Calibrating probabilistic solar-wind forecasts driven by the Wang-Sheeley-Arge model

Space Weather

RESEARCH ARTICLE
10.1029/2025SW004614

Calibrating Probabilistic Solar-Wind Forecasts Driven by the Wang-Sheeley-Arge Model

N. O. Edward-Inatimi[1], M. J. Owens[1], L. Barnard[1], H. Turner[2], M. Marsh[2], S. Gonzi[2], and M. Lang[1]

[1]University of Reading, Reading, UK, [2]UK Met Office, Exeter, UK

Key Points:

- Calibration is a necessary step for probabilistic prediction systems; a calibrated ensemble adds useful information content into a forecast
- Calibrating the WSA-HUXt ensemble improves reliability and reduces continuous ranked probability score with a trade-off in statistical resolution/decisiveness
- Current ensemble forecast methods struggle to capture the true levels of ambient variability during solar maximum

Correspondence to:
N. O. Edward-Inatimi,
n.o.edward-inatimi@pgr.reading.ac.uk

Citation:
Edward-Inatimi, N. O., Owens, M. J., Barnard, L., Turner, H., Marsh, M., Gonzi, S., & Lang, M. (2026). Calibrating probabilistic solar-wind forecasts driven by the Wang-Sheeley-Arge model. *Space Weather*, *24*, e2025SW004614. https://doi.org/10.1029/2025SW004614

Received 9 JUL 2025
Accepted 23 FEB 2026

Author Contributions:
Conceptualization: N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, M. Lang
Data curation: N. O. Edward-Inatimi
Formal analysis: N. O. Edward-Inatimi
Investigation: N. O. Edward-Inatimi
Methodology: N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, M. Lang
Resources: N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, M. Lang
Software: N. O. Edward-Inatimi
Supervision: M. J. Owens, L. Barnard, H. Turner, M. Marsh, S. Gonzi, M. Lang
Validation: N. O. Edward-Inatimi
Visualization: N. O. Edward-Inatimi
Writing – original draft: N. O. Edward-Inatimi

Abstract By spatially perturbing coronal model output within a coupled coronal-heliospheric model we can generate probabilistic predictions of solar-wind speed. We apply these spatial perturbations to the Wang-Sheeley-Arge (WSA) model output to generate large sets of input conditions for the Heliospheric Upwind eXtrapolation with time dependence (HUXt) solar-wind model. The resulting ensemble forecasts at 1 AU contain useful information about likely outcomes and the method allows uncertainty to be better characterized. We tune the scales of perturbations to calibrate the probabilistic predictions. We use the rank histogram and reliability component of the Brier score to demonstrate how increasing levels of perturbation/variability generally improves the reliability of the WSA-HUXt ensemble distribution; the ability of the ensemble to capture the true likelihood of events based on observational frequencies. We use the resolution component of the Brier score to highlight how too large a perturbation harms the statistical resolution of the forecast; the ability of the model to meaningfully distinguish between events beyond a statistical observational baseline (like climatology). This adds a useful constraint on the maximum size of perturbation we should be applying. Additionally, we use continuous ranked probability score to demonstrate how a calibrated ensemble can improve a prediction system, reducing forecast error across all lead times. Finally we demonstrate that the calibrated ensemble provides value for an end-user through a Cost/Loss analysis. In refining this calibration procedure we provide optimal values of the perturbation parameters for use in the operational WSA-HUXt forecast.

Plain Language Summary Ensemble calibration aims to bring forecast probabilities in-line with the true observed frequency of events. We improve upon solar-wind ensemble forecasting by building on recently developed calibration methods and aim to better understand uncertainty in solar wind predictions. The Wang-Sheeley-Arge (WSA) model is commonly used in forecasting and we use it along with the HUXt solar wind model to create an ensemble of forecasts, each with slightly different boundary conditions. The spread of input conditions for the ensemble can greatly impact forecast performance. By controlling the spread of the ensemble input we find that improving how reliable the forecast is—how accurately it reflects real-world conditions—can sometimes reduce how clearly it distinguishes between different possible outcomes, harming statistical resolution. To address this, we fine-tune the level of variation in inner-boundary conditions within a range of calibrated values. This aims to strike a balance between reliability and resolution, helping make forecasts both accurate and useful. The paper also provides the best settings for these variations when using the WSA-HUXt model in a forecasting context.

1. Introduction

The solar wind is an outflow of magnetized plasma from the solar corona; the upper-most region of the solar atmosphere. Solar-wind speed and other characteristics can be modeled through the use of empirical and physics-based coronal-heliospheric models driven by observations of the photospheric magnetic field. The Wang-Sheeley-Arge (WSA) model is a coronal model which reconstructs the magnetic field within the solar coronal and uses empirical relationships to relate the field structure to bulk plasma properties (Arge & Pizzo, 2000). WSA is used operationally by the UK Met Office (UKMO) and US Space Weather Prediction Centre (SWPC) coupled with Enlil, a 3-D Magnetohydrodynamic (MHD) solar-wind model (Odstrcil, 2003). Coupled coronal-heliospheric modeling is an established and pragmatic approach to forecasting (Odstrcil et al., 2002, 2004; Riley et al., 2002), used as a work around for the lack of routine in situ observations of near-Sun solar-wind conditions needed for a true set of inner-boundary conditions (Cranmer et al., 2017). However, due to both the nature of the photospheric magnetic field observations, and the indirect way in which the near-Sun solar wind

Writing – review & editing:
N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, M. Marsh, S. Gonzi, M. Lang

speed is inferred, this approach introduces significant uncertainty into forecasts of near-Earth solar wind conditions (Bertello et al., 2014; Kennis et al., 2024; Perri et al., 2023). Uncertainties can be introduced by the choice of observatory and processing techniques for input magnetograms which can produce highly variable coronal solutions for the same observations (Heinemann et al., 2025; Riley et al., 2014). Additionally, there are dynamical uncertainties introduced by various mesoscale processes (e.g., turbulence, microstreams, switchbacks) which are not explicitly modeled within our chosen solar-wind model and also by model behaviors across stream interaction regions between fast/slow wind and complex coronal hole structures which can be difficult to capture accurately in simulation (Hinterreiter et al., 2019; Jian et al., 2015). Most studies agree that the majority of uncertainty lies within the inner-boundary conditions used to drive heliospheric models. We can directly consider some of the uncertainty within the inner-boundary to generate an ensemble of forecasts at Earth. This can provide a way to systematically quantify the forecast uncertainty by providing probabilistic predictions (Owens & Riley, 2017). However, naive application of ensemble methods can risk inaccurately estimated uncertainties. Hence, one of the aims of calibration is to begin to more carefully characterize the uncertainty being captured by an ensemble prediction of the solar-wind speed (Edward-Inatimi et al., 2024).

Perturbed boundary ensembles combine many individual simulations with adjusted initial, and/or inner-boundary conditions, perturbed away from an assumed ground truth (Wilks, 2019a). This can determine sensitivity to boundary conditions, quantify uncertainty, and produce a probabilistic forecast. Ensemble methods have been explored and employed with success to solar wind forecasts (see Riley et al., 2013, and references therein) and Interplanetary Coronal Mass Ejection (ICME) arrival time forecasts (Barnard et al., 2020; Kumar et al., 2020; Lee et al., 2013; Mays et al., 2015). In operational settings ensemble techniques can provide crucial insights and value for decision makers (Henley & Pope, 2017; Pizzo et al., 2015). However, care must be taken when interpreting the output of an ensemble, which does not always reflect the real-life likelihoods of events. for example, 60% of ensemble members predicting the occurrence of an event does not necessarily mean there is a 60% probability of an event occurring. Ensemble calibration adjusts model inputs/outputs to better align with the known uncertainties and observed frequencies of events. This improves reliability, such that for a well-designed ensemble, the spread of results is well-correlated with forecast uncertainty (Edward-Inatimi et al., 2024). Calibration tunes a prediction system such that it provides the end-user with more realistic probabilities which reflect real-world likelihoods. Hence, a more trustworthy and actionable forecast. In addition to reliability, we must consider the statistical resolution of an ensemble. This is the ability of a model system to differentiate between likely and unlikely events. This statistical definition of resolution is the one we will explore in this paper. For clarity, we will specify when we are referring to a different form of resolution (e.g., temporal/spatial).

Within this coupled coronal-heliospheric model framework, CMEs can be added through time-dependent perturbations to the inner-boundary of the heliospheric model. However, in operational settings, high-confidence in ICME arrival time forecasts is hard to achieve if there is a large amount of uncertainty within the ambient solar wind (Mays et al., 2015; Riley & Ben-Nun, 2021). It is therefore vital to accurately assess the uncertainty present in the ambient solar wind, as a first step toward better CME forecasting. Calibration is the means to achieve this.

