hour data). This can provide useful information if, again, tailored to the specific needs of the operational setting. But there are a number of considerations with applying this approach more generally. Firstly, different time windows will, of course, be more or less appropriate for different forecast applications. Secondly, for a fixed time window, the same peak value can result from a single data spike, multiple peaks, or the whole window being elevated. Thirdly, changing the time resolution of the data can affect the peak values in different ways: The peak value of the single data spike will be dramatically reduced, whereas broader peaks will be less affected. A method to effectively summarise this parameter space for a binary forecast is described in the next section.

3.2 Scale-Selective Metrics

In validation of forecasts from numerical weather prediction (NWP), double penalties are also a ubiquitous issue, resulting from both spatial and temporal offsets between forecast and observation. A particularly apposite example is convective rain, which is inherently patchy on the spatial scales measureable by radar and forecastable by NWP. This can lead to misdiagnosis of forecasts if performed on a point-by-point basis at the grid-cell level. Hypothetical rain observations and forecast for a 10x10 grid are shown in Figure 5. The forecast has little bias over the whole domain (forecast and observation predict 18% and 19% of grid points, respectively, will contain rain) and captures much of the large-scale structure, with a front of rain in the bottom-right corner of the domain. There is, however, little correspondence at the individual grid-point level. Making a simple point-by-point comparison of the forecast and observations reveals *FPR* > *TPR,* meaning it performs worse than climatology. In fact, even a completely null prediction, where rain is never predicted anywhere, is found to be superior in this instance.

Roberts and Lean [2008] and *Roberts* [2008] suggest a scale-selective approach to address this issue. This considers how well the forecast captures the observed rain on increasing larger spatial scales, or "neighbourhood sizes", *n* [Theis et al., 2005]. In the example shown in Figure 5, the available neighbourhood sizes would be *n*=1 (where each neighbourhood is one grid point, resulting in the original distribution of observed and forecast rain), to *n*=2 (where each neighbourhood contains 2x2 grid points), *n*=5 (25 grid points) and *n*=10 (100 grid points, the entire domain). At each *n*, the fraction, *f*, of grid points within each neighbourhood which contains rain is computed. For *n*=1, each *f* will be either exactly 0 or exactly 1. For higher values of *n*, *f* will take a value between 0 and 1. For the example shown, at *n*=10 the observed *f* = 0.19, while the forecast f_F = 0.18. For each n, the fraction MSE, fMSE, can be computed:

$$fMSE(n) = \frac{1}{N_x\, N_y} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} [f(x,y) - f_F(x,y)]^2$$

where *x* and *y* are the neighbourhood number in the x- and y-directions and N_X and N_Y are the total number of neighbourhoods in the x- and y-directions, respectively. Thus for the example shown, N_X =

$N_Y = 10/n$. The fraction skill score (FSS) is computed by comparing the forecast fMSE with the fMSE of a reference (or baseline) forecast, in this case the null rain forecast:

$$FSS(n) = 1 - \frac{fMSE(n)}{MSE_{REF}(n)}$$

Figure 5d shows how the FSS varies with neighbourhood size, *n*. As discussed above, at *n*=1 FSS is negative as the total number of false grid points (i.e., *FN* + *FP*) is higher for the forecast than for the null prediction. But as neighbourhood size increases, FSS becomes increasingly positive, as the forecast captures the large-scale spatial structure of the observed rainfall. At *n*=10, FSS approaches 1 as the forecast bias is very low, whereas the null prediction bias is high. The overall conclusion is that if an operator is interested in spatial scales greater than those represented by single grid points, the forecast is valuable (relative to a null forecast).

This same scale-selective approach can be adapted to the time domain for space-weather purposes. For the *V* time series, the fMSE for neighbourhood size *n* becomes:

$$fMSE(n) = \frac{n}{T} \sum_{t=1}^{T/n} [f(t) - f_F(t)]^2$$

where *f(t) and* $f_F(t)$ are the fraction of observed and forecast hours in time bin *t* for which *V* > 500 km s^{-1}. The top panel of Figure 6 shows the observed and forecast *f* as a function of time for the CR 2049 solar wind speed, with events (red) and non-events (blue) defined using *V* > 500 km s^{-1}. At this 1-hour neighbourhood size, this is equivalent to the original point-by-point analysis (i.e., the same as Figure 1b) and *f* is either exactly 0 or 1. The fMSE of the forecast is 0.370, whereas for the climatological mean, fMSE= 0.3048. Thus for *n* = 1, FSS = -0.21.

The second panel of Figure 6 shows a neighbourhood size of 45 hours. There are still neighbourhoods with *f* = 0 and *f* = 1, but there are now also intermediate values. By eye, the agreement is still far from perfect, but the "smearing" of events in time means that there are fewer intervals which are so starkly wrong, i.e., where $|f - f_F| = 1$. The fMSE of both the forecast and climatological mean have dropped (to 0.184 and 0.230, respectively) and the FSS is now weakly positive (0.2). The third and fourth panels show a neighbour sizes of 105 and 210 hours, respectively. The agreement between forecast and observation has been greatly enhanced, though at these long temporal scales, a lot of information has also been lost.

Figure 7a shows how the FSS varies with *n* and *V* thresholds. In order to avoid aliasing between features in the *V* time series and the neighbourhood boundaries, the boundaries are slid across the time series to consider all possible neighbourhood combinations for a given value of *n*. The mean FSS for a given *n* is shown. For the CR2049, the *V* forecast is generally most valuable for lower *V* thresholds. However, at the very lowest threshold, *V* > 350 km s^{-1}, the forecast has little value as it fails to capture the lowest observed solar wind speeds during this interval. Across *V* thresholds, forecast skill increases very gradually from *n* = 1 hour to *n* =20-30 hours, before rising more sharply. For *V* > 500 km s^{-1}, the forecast becomes more valuable than the climatological mean at neighbourhood sizes of around 20 hours or longer. This time scale is roughly comparable the average timing error for HSEs (see Table 4) and indicates where the false alarm and missed events begin to cancel out, removing the double penalty effect. The fact that most V thresholds converge to FSS ≈ 1 at the maximum neighbourhood size (n = 630 hours) shows there is little bias in the occurrence of such events. For $V > 600$ km s^{-1}, FSS converges to values less than 1, highlighting an occurrence bias in the forecast for such an event definition (with the forecast slightly over-predicting occurrence of $V > 600$ km s^{-1}).

The same basic approach can also be applied to a probabilistic forecast. However, in the fMSE calculation the forecast fraction of hours above the threshold V, f_F, is replaced by p_F, the average probability of V above the threshold in a given neighbourhood. Thus the uncertainty information is preserved, without the need to investigate different probability thresholds. Figure 7b shows how the FSS varies with neighbourhood size and V threshold for the probablistic forecast of CR 2049. The general trends are similar to the deterministic forecast. But it's clear that the probablistic forecast provides significantly higher FSS at lower n, particularly for V thresholds below 500 km s^{-1}. This is because it intrinsically involves an (imperfect) estimate of timing error and thus some reduction of the double penalty. The rapid rise in FSS with n is consequently less apparent. As the probabilistic forecast includes an increased occurrence of low speed solar wind compared to the deterministic forecast, albeit at low probability, the probabilitstic forecast at $V > 350$ km s^{-1} is now valuable relative to the climatological mean. At the very highest event thresholds, $V > 550$ km s^{-1} and $V > 600$ km s^{-1}, there are insufficient events of high probability in this short interval, resulting in low FSS and a high bias for $V > 600$ km s^{-1} occurrence. Thus a great deal of diagnostic information can be obtained from the simple FSS analysis, which is complementary to point-by-point approaches.