To model the solar wind for the calibration study we use WSA as the coronal component of our coupled coronal-heliospheric model. We couple WSA with the Heliospheric Upwind eXtrapolation with Time dependence (HUXt) model to generate solar-wind ensembles at 1 AU. HUXt is a reduced-physics solar-wind model developed by Owens et al. (2020) and Barnard and Owens (2022). Given an inner-boundary of solar-wind speeds near the Sun (typically at 21.5 r_S from WSA), HUXt treats the solar-wind as an incompressible fluid, solving a reduced 1-D Burgers' equation to model the solar-wind flow out to near-Earth space (and/or solar system). HUXt can match the dynamical evolution of global heliospheric MHD models to within 5% at a fraction of the computational cost compared to full 3D MHD models like Enlil (Owens et al., 2020). This makes it an ideal model for the calibration procedure which involves running over large, $O(100)$, sets of initial conditions, typically over a few hundred days. Furthermore, an ensemble regime which couples WSA and HUXt is run operationally by the University of Reading as part of the Space Weather Empirical Ensembles Package (SWEEP) project (see https://www.ralspace.stfc.ac.uk/Pages/SWIMMR.aspx and https://research.reading.ac.uk/met-spate/huxt-forecast/). The calibration procedure outlined in this paper aims to find optimal perturbation scales for this ensemble scheme. Through evaluating key aspects of the ensemble performance, we aim to demonstrate how to optimize the forecasting system based on the goals of the end user. Finally, we will discuss some limitations surrounding the current perturbation scheme and its ability to sample uncertainty within the inner boundary.

2. Data and Models

2.1. Verification Data Set

To calibrate and verify the ensembles we used hourly solar-wind speed data from the High-Resolution OMNI database with ICMEs removed. ICMEs have been removed from the time-series (by converting into data gaps) because we only consider the ambient solar-wind conditions for the calibration; CME perturbations were not included within model conditions. Hence, CMEs were not captured in the ensemble output. The OMNI database is a set of inter-calibrated near-Earth solar-wind observations (King & Papitashvili, 2005). OMNI data are inter-calibrated such that estimates of solar-wind properties at L1 remain consistent when merging data from different satellite instruments. OMNI data are provided through NASA's Space Physics Data Facility (SPDF). ICME crossing times were sourced from the Richardson and Cane (2024) list of Near-Earth Interplanetary Coronal Mass Ejections. We use OMNI data from ranges: June–December 2020 and January–December 2023 (with ICMES removed as described). The hourly-temporal resolution data provides enough forecast samples for good statistics. From here on within this text, all references to resolution refer to statistical resolution rather than physical/temporal.

2.2. HUXt Ensembles

There is uncertainty in the solar-wind speed maps produced by WSA. There are many sources for this uncertainty such as: the assumptions needed to reconstruct the global photospheric field from observations (Heinemann et al., 2025), limited ability to observe and resolve polar field strengths which have big implications on the size and location of the slow-wind band (Petrie, 2015), how discontinuities in the reconstructed coronal field are handled internally by the WSA model (McGregor et al., 2008), and how the speed is derived from an empirical relation to the field topology (McGregor et al., 2011). We approximate the uncertainty by using an ensemble produced by sampling perturbed trajectories of the sub-Earth point (Owens & Riley, 2017). Where the sub-Earth trajectory is the projection of Earth's latitudinal position onto the Sun tracked across longitude. Figure 1 demonstrates how we generate the solar-wind ensembles. We rotate the WSA solution which provides the inner-boundary of near-Sun solar-wind speeds which are then propagated out to Earth by HUXt. By rotating the speed map along the latitudinal and longitudinal axes we sample the spatial uncertainty within the inner boundary. Practically, latitudinal rotations take the form of a sinusoidal augmentation of the sub-Earth path, with a random phase offset setting the axis of tilt (as in Figure 1b). As we are only considering the steady-state solar wind in this study, longitudinal rotations are implemented as a timing error on the forecast output (longitude and time are equivalent for a steady-state solar wind as in Figure 1c). The magnitude of these perturbation are randomly generated from Gaussian distributions centered on zero and with variances of $\sigma_{latitude}$ and $\sigma_{longitude}$. More detail surrounding the perturbation scheme and its implementation can be found in Edward-Inatimi et al. (2024) and Owens and Riley (2017). The spatial perturbation scheme is used for pragmatic reasons—it allows an ensemble to be produced without any changes to the coronal model or magnetogram. Thus, we expect the optimum spatial perturbations to be larger than the true spatial uncertainties in order to account for uncertainty sources that are not explicitly included. The calibration procedure we go on to describe aims to optimize $\sigma_{latitude}$ and $\sigma_{longitude}$.

2.3. WSA Coronal Model

The Wang-Sheeley-Arge (WSA) model combines photospheric magnetic-field observations, a magnetic field model, and empirical relationships to predict near-Sun solar-wind conditions (Arge & Pizzo, 2000). The coronal magnetic field structure is modeled using a Potential-Field Source-Surface (PFSS) from 1 to 2.5 r_S and Schatten Current Sheet (SCS) model from 2.5 to 21.5 r_S. PFSS solves the magnetic field structure whilst maintaining $\nabla \times \mathbf{B} = 0$ and $\nabla \cdot \mathbf{B} = 0$, and SCS applies additional constraints by enforcing a fully radial field at the outer boundary. Thus, WSA assumes the magnetic-field structure within the solar corona is a current-free extension of photospheric field. WSA traces these field lines and evaluates the expansion of open magnetic-field lines (lines which do not form closed loops within a set radius from the photospheric inner-boundary) and the distance from the coronal hole boundary. Solar wind speed is estimated using an empirical relationship which relates these two properties of the coronal magnetic field topology (Riley et al., 2001, 2015; Wang & Sheeley, 1990). These properties are used to estimate solar-wind speed at 1 AU using the WSA equation (Arge & Pizzo, 2000; Sheeley, 2017). As HUXt involves an acceleration of the solar wind between 21.5 and 215 r_S (Riley & Lionello, 2011), we must decelerate the WSA speeds for use as the HUXt inner boundary conditions (Barnard & Owens, 2022).

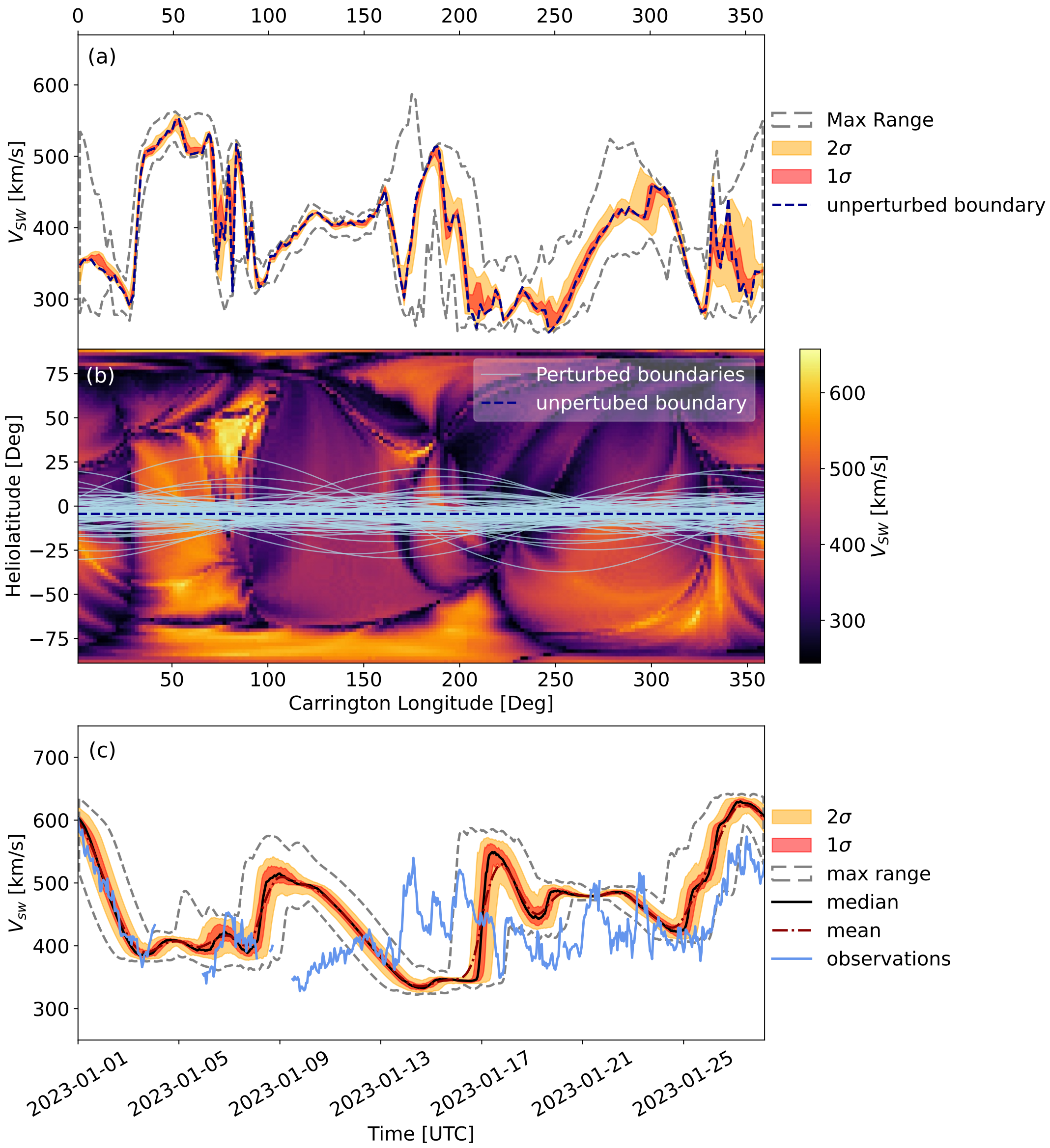

Figure 1. Example of perturbations applied to sub-Earth trajectory used to generate inner-boundaries for HUXt from a WSA solution driven by hourly GONG observations. Panel (a) shows the distribution of inner-boundary velocities across longitude as sourced from panel (b) a WSA solution with set of perturbed sub-Earth trajectories/boundaries plotted in light blue. The unperturbed sub-Earth path is marked with the dark blue dashed line. Panel (c) displays the resulting HUXt ensemble output from the set of inner-boundary conditions generated above with additional longitudinal perturbations added as a timing error on individual ensemble members. For both inner-boundaries and HUXt ensemble distributions 1σ and 2σ spreads are shaded in red and orange. Max range marked by gray dashed lines. Ensemble mean shown with a dashed-dotted line, median shown with a solid black line. Observations sourced from OMNI plotted in light blue.