3.3 B_Z Forecasts

Thus far, these issues have been illustrated exclusively with an example of a solar wind speed forecast. V is one of the more accurately forecast solar wind parameters [e.g., *Owens et al.*, 2008; *MacNeice*, 2009] and is always positive in value. Perhaps the solar wind parameter of greatest importance for space weather is the out-of-ecliptic component of the heliospheric magnetic field (HMF), B_Z, which is fundamentally less predictable than V [*Lockwood et al.*, 2016], both due to its stochastic nature and the difficulty in making remote observations of this parameter [*DeForest et al.*, 2017]. Validation of B_Z forecasts is complicated by the bipolar variations associated with geoeffective coronal mass ejections, which will be particularly susceptible to double penalties. This is illustrated in Figure 8, where a hypothetical forecast of B_Z for the Bastille Day interplanetary coronal mass ejection (ICME), in July 2000, has been produced by smoothing and shifting the observed time series by 18 hours, representative of current ICME forecast timing errors [*Tucker-Hood et al.*, 2015; *Riley et al.*, 2018]. By accurately reproducing the magnitude and direction of the magnetic field within the ICME and sheath region, such a forecast is a far more accurate than any current capability [e.g., *Savani et al.*, 2015]. Yet all the point-by-point metrics, whether they be error functions (even $r_L = -0.1$) or binary metrics, show the forecast to be significantly worse than assuming B_Z is approximately zero at all times. (The total area under the ROC curve is slightly larger than 0.5 but the sampling of B_Z space is uneven. For negative B_Z thresholds, the conditions of interest for space weather, the forecast lies below the y=x line and hence is deemed worse than the climatological mean.)

A features-based metric, equivalent to the high-speed enhancements, would clearly work well in this instance. But the difficultly is in rigorously defining a useable definition: Timescales which would pick out a feature in the body of this ICME may exclude negative B_Z intervals in other ICMEs or in the ICME sheath, which involve higher-frequency variations. The more feature-agnostic approach of the scale-selective fraction skill score is preferable. Figure 9 shows the fraction skill score of the B_Z forecast over a range of time scales (or neighbourhood sizes) and for a range of B_Z thresholds. For neighbourhoods smaller than 10 hours, the forecast is “worse” than assuming $B_Z \approx 0$ at all times, as the point-by-point analyses concluded. But as the time scale is increased to around 10-30 hours, the forecast is shown to be skilful relative to the climatological mean, as one would likely conclude by eye.

4 Summary

This study briefly reviewed some of the commonly used metrics for space-weather forecast and model validation. Simple error functions, like root-mean-square (RMS) and mean-absolute error (MAE), are the mainstay of forecast validation. They compare forecasts and observations on a strictly point-by-point basis. They are undoubtedly a valuable tool for forecast comparison. But there are limitations in their use as forecast diagnostics and they can, in some circumstances, give misleading results about the value or usefulness of a forecast. In particular, by treating each time point entirely independently, timing uncertainties are not explicitly accounted for. Thus when timing errors are present in the forecasts, they can be hit with “double penalties”, for both missing the observed event and issuing a false alarm. While there is nothing inherently wrong with this form of assessment, it can systematically favour overly conservative forecasts, which may not be beneficial. Binary metrics, in which a forecast is converted to series of “yes/no” predictions, reduce the emphasis on event magnitude and hence somewhat reduce the effect of double penalties for timing errors. These kinds of approaches are summarised by the receiver operator characteristic (ROC) and the Cost-Loss analysis. These can provide useful insight into the operational circumstances in which a particular forecast is most useful (e.g., in settings where false alarms are not a major issue).

A neat, simple, solution to the double penalty problem is for all forecasts to include an accurate assessment of uncertainty. As shown here, even relatively coarse estimates of uncertainty can add value to existing forecasts. But there are a number of reasons why this is not always practical. Instead, this study has advocated a more pragmatic solution of “time window” metrics alongside the more traditional point-by-point approaches. Defining discrete, extended features in the observed and forecast time series allows direct comparison of their timing and magnitude. This is a powerful analytical tool, but requires a rigorous *a prior* definition of an event, which is robust to event-to-event variability, and between observations and forecast. An alternative is to use a scale-selective approach, wherein agreement between forecast and observation is considered at a range of time scales. As the time scales become increasingly coarse, false alarms and missed events increasingly cancel out, reducing the double penalty effect. This allows an assessment of the time scales at which the forecast provides an acceptable level of accuracy.

Part of the job of a metric is to summarise a complex parameter space: Different parameter and forecast probability thresholds, different spatial and temporal scales and different operational sensitivities. The examples shown here consider only the simplest case of solar wind time series. Validation in other domains of the space-weather system also has to deal with intrinsically higher dimensionality. For example, in radiation belt forecasting, in addition to temporal variations, there is a great deal of spatially variability in all three directions (radially from the Earth, and in geomagnetic latitude and magnetic local time), as well as in particle energy space [e.g., *Shprits et al.*, 2015]. Often this dimensionality is reduced by averaging over particle drift and bounce motions, but the situation nevertheless remains more complex than a single time series. But the same fundamental issues are still present, just in a more multifarious way.

Finally, it is worth reiterating that these more sophisticated methods of forecast and model validation are intended to compliment, not replace, existing metrics. Error functions should undoubtedly continue to be a standard space-weather metric. In additional to continuing the legacy, they are simple to implement and interpret, as well as enabling easy inter-comparison of different forecasts and models. But a more diagnostic picture of *why* a forecast is accurate or fails is invaluable too.

Acknowledgements

Omni data were provided by the Space Physics Data Facility and National Space Science Data Center. MAS model output is available from the Predictive Science Inc. website: http://www.predsci.com/mhdweb/home.php. MO is part-funded by Science and Technology Facilities Council (STFC) grant numbers ST/M000885/1 and ST/R000921/1, and Natural Environment Research Council (NERC) grant number NE/P016928/1.

Bibliography

DeForest, C., C. de Koning, and H. Elliott (2017), 3D Polarized Imaging of Coronal Mass Ejections: Chirality of a CME, *The Astrophysical Journal*, *850*(2), 130, doi:10.3847/1538-4357/aa94ca.

Dungey, J. W. (1961), Interplanetary magnetic field and the auroral zones, *Phys. Rev. Lett.*, *6*, 47, doi:10.1103/PhysRevLett.6.47.

Ebert, E. E., and W. A. Gallus Jr (2009), Toward Better Understanding of the Contiguous Rain Area (CRA) Method for Spatial Forecast Verification, *Weather and Forecasting*, *24*(5), 1401-1415, doi:10.1175/2009waf2222252.1.

Finley, J. P. (1884), Tornado predictions, *Amer. Meteor. J*, *1*, 85-88.