3. Methods

3.1. Calibration

Ensembles were generated using daily-updated WSA solutions from June–December 2020 ($N_{days} = 195$) and January–December 2023 ($N_{days} = 365$) using GONG magnetograms (Harvey et al., 1996). These two periods allow us to capture model behavior from around solar minimum and maximum respectively. Note: WSA + HUXt forecasts from January–May 2020 show consistent, highly skewed distributions with an average error of +114 km/s and mean bias of +106 km/s. We believe this was caused by a systematic bias within the input magnetograms for WSA during this period hence the data during this time were not included in our calibration, as

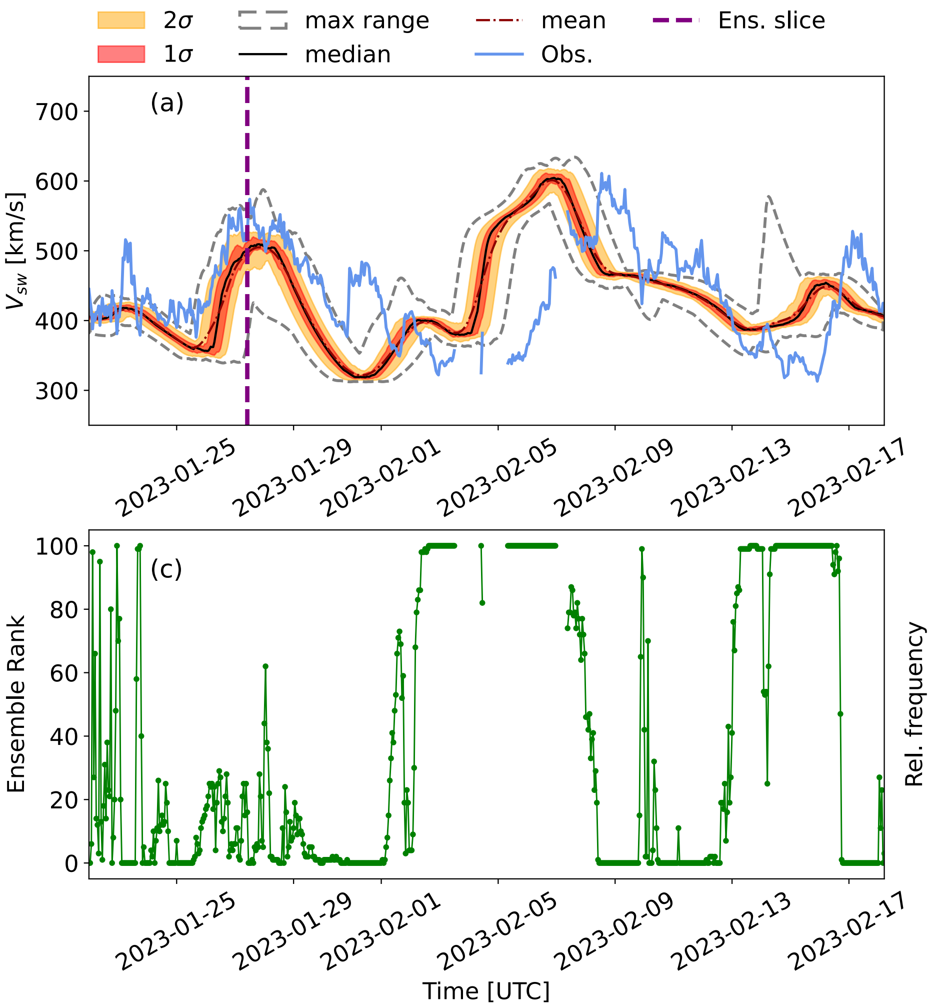

Figure 2. Multipanel figure which demonstrates how the rank histogram is constructed. Panel (a) shows an example ensemble forecast with observations overplotted in blue. Panel (b) shows the ensemble distribution at the marked time in (a) with observation marked with blue line. Panel (c) shows a timeseries of evaluated ranks with respect of the above observations for each timestep of the ensemble forecast. Finally, panel (d) shows the resulting rank histogram from the evaluated timeseries of ranks in (c).

the results would not be generally applicable. Figure A1 shows the model error and bias for all deterministic WSA + HUXt model runs across 2020. The shaded region highlights data excluded from the calibration study. For every WSA solution we generate 100 individual ensemble members of HUXt output from the spatial perturbations applied to the WSA map. We repeat this process for every WSA map across each study period. This produces a set of N_{days} 100-member ensembles we can evaluate. A unique ensemble set is generated with respect to a fixed $\sigma_{latitude}/\sigma_{longitude}$ pair which defines the scale of spatial perturbation applied to the source WSA maps. To explore the calibration, we evaluate all unique pairs of $\sigma_{latitude}/\sigma_{longitude}$ between 0° and 40° in 1° increments resulting in 41 × 41 (1,681) unique ensemble sets for each study period.

Calibration was achieved through finding the optimal scale of perturbation to WSA which balanced HUXt forecast reliability and resolution. Gaussians with standard deviations $\sigma_{latitude}$ and $\sigma_{longitude}$ control the extent of spatial perturbations (i.e., the magnitude of the rotations applied to the WSA output prior to HUXt input) used to generate perturbed-boundary ensembles. We explore this parameter space to find the optimal scales which improve the ensemble distribution, as measured by a range of forecast performance metrics. We detail the performance metrics used in the following sections. These metrics have been chosen to provide a broad overview of different aspects of the ensemble performance. We wish to understand how performance improves/changes not only to better match observations (reducing forecast error and improving reliability) but as a decision making tool

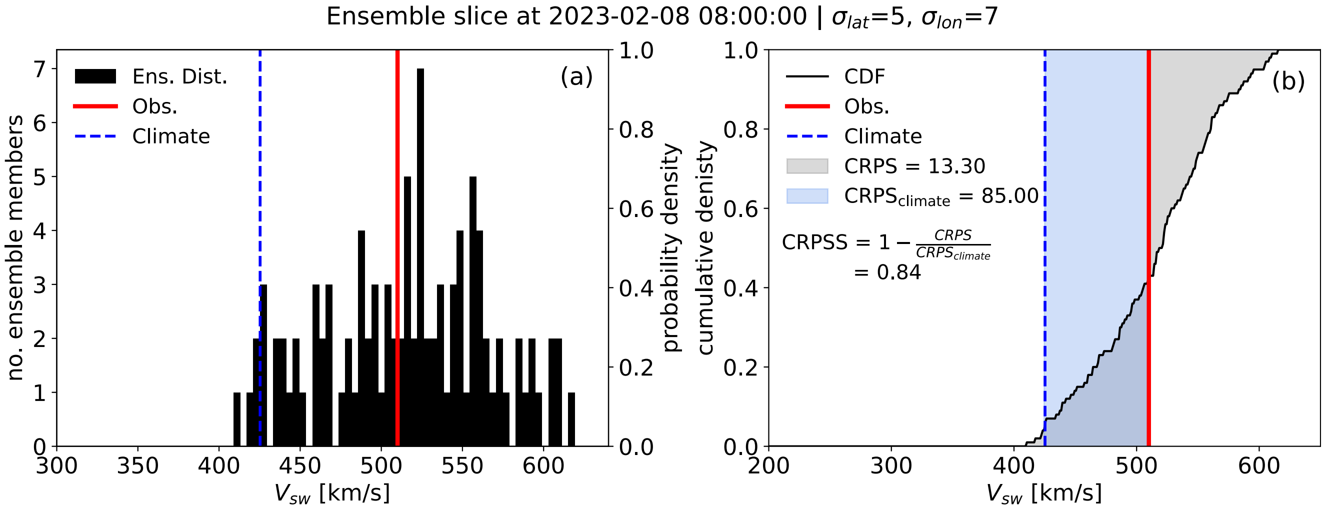

Figure 3. A demonstration of how the continuous ranked probability score (CRPS) gets evaluated at one timestep of the ensemble. The left panel (a) shows histogram of HUXt ensemble slice with the observed wind speed marked by a red line and the climatology forecast in the blue dashed line. The right panel (b) shows the resultant cumulative distribution function (CDF). The gray shaded area highlights the bounded area between the observed CDF (red line) which makes up the estimate of the CRPS, the blue shaded area shows the trapped area between the observation on climatology prediction.

(improving resolution and forecast value). The optimal parameters found through this procedure should better represent the scales of spatial uncertainty within the inner-boundary conditions; whilst still allowing the ensemble to capture and distinguish between events.

3.2. Rank Histogram

As a first step, the ensemble distribution was examined through a rank histogram. A forecasts' rank is the number of members within an ensemble that overestimate an observation. We can evaluate the ensemble ranks for a complete time series as demonstrated in Figure 1 (panels a to c). Then, we can plot the distribution of these ranks as in Figure 2d which yields a rank histogram. The rank histogram evaluates the direct impact of the perturbation scale on the ensemble distribution. A flat rank histogram indicates an ideal distribution where an observation can lie anywhere within the ensemble spread, equally, at all ranks. This is an indication of a well-calibrated ensemble which is accurately sampling the uncertainties. The example shown in Figure 2d is particularly noisy due to the small sample size of the ensemble snippet. The interpretation of the histogram improves as more data is added. We note that the large spike in frequencies at ranks 0 and 100 usually persist, even with larger data sets, due to imposed maximum/minimum speed limits based on model settings. The distance from the perfect distribution can be further quantified using an augmented χ^2 parameter (Wilks, 2019b) such that finding the ranges of perturbation scales which minimize χ^2 can be used as a basis for calibrating the ensemble. However, calibration through use of the rank histogram alone has limits because the rank histogram can only describe the ensemble distribution in the aggregate and says little of the effectiveness of said distribution. Effectiveness in this context refers to the model performance/forecast evaluation. Additionally, whilst rank histograms are useful indicators of calibration, model systems are not perfect and the 'best' rank histogram will likely still show signatures of bias based on the fundamental characteristics/limitations of the model.

3.3. Brier Score Composition

The brier score (BS) is a useful single-metric indicator of forecast accuracy/performance. In its simplest form, it is a measure of the mean squared error between forecast probabilities and observed events (Brier, 1950; Wilks, 2010). BS can be decomposed into three components: reliability (REL), resolution (RES), and uncertainty (UNC) which provide a more nuanced picture of forecast performance (Murphy, 1973).

$$BS = REL - RES + UNC = \frac{1}{N}\sum_{k=1}^{K} n_k\left(\mathbf{f_k} - \bar{\mathbf{o}}_\mathbf{k}\right)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{\mathbf{o}}_\mathbf{k} - \bar{\mathbf{o}})^2 + \bar{\mathbf{o}}(1 - \bar{\mathbf{o}}) \tag{1}$$

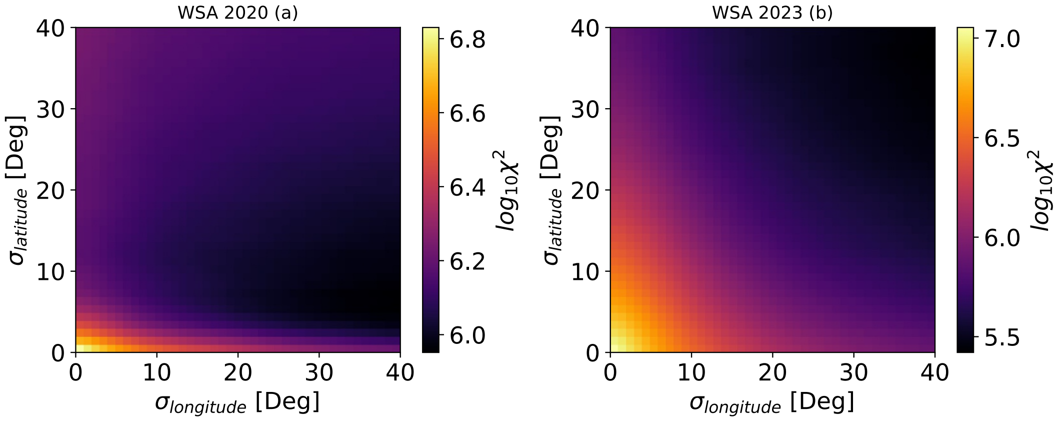

Figure 4. Colormaps showing the impact of perturbation scale on WSA + HUXt ensemble distributions through evaluation of rank histograms using χ^2. Colormaps display $\log_{10}\chi^2$ of rank histograms with respect to the Gaussian variances $\sigma_{latitude}$ and $\sigma_{longitude}$ which control the spread of input conditions for the evaluated ensemble. Panels from left to right, (a) WSA solutions from 2020 representing activity during solar minimum (b) WSA solutions from 2023 representing solar maximum.

where N is the number of forecasts, K is the number of probability bins (i.e., number of unique forecasts), n_k is the number of forecasts inside the Kth bin. f_k is the forecast probability, $\bar{o}$ is the climatological probability calculated as the relative frequency of observations above a threshold over a long time period, and $\bar{o}_k$ is the observed frequency of events associated with forecast probabilities f_k. REL quantifies how well forecast probabilities match the observed frequencies of events. It is a direct indicator of the calibration and is calculated above in Equation 1 as the mean-squared difference between binned forecast probabilities with the respective observed frequencies. A lower value of REL indicates a model which accurately represents the true observed frequency (i.e., binned forecast probability of 30% are associated with events which occurred 30% of the time). RES describes the ability of a model to discriminate between classes of events beyond climatology. It is a measure of forecast decisiveness and is calculated as the mean-square difference of the number of forecasts which fall in line with the observed probabilities subtracted against climatology. A higher value of RES indicates a forecast which can better distinguish between events and non-events by assigning unique probabilities which can be very different from the climatology. We would expect a low-RES ensemble to always assign probabilities to events which reflect a climatology regardless of if an event/non-event is more likely, whereas a high-RES ensemble could assign distinct probabilities for different events which may be very different from the observed frequency. Lastly, UNC is a measure of the baseline uncertainty of an event. It is calculated as above, hence, UNC is maximized when there is always around a 50% chance of an event occurring for example, an event threshold set to the median observed wind speed (approx. 400 km/s) produces an uncertainty of ∼0.25 (a 25% uncertainty), whilst a rarer event threshold of 600 km/s produces a much lower uncertainty of ∼0.05 (a 5% uncertainty). These three components help unpack different aspects of the forecast behavior which contributed to the overall forecast performance. As a rule of thumb, improving one of these components reduces the performance of the other (Toth et al., 2006). So we use these to find a balance between reliability and resolution which produces the most valuable forecast. The REL and RES results we present in later sections were computed using the 90th percentile of observed wind speeds during 2020 and 2023 as the event threshold (507 and 553 km/s respectively).

3.4. Continuous Ranked Probability Score (CRPS)

To evaluate BS we treat the ensemble as a binary classifier (e.g., evaluating the fraction of ensemble members above a certain wind speed threshold). This is a useful simplifying step but in reality the solar-wind speed is a continuous variable. The CRPS is a proper scoring technique which directly compares the ensemble distribution of a given variable with observations. It allows us to treat the solar-wind as a fully continuous variable and provides a comprehensive evaluation of a probabilistic forecast by capturing more information about the forecast distribution than a simple BS. CRPS compares the ensemble cumulative distribution function (CDF) against an observation, through evaluating the bounded area between the ensemble CDF and a step function defined by an observation. The way this is done is outlined in Figure 3.

The formal definition of CRPS as defined by Matheson and Winkler (1976), Hersbach (2000, and references therein) can be written as:

$$CRPS = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx \tag{2}$$

where P and P_a are the ensemble and observed CDFs. P is generated from a probability distribution function. The observed CDF P_a can be described as a step function:

$$P_a(x) = H(x - x_a) = \begin{cases} 0 & \text{for } x < x_a \\ 1 & \text{for } x \geq x_a \end{cases} \tag{3}$$

where x_a is set to the value of the observed windspeed. We have 100 ensemble members and so the CDFs are reasonably well characterized allowing us to directly evaluate the area between the curves using simple trapezoidal integration algorithms instead of relying on estimation methods such as summing over discrete elements, or fitting a CDF through quantiles, which is often necessary when there are not enough ensemble members to provide a smooth empirical CDF (Zamo & Naveau, 2018).

CRPS is evaluated at each time step of the ensemble. Furthermore, the CRPS for a certain period of interest is the mean value within that window. Notably, the CRPS is a generalization of MAE as applied to a probabilistic forecast. This is most obvious when evaluating CRPS for a point forecast (e.g., deterministic forecast or climatology) where Equation 2 collapses into the mean absolute difference between the predicted value and observation. CRPS can be converted into a skill score with respect to a reference CRPS. We use climatology as the reference, such that a score above 0 indicates forecast performance above climatology, and a score of 1 indicating a perfect forecast:

$$CRPSS = 1 - \frac{CRPS}{CRPS_{climate}} \tag{4}$$

3.5. Cost Loss Analysis

Cost/loss analysis evaluates the usefulness of forecasts in a range of hypothetical applications, where the appetite for risk varies substantially. The Potential Economic Value (PEV) of forecasts compares the cost of taking preventive action against the potential losses if no action is taken, for example, the cost of putting a satellite into a safe operating mode versus the cost of damage to said satellite if no action is taken and found to be needed. The analysis helps determine optimal decision thresholds by considering both the cost of protection (C) and the potential loss (L) that could occur if an event happens without protection. This framework is useful in operational forecasting, as it provides a quantitative way to assess how ensemble forecasts impact decisions under uncertainty, whereas the CRPS only assess the accuracy of the ensemble distribution, independent of user-context. This allows us to evaluate if the calibration produces a more actionable forecast. We compute the PEV of a forecast across the range of C/L regimes. PEV is calculated as:

$$PEV = 100\frac{E_c - E}{E_c - E_0} \tag{5}$$

where E is the total expense from the forecast, E_c is the total expense from climatology, and E_0 is the expense from a perfect forecast. A PEV over 0 indicates economic value beyond a climatological forecast. A more complete overview of this method as applied to solar-wind ensembles can be found in Owens and Riley (2017). We evaluate the most calibrated ensembles along with some useful comparisons to demonstrate how the perturbation scale has a direct impact on forecast value.