Jian, L. K., P. J. MacNeice, M. L. Mays, A. Taktakishvili, D. Odstrcil, B. Jackson, H.-S. Yu, P. Riley, and I. V. Sokolov (2016), Validation for global solar wind prediction using Ulysses comparison: Multiple coronal and heliospheric models installed at the Community Coordinated Modeling Center, *Space Weather*, *14*(8), 592-611, doi:doi:10.1002/2016SW001435.

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*, doi:10.1029/2004JA010649.

Linker, J., Z. Mikic, D. A. Biesecker, R. J. Forsyth, W. E. Gibson, A. J. Lazarus, A. Lecinski, P. Riley, A. Szabo, and B. J. Thompson (1999), Magnetohydrodynamic modeling of the solar corona during whole sun month, *J. Geophys. Res.*, *104*, 9809-9830.

Lockwood, M., M. J. Owens, L. A. Barnard, S. Bentley, C. J. Scott, and C. E. Watt (2016), On the origins and timescales of geoeffective IMF, *Space Weather*, *14*(6), 406-432, doi:10.1002/2016SW001375.

MacNeice, P. (2009), Validation of community models: Identifying events in space weather model timelines, *Space Weather*, *7*(6), doi:10.1029/2009sw000463.

P. MacNeice, L. Jian, S.K. Antiochos, C.N. Arge, C.D. Bussy-Virat, M.L. DeRosa, B.V. Jackson, J.A. Linker, Z. Mikic, M.J. Owens, A.J. Ridley, P. Riley, N. Savani, I. Sokolov, Space Weather, doi:10.1029/2018SW002040, 2018

Mason, I. (1982), A model for assessment of weather forecasts, *Aust. Meteor. Mag*, *30*(4), 291-303.

Mason, S. J., and N. E. Graham (2002), Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145-2166, doi:doi:10.1256/003590002320603584.

Mays, M. L., et al. (2015), Ensemble Modeling of CMEs Using the WSA–ENLIL+Cone Model, *Solar Physics*, *290*(6), 1775-1814, doi:10.1007/s11207-015-0692-1.

McCloskey, A. E., P. T. Gallagher, and D. S. Bloomfield (2018), Flare forecasting using the evolution of McIntosh sunspot classifications, *Journal of Space Weather and Space Climate*, *8*, A34, doi:10.1051/swsc/2018022.

Murphy, A. H. (1977), The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation, *Mon. Weather Rev.*, *105*(7), 803-816.

Murphy, A. H. (1993), What is a good forecast? An essay on the nature of goodness in weather forecasting, *Wea. Forecasting*, *8*, 281-293.

Murphy, A. H. (1996), The Finley Affair: A Signal Event in the History of Forecast Verification, *Weather and Forecasting*, *11*(1), 3-20, doi:10.1175/1520-0434(1996)011<0003:tfaase>2.0.co;2.

Murray, S. A. (2018), The importance of ensemble techniques for operational space weather forecasting, *Space Weather*, doi:10.1029/2018SW001861

Murray, S. A., S. Bingham, M. Sharpe, and D. R. Jackson (2017), Flare forecasting at the Met Office Space Weather Operations Centre, *Space Weather*, *15*(4), 577-588, doi:doi:10.1002/2016SW001579.

Owens, M., and P. Riley (2017), Probabilistic Solar Wind Forecasting Using Large Ensembles of Near-Sun Conditions With a Simple One-Dimensional “Upwind” Scheme, *Space Weather*, *15*(11), 1461-1474, doi:10.1002/2017SW001679.

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

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 J.*, *11*, 225-236, doi:10.1002/swe.20040.

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

Owens, M. J., P. Riley, and T. S. Horbury (2017), Probabilistic Solar Wind and Geomagnetic Forecasting Using an Analogue Ensemble or “Similar Day” Approach, *Solar Physics*, *292*(5), 69, doi:10.1007/s11207-017-1090-7.

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

Peterson, W., T. Birdsall, and W. Fox (1954), The theory of signal detectability, *Transactions of the IRE professional group on information theory*, *4*(4), 171-212, doi:10.1109/TIT.1954.1057460.

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:doi:10.1002/2016SW001390.

Richardson, D. S. (2000), Skill and relative economic value of the ECMWF ensemble prediction system, *Q.J.R. Meteorol. Soc.*, *126*(563), 649-667.

Riley, P., J. A. Linker, R. Lionello, and Z. Mikic (2012), Corotating interaction regions during the recent solar minimum: The power and limitations of global MHD modeling, *Journal of Atmospheric and Solar-Terrestrial Physics*, *83*, 1-10, doi:10.1016/j.jastp.2011.12.013.

Riley, P., J. A. Linker, and Z. Mikić (2013a), On the application of ensemble modeling techniques to improve ambient solar wind models, *Journal of Geophysical Research: Space Physics*, *118*(2), 600-607, doi:10.1002/jgra.50156.

Riley, P., J. A. Linker, Z. Mikič, G. P. Zank, J. Borovsky, R. Bruno, J. Cirtain, S. Cranmer, H. Elliott, and J. Giacalone (2013b), Ensemble modeling of the ambient solar wind, paper presented at AIP Conference Proceedings, AIP.

Riley, P., et al. (2018), Forecasting the Arrival Time of Coronal Mass Ejections: Analysis of the CCMC CME Scoreboard, *Space Weather*, doi:10.1029/2018SW001962.

Roberts, N. (2008), Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model, *Meteorological Applications*, *15*(1), 163-169, doi:doi:10.1002/met.57.

Roberts, N. M., and H. W. Lean (2008), Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, *Monthly Weather Review*, *136*(1), 78-97, doi:10.1175/2007mwr2123.1.

Savani, N. P., A. Vourlidas, A. Szabo, M. L. Mays, I. G. Richardson, B. J. 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.

Shprits, Y. Y., A. C. Kellerman, A. Y. Drozdov, H. E. Spence, G. D. Reeves, and D. N. Baker (2015), Combined convective and diffusive simulations: VERB-4D comparison with 17 March 2013 Van Allen Probes observations, *Geophysical Research Letters*, *42*(22), 9600-9608, doi:doi:10.1002/2015GL065230.

Siscoe, G. (2007), Space weather forecasting historically viewed through the lens of meteorology, in *Space Weather- Physics and Effects*, edited, p. 5, doi:10.1007/978-3-540-34578-7-2.

Slingo, J., and T. Palmer (2011), Uncertainty in weather and climate prediction, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4751-4767, doi:10.1098/rsta.2011.0161.

Spence, H., D. Baker, A. Burns, T. Guild, C.-L. Huang, G. Siscoe, and R. Weigel (2004), Center for integrated space weather modeling metrics plan and initial model validation results, *J. Atmos. Sol. Terr. Phys.*, *66*, 1491-1498.

Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192, doi:doi:10.1029/2000JD900719.

Theis, S. E., A. Hense, and U. Damrath (2005), Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach, *Meteorological Applications*, *12*(3), 257-268, doi:doi:10.1017/S1350482705001763.

Thornes, J. E., and D. B. Stephenson (2001), How to judge the quality and value of weather forecast products, *Meteorological Applications*, *8*(3), 307-314, doi:10.1017/S1350482701003061.

Tucker-Hood, K., et al. (2015), Validation of a priori CME arrival predictions made using real-time heliospheric imager observations, *Space Weather*, *13*(1), 35-48, doi:10.1002/2014SW001106.

Table 1: Point-by-point metrics for the solar wind speed forecasts shown in Figure 1. (a) Mean-square error, (b) root-mean-square error, (c) mean absolute error, (d) Pearson (linear) correlation coefficient, (e) Spearman (rank-order) correlation coefficient, (f) receiver operator characteristic area under curve.

	MSE[a] [km^2 s^{-2}]	RMS[b] [km s^{-1}]	MAE[c] [km s^{-1}]	r_L [d]	r_S [e]	ROC area under curve (AUC)[f]
Forecast	1.30x10^4	114.0	84.6	0.28	0.06	0.68
Climatological mean	0.98x10^4	98.9	85.0	0.00	0.00	0.50

Table 2: A contingency table for the forecast of solar wind speed events in CR 2049 defined by a threshold of *V* > 500 km s^{-1}. *TP, FP, TN* and *FN* are the numbers of true positive, false positive, true negative and false negative intervals, respectively. *P* and P_F are the number of observed and forecast events, while *N* and N_F are the number of observed and forecast non-events.

		Event in forecast? (i.e., V_F > 500 km s^{-1})		
		Yes	No	Total
Observed event? (i.e., *V* > 500 km s^{-1})	Yes	*TP* = 68	*FN* = 124	*P* = 192
	No	*FP* = 109	*TN* = 354	*N* = 463
	Total	P_F = 177	N_F = 478	655

Table 3: The same as Table 2, but for the climatological mean of solar wind speed for CR 2049.

		Event in climatological mean? (i.e., V_F > 500 km s^{-1})		
		Yes	No	Total
Observed event? (i.e., V > 500 km s^{-1})	Yes	TP = 0	FN = 192	P = 192
	No	FP = 0	TN = 463	N = 463
	Total	P_F = 0	N_F = 655	655

Table 4: Results of the high-speed enhancement analysis applied to the observed and forecast solar wind speed for CR 2049, mid-Oct to mid-Nov 2006. In both case, 3 HSEs were identified. Δ indicates the (observed – forecast) value.

HSE	Observed	Forecast	ΔT [days]	| ΔT| [days]	V_{MAX} obs [km s^{-1}]	V_{MAX} for [km s^{-1}]	ΔV_{MAX} [km s^{-1}]	$\|\Delta V_{MAX}\|$ [km s^{-1}]
1	2006-10-20 T21:00	2006-10-19 T02:00	1.79	1.79	630	547	82.9	82.9
2	2006-10-28 T16:00	2006-10-26 T12:00	2.17	2.17	580	612	-32.1	32.1
3	2006-11-09 T23:00	2006-11-09 T6:00	0.71	0.71	633	638	-5.1	5.1
Mean	-	-	1.56	1.56	614	599	15.2	40.1

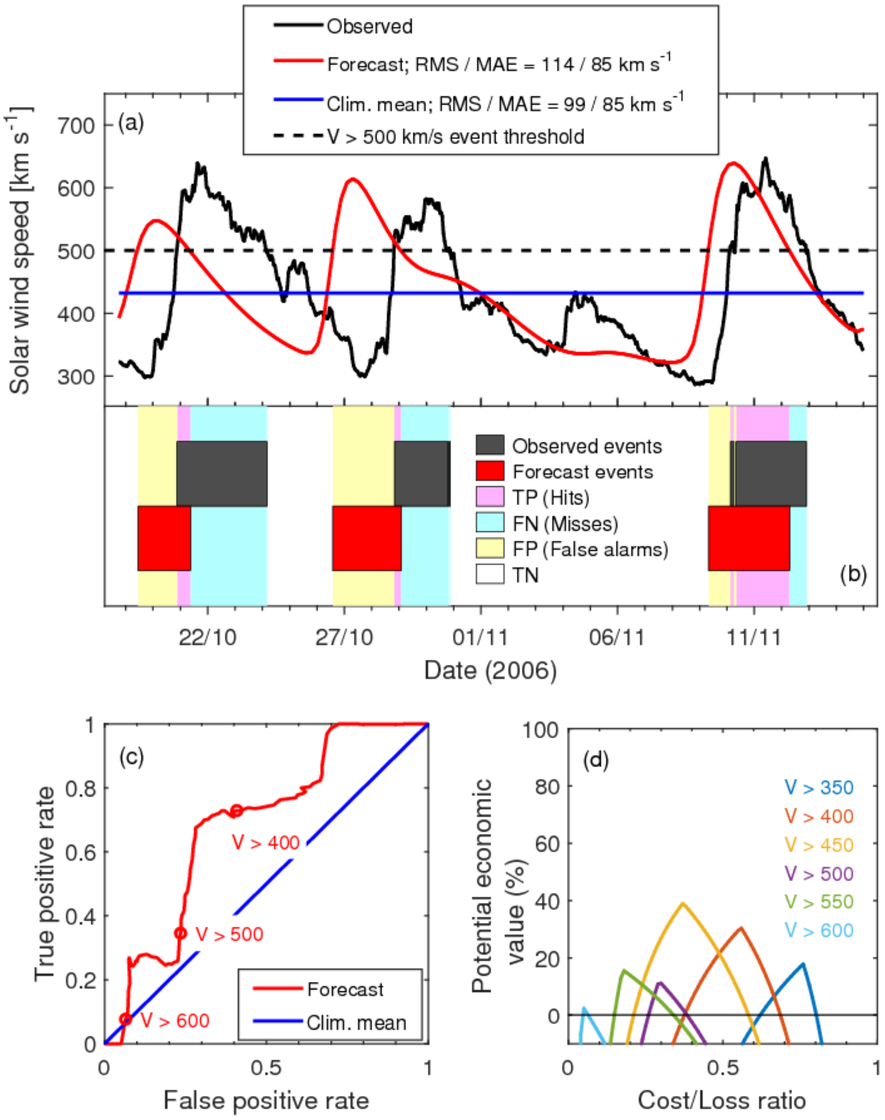

Figure 1: An example of a deterministic solar wind speed forecast and associated "point-by-point" metrics. (a) Time series of hourly means of near-Earth solar wind speed, *V*, for CR 2049, spanning mid-Oct to mid-Nov 2006, as observed (black) and forecast (red). The climatological mean for this interval (blue) is also shown. (b) Solar wind speed events defined using a threshold of *V* > 500 km s^{-1}. (c) The receiver operator characteristic (ROC), which plots the true positive rate against the false positive rate for a range of solar wind speed event definitions. (d) The potential economic value of the forecast at various *V* thresholds and cost/loss ratios. See text for more detail.