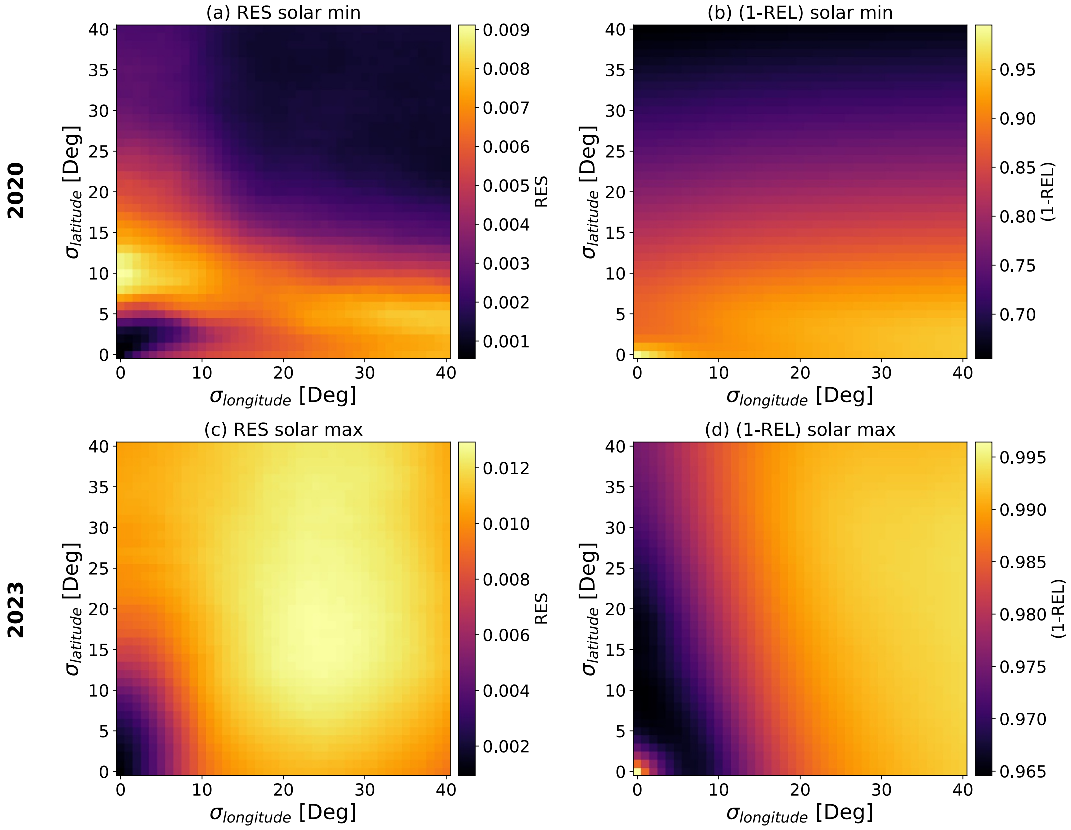

Figure 5. Colormaps of the resolution (RES) and reliability (REL) component of brier score with respect to Gaussian variances $\sigma_{latitude}$ and $\sigma_{longitude}$ which control the spread of input conditions for the evaluated ensemble. Panels (a and b) show results from 2020 (low solar activity period), (c and d) show results from 2023 (high solar activity period). REL component has been plotted (1-REL) to match RES component behavior that is, closer to 1 indicates a higher performance.

4. Results

4.1. Rank Histogram Analysis

Figure 4 shows the results of the rank histograms analysis of WSA ensembles generated with different magnitudes of spatial perturbations. Lower values indicate better calibrated ensembles. Figure 4a shows results for 2020, which represent periods of low activity. There is a well-defined lower bound in $\sigma_{latitude}$ of 4°, above which the ensemble calibration improves, as seen through the sharp drop in χ^2. However, there is no such cut-off in $\sigma_{longitude}$. Figure 4b shows results from 2023, which act as a representation for solar maximum. More extreme levels of perturbation are required to achieve an improved ensemble calibration during this interval. However, the scales of perturbation required to achieve the best calibration are likely unphysical and instead indicate that there is a significant limitation within the spatial perturbation scheme. that is, at solar maximum, there are other significant sources of uncertainty exist which are not being accounted for.

We note that the behavior of the rank histograms with latitudinal and longitudinal perturbations agrees well with those obtained for the Magnetohydrodynamic Algorithm outside a Sphere (MAS; Linker et al., 1999) model for both solar minimum and maximum (Edward-Inatimi et al., 2024).

However, the rank histogram analysis only tells half of the story. Calibration is unbounded at high $\sigma_{longitude}$ because the rank histogram describes the ensemble distribution naively and says little of actual model skill and ability to distinguish between events and non-events. In addition to the largest values likely being unphysical, there exists a trade-off between perturbation scales and the forecast resolution. Calibration acts to balance the ensemble reliability, the ability to accurately recreate real-world likelihoods, versus resolution, the ability for the forecast to clearly distinguish between events and non-events. Hence, the optimal perturbation scale will be found

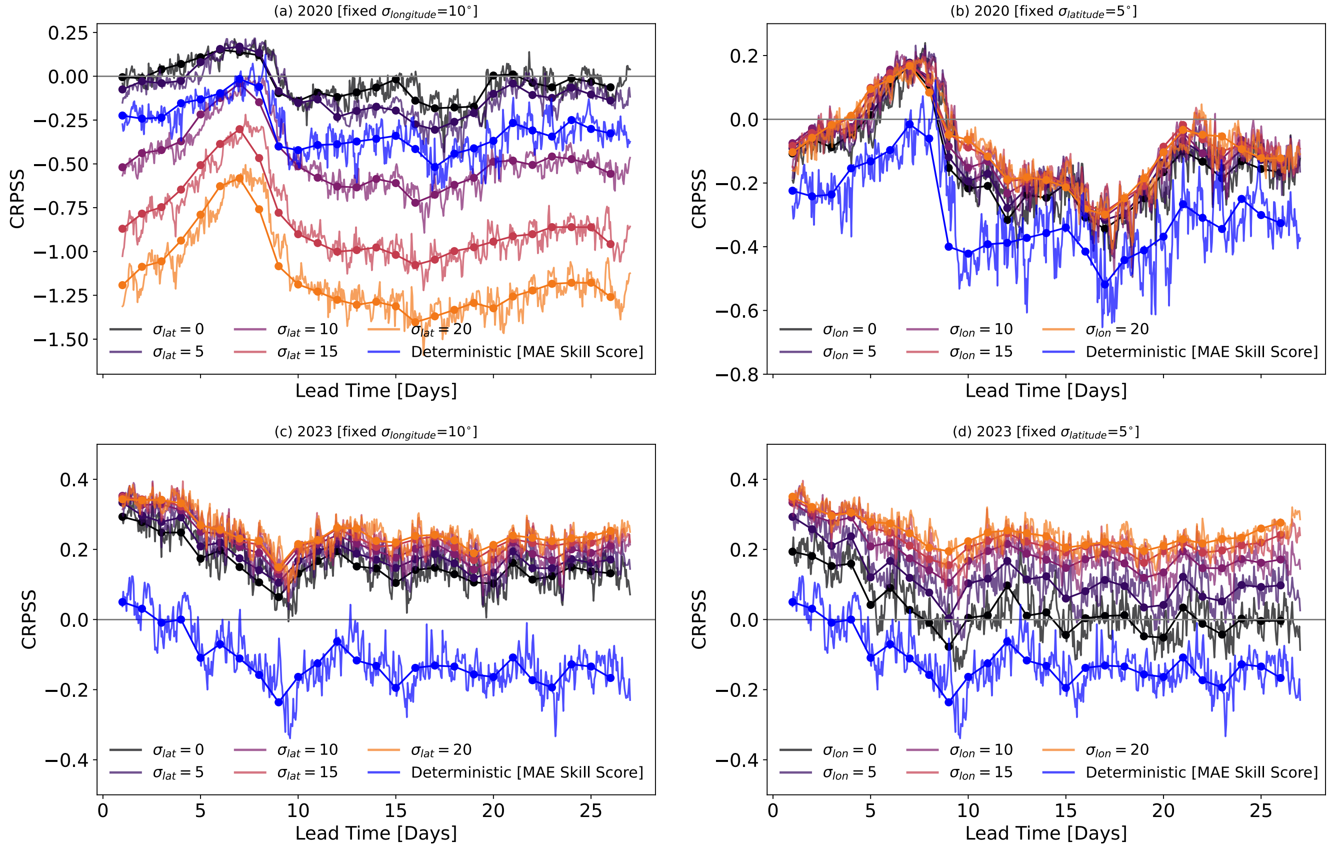

Figure 6. Mean CRPSS (purple to orange lines) and daily mean CRPSS (markers) for 2020 (panels a and b) and 2023 (panels c and d) WSA-HUXt ensembles. Each panel holds either $\sigma_{latitude}$ or $\sigma_{longitude}$ fixed to a calibrated value whilst varying the other to examine the impact of each parameter and the calibration on the ensemble performance across lead time. CRPSS above the 0 line indicate positive forecast skill relative to climatology. All sets are compared with an unperturbed deterministic WSA-HUXt forecast (blue) as an additional baseline.

in the levels which optimize reliability without reducing the resolution below some threshold of usefulness (e.g., below climatology). Finally, we want to bring in our knowledge of the system, to ensure that the optimal scales of perturbation remain within physically reasonable/justifiable sizes.