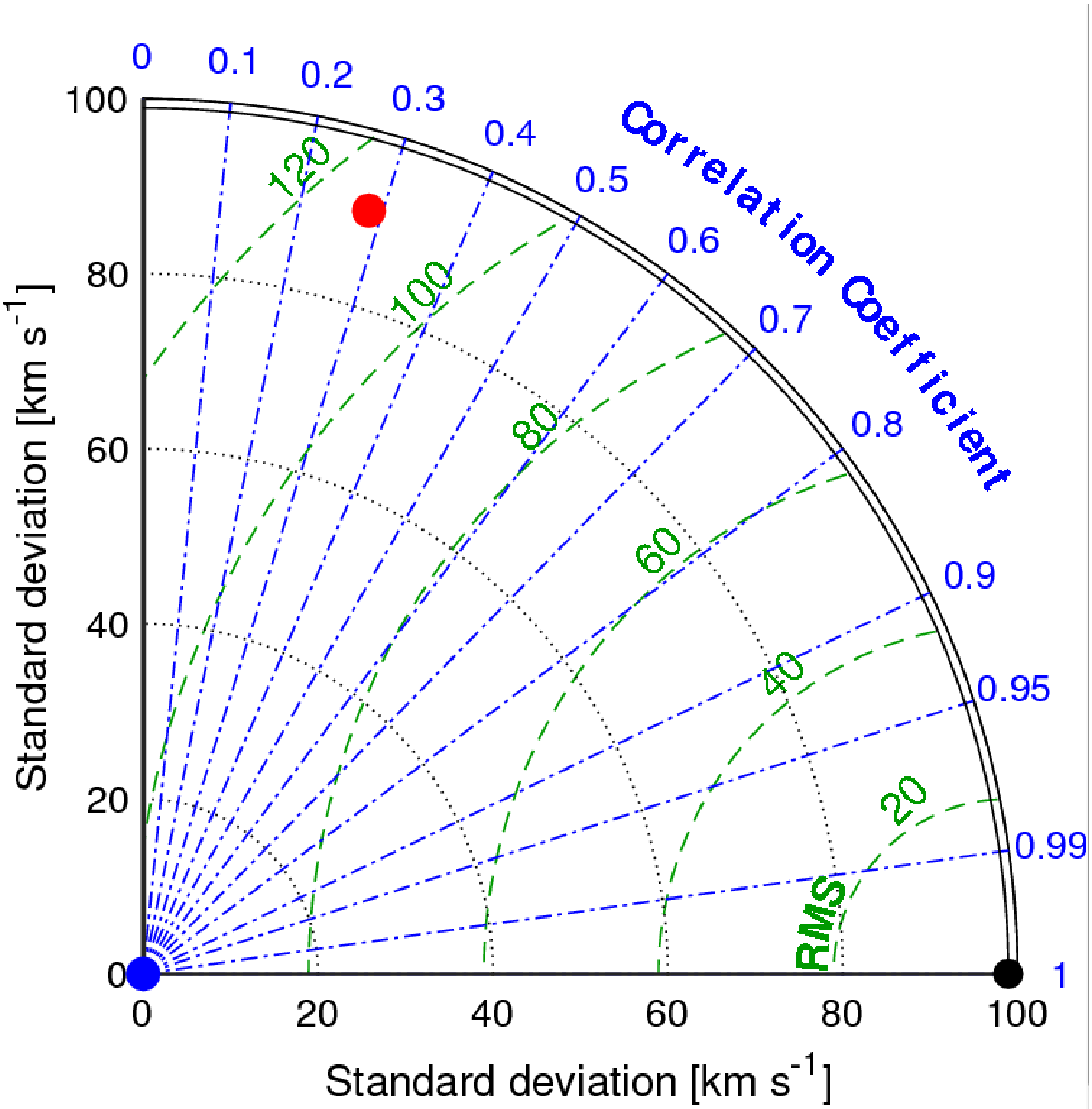

Figure 2: A Taylor diagram of the solar wind speed time series shown in Figure 1. The radial distance from the origin shows the standard deviation of the time series, while the azimuthal angle about the origin shows the linear correlation coefficient (note non-linear scale) with the observed time series. Green dashed circles show contours of constant RMS error (with forecast and observation mean subtracted). The black, red and blue points show the observed, forecast and climatological *V*, repsectively.

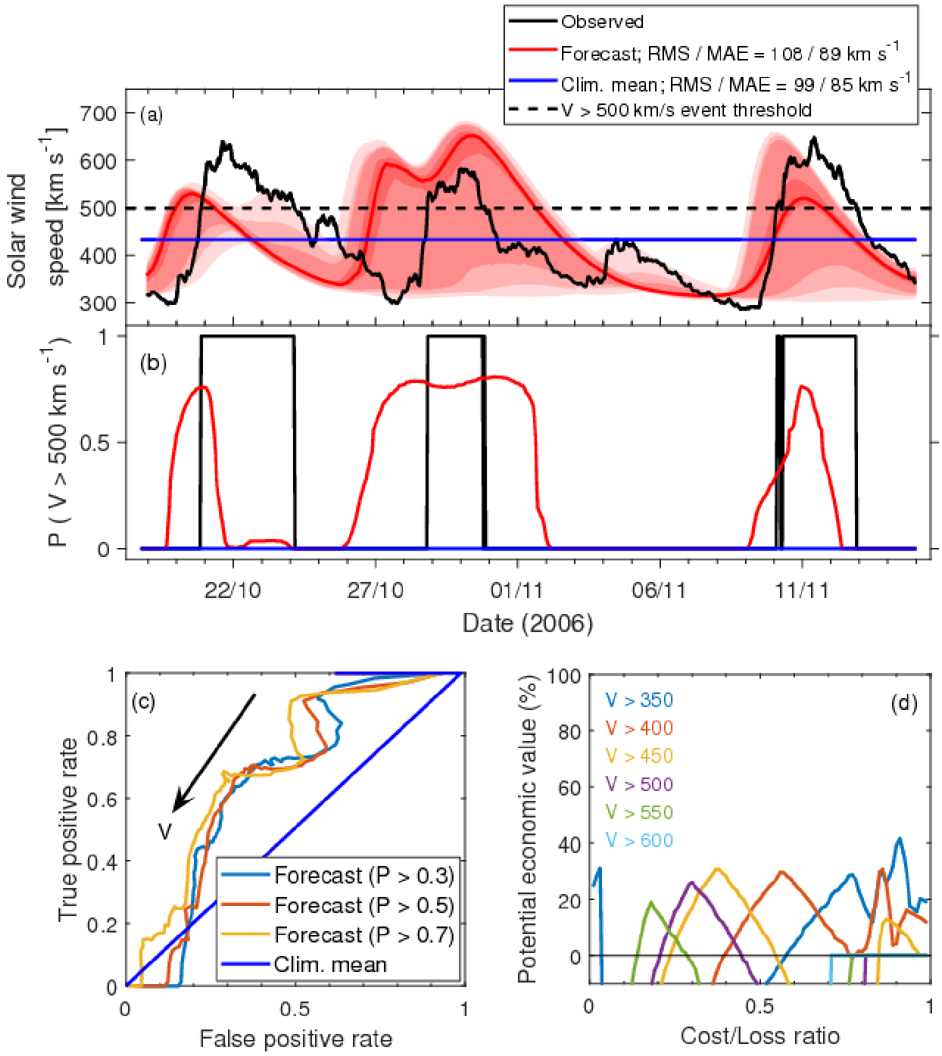

Figure 3: An example of a probabilistic solar wind speed forecast and associated point-by-point metrics. Panel (a) shows the time series of hourly means of near-Earth solar wind speed for CR2049, mid-Oct to mid-Nov 2006, as observed (black) and forecast by the ensemble median (red), with pink-shaded areas showing 68, 90, 95 and 99.8 percentiles of the forecast ensemble. The climatological mean for this interval (blue) is also shown. A threshold of *V* > 500 km s^{-1} is to define events in the time series (black dashed line), which are represented in panel (b) as a probability of occurrence. Panel (c) shows the receiver operator characteristic (ROC) for three different probability thresholds. Panel (d) shows the cost-loss curves for the forecast at various action thresholds of *V*.