4.2. Brier Score Components

Figure 5 shows colormaps of the RES and REL components of the BS during 2020 and 2023. Here we have plotted REL as (1-REL) such that high values are better for each metric. In Figure 5a we see a well defined locus of high RES during 2020 which begins at scales $\sigma_{latitude} > 5°$ and $\sigma_{longitude} \gtrsim 10°$. There is also a localized maxima around $\sigma_{latitude} = 10°$ and $\sigma_{longitude} = 0°$. Outside of these areas of parameter space, the resolution drops off again; this is likely related to the width of the slow wind band at solar minimum. As the scale extends beyond the average width of the slow wind band, as both unrealistic levels of variability and a systematic high-speed bias are introduced into the inner-boundary conditions (and subsequent HUXt ensemble), as highly perturbed sub-earth trajectories sample higher latitudes and will hence faster wind. Consequently, the resolution component is greatly penalized as the ensemble gets worse at distinguishing between events (since the ensemble is more likely to predict high-speed streams when none are observed).

Figure 5b shows that only small perturbations can retain some level of reliability with zero perturbation producing the most reliable forecast. REL remains roughly constant across $\sigma_{longitude}$ indicating that the timing of events has less of an impact than the ambient wind structure itself. The plotted 1-REL monotonically decreases (i.e., reliability worsens) with $\sigma_{latitude}$. From this we can conclude that smaller latitudinal perturbations are needed for low activity periods. Thus $\sigma_{latitude}$ and $\sigma_{longitude}$ values below 15° best balance reliability and resolution at solar minimum.

The solar maximum (2023) data shown in panels c and d highlight how REL and RES often work against each other. The regions of good reliability below 15° share a limited overlap with regions of good resolution. There is a small region of low perturbation scales (a few degrees for both $\sigma_{latitude}$ and $\sigma_{longitude}$) which produces more reliable ensembles during 2023. But this is associated with very low resolution in 2023. At larger perturbation scales, the 2023 REL mirrors the 2023 rank histogram analysis, showing that the levels of variability required to create a more reliable forecast can only be achieved through very large perturbations. Hence, calibration during high-activity periods is heavily constrained by the overlap of the RES component between 5 and 10°, and the REL component between 0 and 5°.

4.3. CRPSS

The BS components provide a broad overview of the forecast performance. However, we can now use the CRPSS to evaluate the forecast performance with respect to lead time to see how the perturbation scales impact the forecast skill at different points within across forecast window. We evaluate the CRPS for each timestep across every ensemble and then average along each timestep to get the mean CRPS behavior across lead time. Furthermore, the daily lead-time behavior was calculated by binning the CRPS into daily sets and averaging. CRPSS is calculated using the climatology for the respective periods of each ensemble set (i.e., mean wind-speed June–December 2020, and January–December 2023).

Based on the BS components a calibrated ensemble with $\sigma_{latitude}, \sigma_{longitude} = 5°, 10°$ was chosen as a baseline. We compare the behavior across lead time with increasingly perturbed forecasts. The unperturbed deterministic forecast is also shown. Figure 6a shows that increasing $\sigma_{latitude}$ beyond the calibrated range (>5°) creates a much bigger penalty on forecast performance during 2020. As explained above, this is likely due to the latitudinal perturbation scale increasing beyond the width of the slow wind band and introducing much higher levels of variability and fast wind than are present at Earth. Figure 6b shows that the longitudinal perturbation has a lesser effect on CRPSS. Figures 6c and 6d show that, for the 2023 ensembles, any level of perturbation improves performance over a deterministic model. Increasing the parameter further beyond the calibration value of 10° does increase CRPSS, but it shows diminishing improvement at increasingly large values. It is likely that increased perturbation scales still do not adequately capture the higher levels of variability during solar maximum.

These behaviors mean that the range of calibrated $\sigma_{longitude}$ can be more easily constrained by the levels which provide the best resolution without harming the skill by a great degree. During 2023 increasing $\sigma_{longitude}$ has limited impact at short lead times out to 3 days. After that, larger $\sigma_{longitude}$ consistently improves the forecast skill.

During 2020 CRPSS gets maximized at roughly 6–7 days. Ambient solar wind typically takes between 3 and 5 days to propagate from the corona into near-Earth space. As such, the observations used to construct the daily magnetogram used as input into WSA best represent the structures modeled (and observed) at Earth after that 3–5 days window as shown in Owens et al. (2024). Thus the increase in skill around 6–7 days is somewhat counter-intuitive. It may simply be the result of the limited 7-month period under consideration.

4.4. Cost/Loss

As was done for the REL and RES calculations, we use the 90th percentile of observed wind speeds during 2020 and 2023 as the event threshold (507 and 553 km/s respectively). Following the same procedure used with CRPSS, PEV was first evaluated at a fixed $\sigma_{latitude}$ of 5° and varying $\sigma_{longitude}$ to $5°$, $10°$, and $15°$. Second, $\sigma_{longitude}$ was then fixed to 10° with $\sigma_{latitude}$ varied between $5°, 15°$, and $20°$. An unperturbed deterministic forecast was also evaluated for a baseline comparison. This produced Figure 7.

During 2020 (panels a and b), not much value is gained beyond climatology at any cost/loss ratio. The ensemble also generally has lower PEV than the deterministic model. This was an unexpected result given that the CRPSS indicated that, within the calibrated scales, the ensemble outperformed the deterministic run. It should be noted, at larger perturbation scales, the ensemble show much higher PEV than the deterministic model, approaching climatology in the high C/L regime where false positives incur the highest penalties. This highlights a crucial component of the calibration in which an end-user decision needs to be made between the accuracy of the model versus the decisiveness of the output.

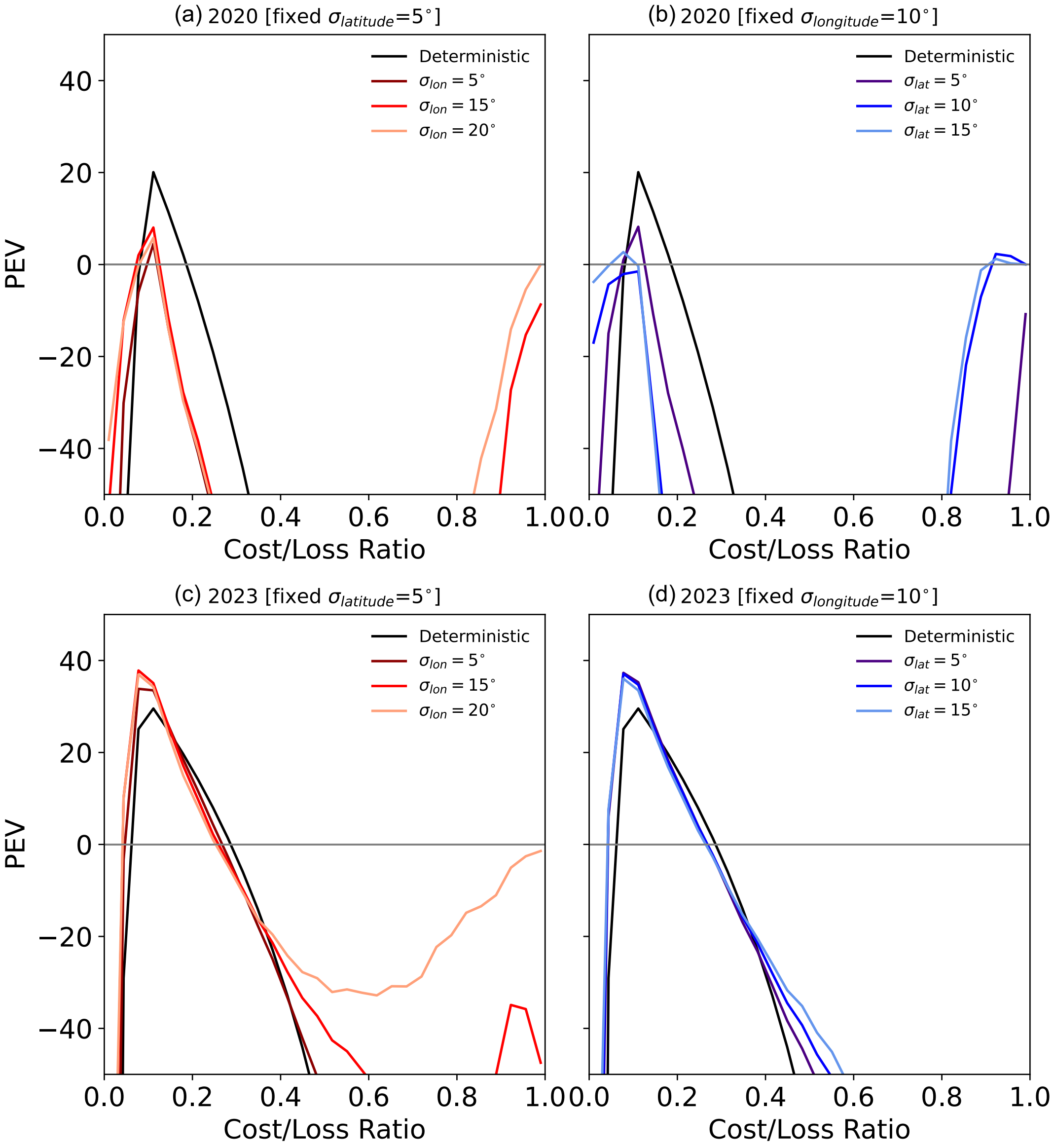

Figure 7. Plots of potential economic value (PEV) as a function of different cost/loss ratios for a range of forecasts for 2020 (a and b) and 2023 (c and d). Panels (a and c) hold $\sigma_{latitude}$ fixed at 5° whilst increasing $\sigma_{longitude}$. Panels b and d hold $\sigma_{longitude}$ fixed at 5° whilst increasing $\sigma_{latitude}$. The PEV for an unperturbed deterministic WSA-HUXt forecast is plotted in black as a comparison point.