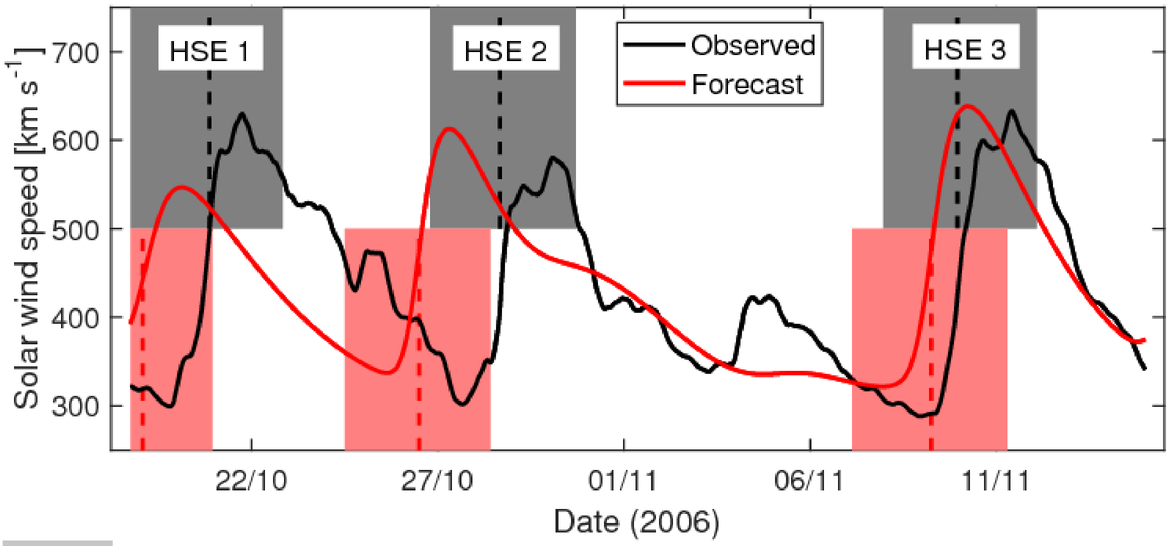

Figure 4: High-speed enhancement (HSE) analysis applied to the solar wind speed observed (black) and forecast (red) for CR2049, mid-Oct to mid-Nov 2006. All data has been 8-hour smoothed. Black- and red-shaded intervals show times when observed and forecast *V* meet the criteria for a HSE, respectively. The dashed vertical lines show the times of maximum *V* gradient.

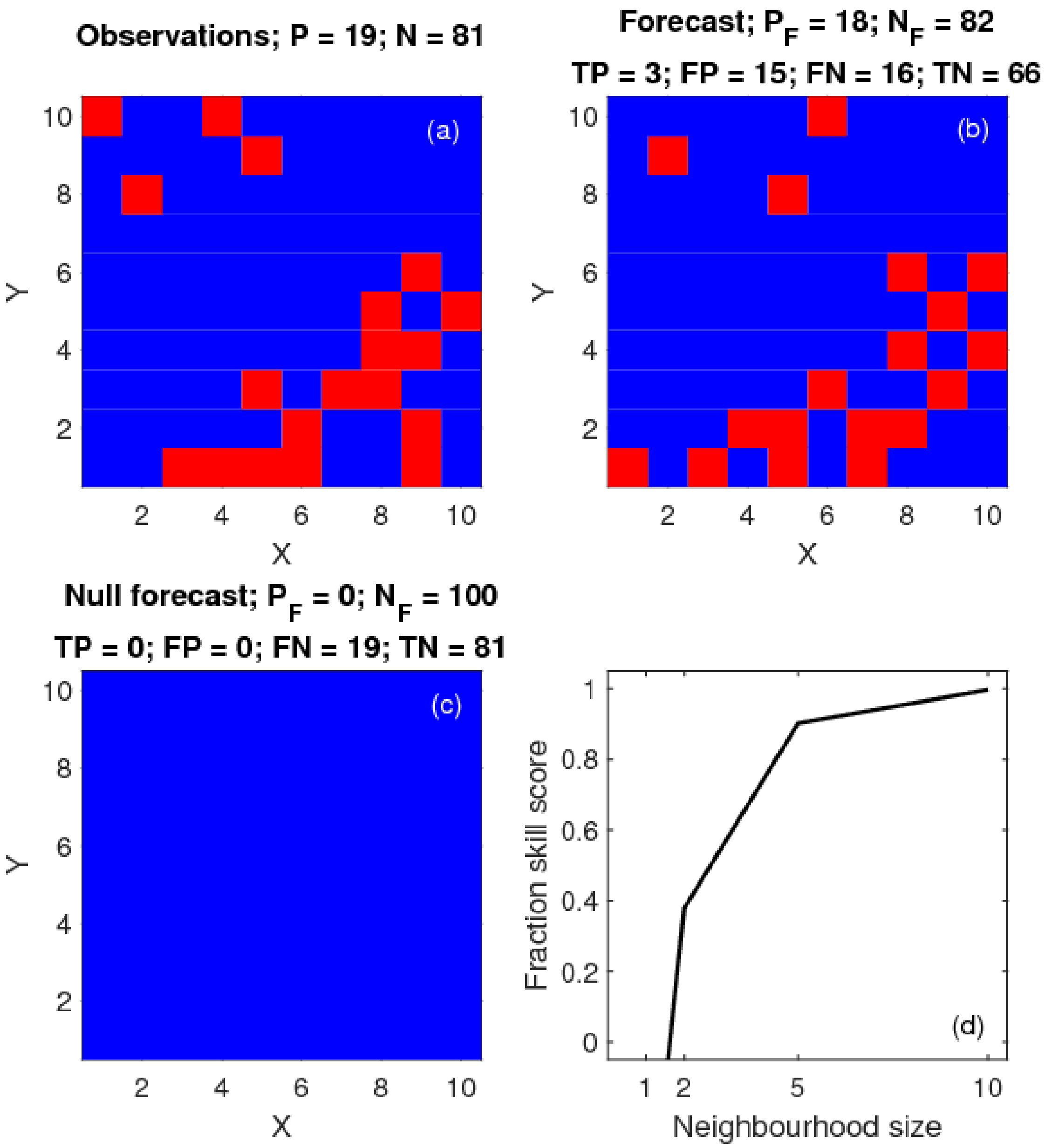

Figure 5: Spatial distributions of hypothetical (a) observed and (b) forecast rain. Red is a positive observation/forecast at a given position, blue is negative. A null forecast (c) predicts no rain anywhere. Panel (d) shows the fraction skill score for different spatial scales (or neighbourhood sizes).