During 2023 (panels c and d) the ensembles shows higher PEV than the deterministic model for all values of $\sigma_{longitude}$ and $\sigma_{latitude}$ considered. However, lower perturbation scales (closer to the ideal calibrated values) appear to provide subtly larger PEV, except in the high cost/loss regime.

5. Discussion

Combining the information from the four avenues of investigation—using the rank histogram analysis to set a lower limit on the perturbation scales, BS components to identify ranges which provide better reliability and resolution, CRPSS to evaluate the impact of perturbation scales at increasing lead times, and finally a Cost/Loss analysis to check if the calibrated scales provide more PEV over under/over perturbed ensembles. We can now construct a set of parameter ranges which produce a more calibrated ensemble: $4° \leq \sigma_{latitude} \leq 10°$ and $6° \leq \sigma_{longitude} \leq 15°$

Based on the CRPSS, even small levels of perturbation improved the forecast, particularly during high activity periods. This indicates that the ensemble methods are indeed capturing useful information about the uncertainty which aids the overall forecast accuracy. As expected, the forecast performance drops with increasing lead time. However, the drop in performance is reduced with larger longitudinal perturbation scales, which allow for the temporal uncertainty to be better characterized. In operational forecasting, this is something which can be exploited through the use of multi-model ensembles. A set of ensemble outputs, each tuned for different lead times, can be combined through some statistical post-processing to produce a more comprehensive forecast for lead-times beyond 1–3 days (see Allen et al., 2020; Roberts et al., 2023, and references therein).

Typically in forecasting, resolution is prioritized over reliability, meaning a sharper ensemble distribution is prioritized over a perfectly accurate representation of the uncertainties/likelihoods of events (Toth et al., 2006). In refining the calibration procedure, it is clear that the spatial perturbations that we currently employ for pragmatic reasons are insufficient. Other methods for perturbing the inner boundary need to be adopted. The rank histogram analysis clearly demonstrates that the current perturbation scheme struggles to recreate the true levels of variability, especially during solar maximum (Figure 4b). This indicates, at present, there are other aspects of uncertainty not being accurately represented within the ensemble. The empirical relationships used to estimate the near-Sun solar-wind speed from the magnetic topology is a clear source of uncertainty. Hence, future work could be done to constrain the uncertainty within the WSA equation through direct perturbation of the free parameters within that equation. This would be a natural extension of the ensemble methods outlined in Reiss et al. (2020). As per Edward-Inatimi et al. (2024), the analysis was limited to ambient conditions. Next steps should also involve introducing CME perturbations into the ensemble to see how the calibration impacts CME arrival time forecasts. This could aid in determining the contribution of ambient wind uncertainty on CME arrival time errors. Once CMEs are added, we can then begin to fully explore post-processing methods which can be used to augment the ensemble output directly to improve the calibration.

Ultimately, it is the priorities of the end-user that should inform the exact scales of spatial perturbation (within the bounds of the optimal scales found in this study). If physical accuracy is a priority then we recommend using larger perturbation scales which better represent the true scales of uncertainty within the coupled coronal-heliospheric model framework. In forecasting applications, where clear decision thresholds often take priority, we recommend selecting parameters toward the lower bounds of the calibration as they will allow for a more decisive forecast which provides more economic value. For the WSA-HUXt operational ensemble (https://research.reading.ac.uk/met-spate/huxt-forecast/) we have adopted $\sigma_{lat} = 5°$, $\sigma_{lon} = 8°$.

6. Conclusion

We have gone through a calibration procedure for a solar-wind ensemble forecast using WSA solutions as the inner-boundary for the HUXt solar-wind model. An ensemble of HUXt simulations are generated by spatially perturbing the WSA solutions. The extent of perturbations are controlled by two perturbation scale parameters, $\sigma_{latitude}, \sigma_{longitude}$. The calibration tunes these parameters and balance changes the model's ability to produce realistic probabilities with the need for decisive and actionable forecasts. Through this calibration process we have unpacked various aspects of the forecast performance:

1. We use the rank histogram to demonstrate how the size of spatial perturbation impacts the ensemble distribution, ultimately finding that spatial perturbations alone do not adequately capture levels of uncertainty during solar maximum. We intend to investigate using the empirical ensemble methods discussed in Reiss et al. (2020) to begin to tackle this problem.
2. We use reliability and resolution components of BS to identify the aspects of the forecast performance that get optimized at different perturbation scales. We find REL and RES components often work against each other. Through exploring the $\sigma_{latitude}, \sigma_{longitude}$ parameter space we could constrain the ranges of parameters which balance/find a compromise between the two components. We find during solar minimum, especially when there is a narrow slow-wind band, larger perturbation scales risk producing much higher levels of variability than really exist. This added an additional constraint on the calibration to keep scales within physically consistent ranges as $\sigma_{latitude}, \sigma_{longitude}$ beyond 15° showed diminishing improvements in forecast performance, often harming the resolution/decision making ability of the ensemble.

3. We use CRPS to further evaluate the ensemble performance and provide information about ensemble skill at various lead times. When considering the ensemble distribution as a whole, any amount of perturbation improved the overall forecast performance. Increased perturbation scales have diminishing improvements on performance out to around 3-day lead times. Beyond which, larger perturbation scales better capture uncertainties at lead times out to 10 days. However, larger perturbation scales harm resolution at shorter lead-times.
4. Finally, we use a cost/loss analysis to highlight how the calibrated ensemble provided more value than uncalibrated ensembles and a deterministic forecast (specifically for rare events) during high-activity periods.

Calibration is required as optimal perturbation scales are heavily model dependant. WSA is the coronal model currently used operationally by both UKMO and SWPC coupled with Enlil, and the SWEEP coupled WSA-HUXt ensemble forecast. As a result of the outlined calibration procedure, we adopt parameters $\sigma_{latitude} = 5°$ and $\sigma_{longitude} = 8°$ for the WSA-HUXt operational scheme. For our purposes we find these parameters strike the ideal balance between reliability and resolution within the forecast.

Appendix A: Forecast Bias During Early-2020

Figure A1 shows the period of model runs during 2020 which were not included in the calibration study due to anomalous behavior seen within the bias and average error during Jan-May.

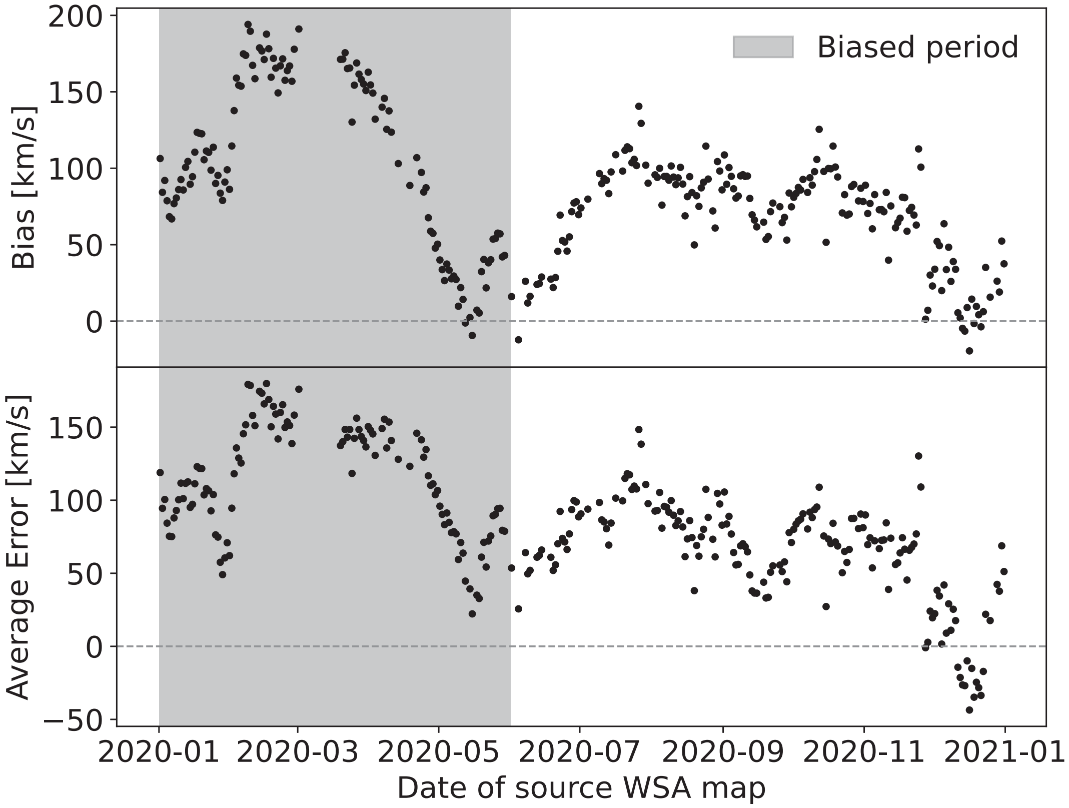

Figure A1. Plot of WSA-HUXt forecast bias calculated as $\mu_{forecast} - \mu_{observations}$ on the top panel and average error calculated as $\overline{v_{forecast} - v_{observation}}$ on the bottom panel for all forecasts run against the dates of source WSA maps used to drive the forecast. Gray shaded region shows portion of data not included within calibration study due to higher levels of bias than expected.