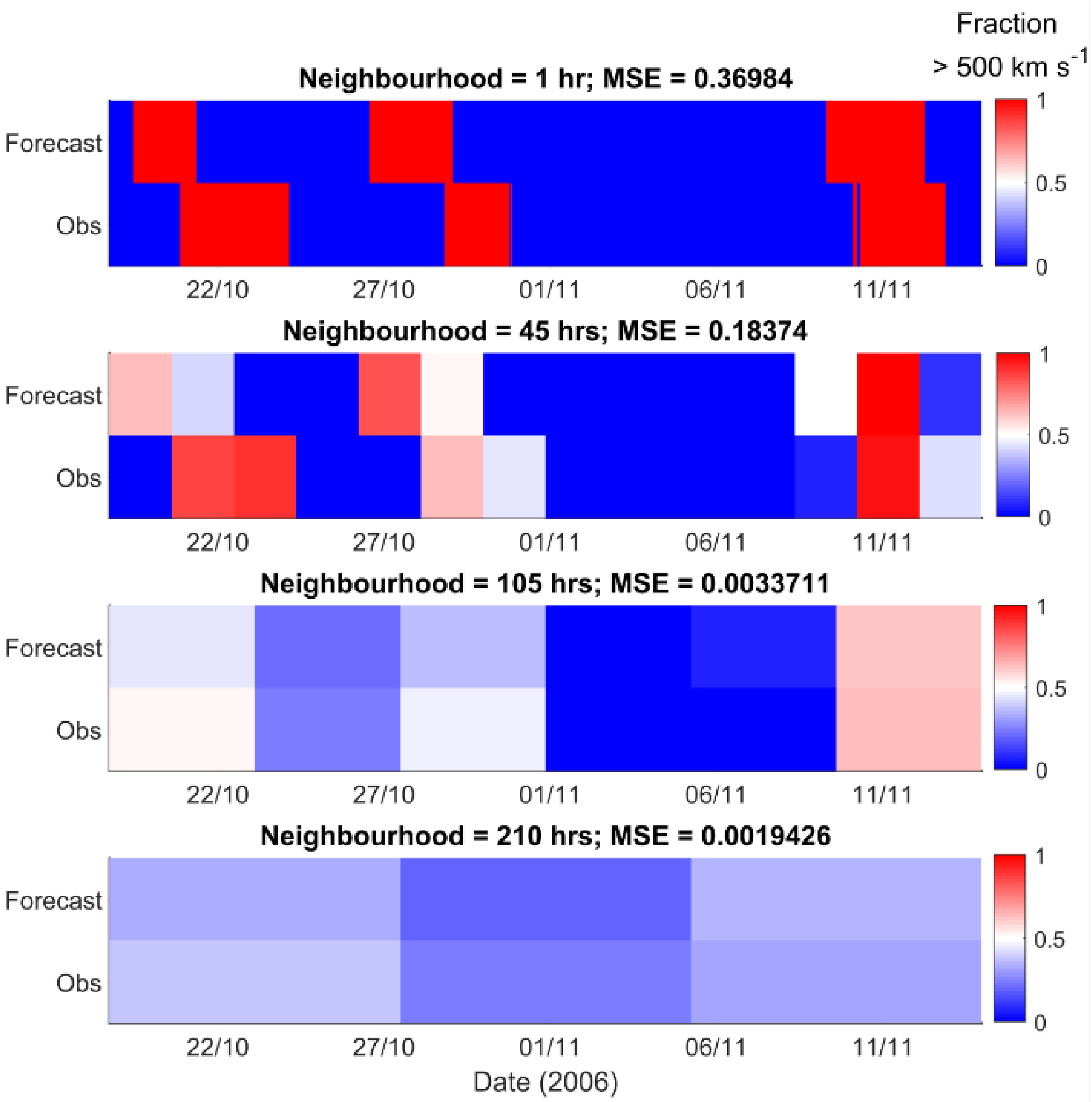

Figure 6: Scale-selective metrics applies the observed and forecast solar wind speed for CR2049, mid-Oct to mid-Nov 2006. The colour scale shows the fraction of individual hours within a neighbourhood which exceed a speed threshold of 500 km s^{-1}, from 0 (blue) to 1 (red). The top panel shows a neighbour size of 1 hour and thus is simply the threshold applied to the original observations and forecast (i.e., the same as Figure 1b). The second, third and fourth panels show neighbour sizes of 45, 105 and 210 hours.

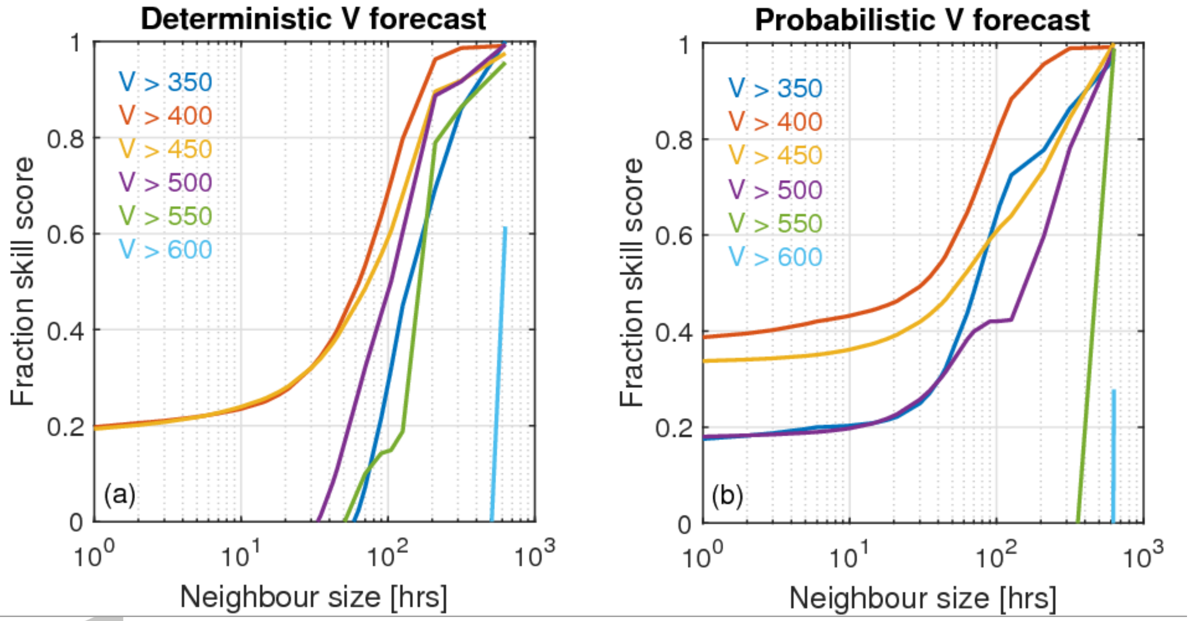

Figure 7: Fraction skill score for the forecast solar wind speed for mid-Oct to mid-Nov 2006 for a range of speed thresholds and neighbour sizes. The climatological mean is used as the baseline forecast. (a) The deterministic forecast of *V*, (b) the probabilistic forecast of *V*.

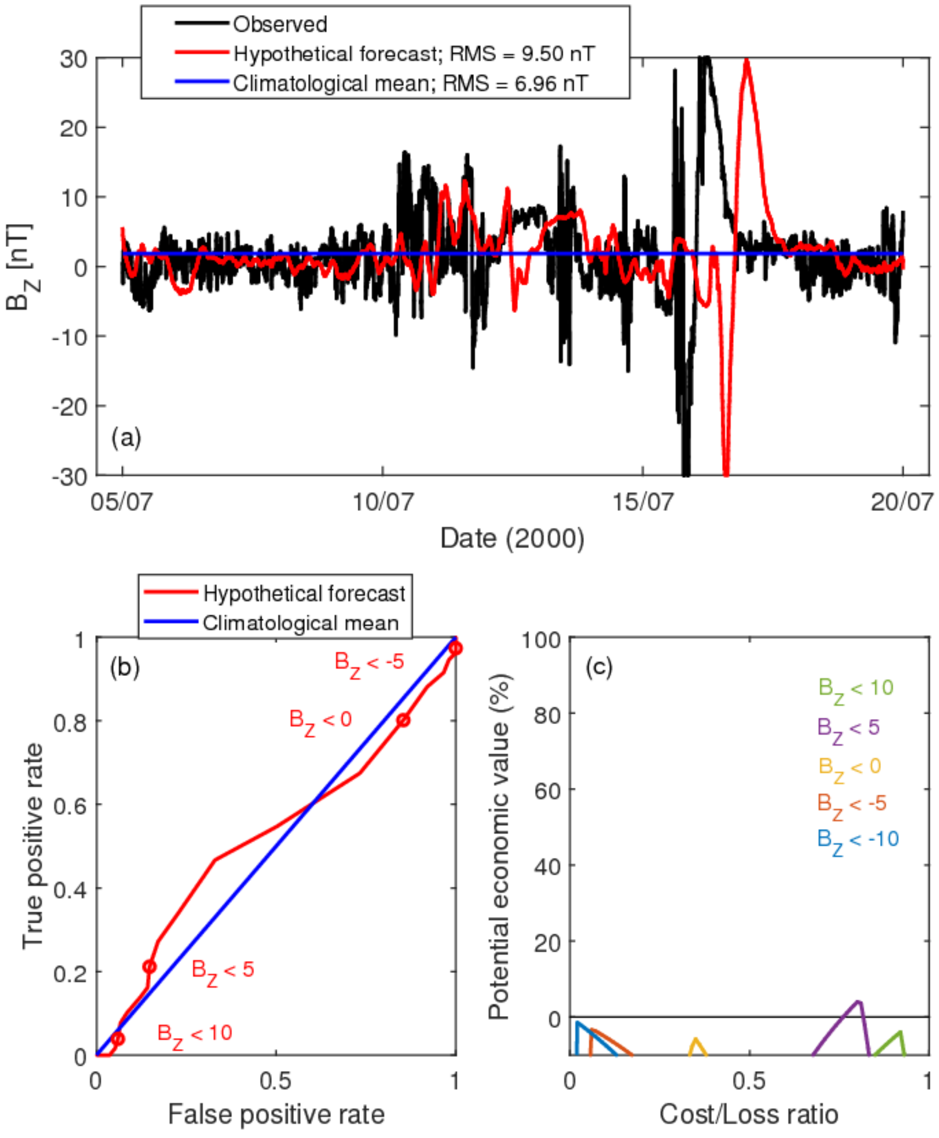

Figure 8: Point-by-point metrics for a hypothetical forecast of the out-of-ecliptic HMF component, B_Z. (a) Time series of observed B_Z (black) for 15 days around the "Bastille Day" CME of July 2000. A hypothetical forecast (black) has been produced by smoothing and shifting the observations by 18 hours. The blue line shows the mean B_Z for this period (b) The ROC for the forecast and climatological mean. (c) Cost-loss analysis for different B_Z thresholds.

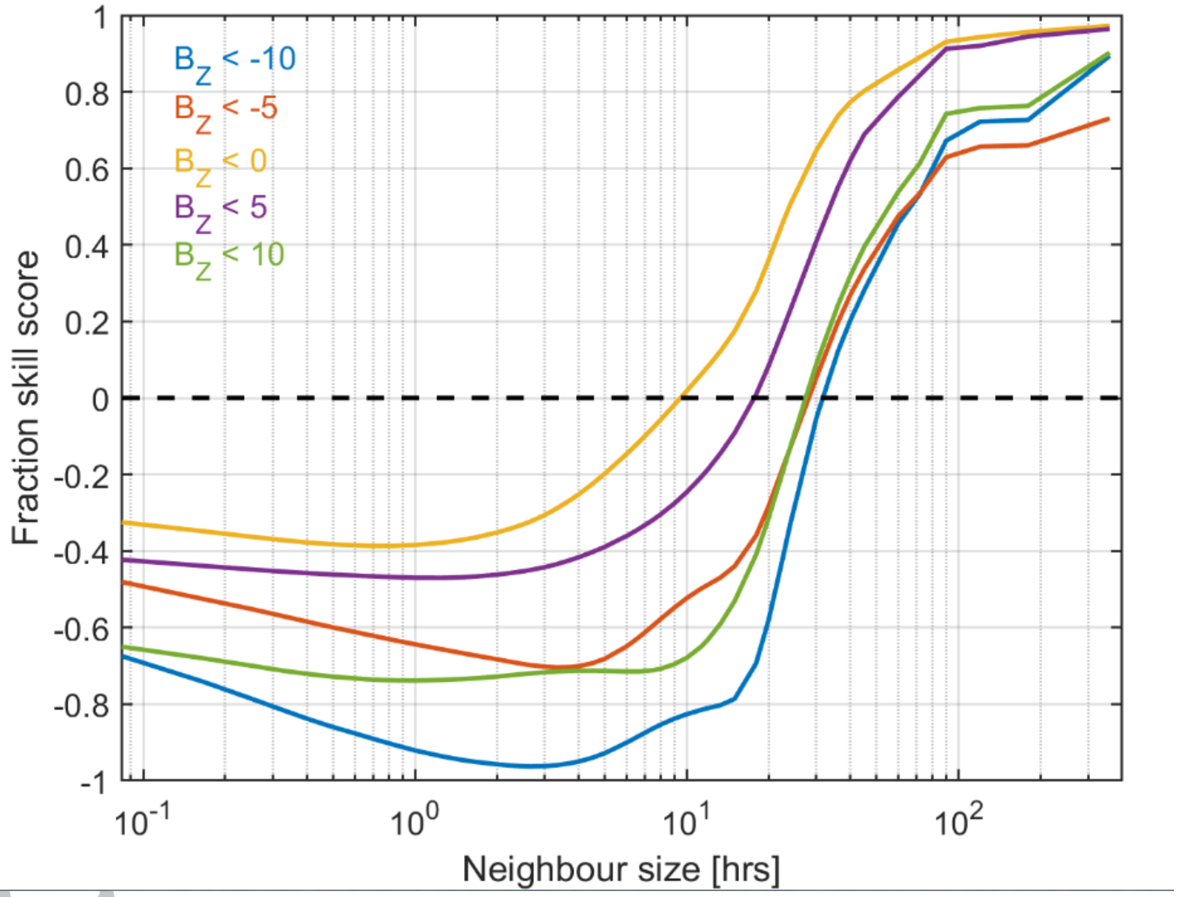

Figure 9: Fraction skill score for the B_Z forecast for 15 days around the "Bastille Day" CME of July 2000, for a range of B_Z thresholds and neighbour sizes. The climatological mean is used as the baseline forecast.