Acronyms

BS	Brier Score
CDF	Cumulative Distribution Function
(I)CME	(Interplanetary) Coronal Mass Ejection
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
HUXt	Heliospheric Upwind eXtrapolation with Time dependence
MHD	Magnetohydrodynamics
OMNI	OMNI has no special abbreviation, just "variety"
PEV	Potential Economic Value
REL	Reliability
RES	Resolution
ROC	Receiver Operating Characteristic
WSA	Wang-Sheeley-Arge

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

- Verification data was sourced from OMNI solar-wind observations which can be found: https://omniweb.gsfc.nasa.gov/ftpbrowser/pla_cross_wind_ace.html.
- List of Near-Earth Interplanetary Coronal Mass Ejections Since January 1996 used to remove ICME crossings from OMNI time series found at Richardson and Cane (2024).
- HUXt is an open-source solar-wind model available at: Owens and Barnard (2024) (in this study we used HUXt code version V4.1.1).
- In this study we used model output from WSA V4.5 provided by the UK Met Office. Standard WSA output driven by GONG magnetograms can be publicly accessed through the Community Coordinated Modeling Center (CCMC) at https://iswa.gsfc.nasa.gov/iswa_data_tree/model/solar/WSA5.4/
- Python notebooks written to run WSA + HUXt ensembles and evaluate the calibration can be found at Edward-Inatimi (2026).

Acknowledgments

We would like to thank the data providers for this project. Nathaniel Edward-Inatimi is funded through SCENARIO Grant NE/S007261/1. Harriet Turner is funded through NERC Grant NE/Y001052/1. Luke Barnard is supported by a UKRI Future Leaders Fellowship (MR/Y021207/1). This work was also part-funded by Science and Technology Facilities Council (STFC) Grant ST/V000497/1.

References

Allen, S., Ferro, C. A. T., & Kwasniok, F. (2020). Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, *146*(731), 2576–2596. https://doi.org/10.1002/qj.3806

Arge, C. N., & Pizzo, V. J. (2000). Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. *Journal of Geophysical Research*, *105*(A5), 10465–10479. https://doi.org/10.1029/1999JA000262

Barnard, L., & Owens, M. J. (2022). HUXt—An open source, computationally efficient reduced-physics solar wind model, written in Python. *Frontiers in Physics*, *10*, 1005621. https://doi.org/10.3389/fphy.2022.1005621

Barnard, L., Owens, M. J., Scott, C. J., & de Koning, C. A. (2020). Ensemble CME modeling constrained by heliospheric imager observations. *AGU Advances*, *1*(3), e2020AV000214. https://doi.org/10.1029/2020AV000214

Bertello, L., Pevtsov, A. A., Petrie, G. J. D., & Keys, D. (2014). Uncertainties in solar synoptic magnetic flux maps. *Solar Physics*, *289*(7), 2419–2431. https://doi.org/10.1007/s11207-014-0480-3

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2

Cranmer, S. R., Gibson, S. E., & Riley, P. (2017). Origins of the ambient solar wind: Implications for space weather. *Space Science Reviews*, *212*(3–4), 1345–1384. https://doi.org/10.1007/s11214-017-0416-y

Edward-Inatimi, N. O. (2026). University-of-Reading-Space-Science/WSA_calibration: Code for publication (zenodo release) [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.18337047

Edward-Inatimi, N. O., Owens, M. J., Barnard, L., Turner, H., Marsh, M., Gonzi, S., et al. (2024). Adapting ensemble-calibration techniques to probabilistic solar-wind forecasting. *Space Weather*, *22*(12), e2024SW004164. https://doi.org/10.1029/2024SW004164

Harvey, J. W., Hill, F., Hubbard, R. P., Kennedy, J. R., Leibacher, J. W., Pintar, J. A., et al. (1996). The Global Oscillation Network Group (GONG) project. *Science*, *272*(5266), 1284–1286. https://doi.org/10.1126/science.272.5266.1284

Heinemann, S. G., Pomoell, J., Caplan, R. M., Owens, M. J., Jones, S., Upton, L., et al. (2025). Quantifying uncertainties in solar wind forecasting due to incomplete solar magnetic field information. *The Astrophysical Journal*, *986*(2), 166. https://doi.org/10.3847/1538-4357/adcf9e

Henley, E. M., & Pope, E. C. D. (2017). Cost-loss analysis of ensemble solar wind forecasting: Space weather use of terrestrial weather tools. *Space Weather*, *15*(12), 1562–1566. https://doi.org/10.1002/2017SW001758

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2

Hinterreiter, J., Magdalenic, J., Temmer, M., Verbeke, C., Jebaraj, I. C., Samara, E., et al. (2019). Assessing the performance of EUHFORIA modeling the background solar wind. *Solar Physics*, *294*(12), 170. https://doi.org/10.1007/s11207-019-1558-8

Jian, L. K., MacNeice, P. J., Taktakishvili, A., Odstrcil, D., Jackson, B., Yu, H.-S., et al. (2015). Validation for solar wind prediction at Earth: Comparison of coronal and heliospheric models installed at the CCMC. *Space Weather*, *13*(5), 316–338. https://doi.org/10.1002/2015SW001174

Kennis, S., Perri, B., & Poedts, S. (2024). Magnetic connectivity from the sun to the Earth with MHD models I. Impact of the magnetic modelling for connectivity validation. Retrieved from https://arxiv.org/abs/2409.20217

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ACE plasma and magnetic field data. *Journal of Geophysical Research*, *110*(A2), A02104. https://doi.org/10.1029/2004JA010649

Kumar, S., Paul, A., & Vaidya, B. (2020). A comparison study of extrapolation models and empirical relations in forecasting solar wind. *Frontiers in Astronomy and Space Sciences*, *7*, 572084. https://doi.org/10.3389/fspas.2020.572084

Lee, C. O., Arge, C. N., Odstrčil, D., Millward, G., Pizzo, V., Quinn, J. M., & Henney, C. J. (2013). Ensemble modeling of CME propagation. *Solar Physics*, *285*(1–2), 349–368. https://doi.org/10.1007/s11207-012-9980-1

Linker, J. A., Mikić, Z., Biesecker, D. A., Forsyth, R. J., Gibson, S. E., Lazarus, A. J., et al. (1999). Magnetohydrodynamic modeling of the solar corona during Whole Sun Month. *Journal of Geophysical Research*, *104*(A5), 9809–9830. https://doi.org/10.1029/1998JA900159

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Mays, M. L., Taktakishvili, A., Pulkkinen, A., MacNeice, P. J., Rastätter, L., Odstrcil, D., et al. (2015). Ensemble modeling of CMEs using the WSA-ENLIL+Cone model. *Solar Physics*, *290*(6), 1775–1814. https://doi.org/10.1007/s11207-015-0692-1

McGregor, S. L., Hughes, W. J., Arge, C. N., & Owens, M. J. (2008). Analysis of the magnetic field discontinuity at the potential field source surface and Schatten Current Sheet interface in the Wang–Sheeley–Arge model. *Journal of Geophysical Research*, *113*(A8), A08112. https://doi.org/10.1029/2007JA012330

McGregor, S. L., Hughes, W. J., Arge, C. N., Owens, M. J., & Odstrcil, D. (2011). The distribution of solar wind speeds during solar minimum: Calibration for numerical solar wind modeling constraints on the source of the slow solar wind. *Journal of Geophysical Research*, *116*(A3), A03101. https://doi.org/10.1029/2010JA015881

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, *12*(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012⟨0595:ANVPOT⟩2.0.CO;2

Odstrcil, D. (2003). Modeling 3-D solar wind structure. *Advances in Space Research*, *32*(4), 497–506. https://doi.org/10.1016/S0273-1177(03)00332-6

Odstrcil, D., Linker, J. A., Lionello, R., Mikic, Z., Riley, P., Pizzo, V. J., & Luhmann, J. G. (2002). Merging of coronal and heliospheric numerical two-dimensional MHD models. *Journal of Geophysical Research*, *107*(A12), SSH 14-1–SSH 14-11. https://doi.org/10.1029/2002JA009334

Odstrcil, D., Pizzo, V. J., Linker, J. A., Riley, P., Lionello, R., & Mikic, Z. (2004). Initial coupling of coronal and heliospheric numerical magnetohydrodynamic codes. *Journal of Atmospheric and Solar-Terrestrial Physics*, *66*(15), 1311–1320. https://doi.org/10.1016/j.jastp.2004.04.007

Owens, M. J., & Barnard, L. (2024). University-of-Reading-Space-Science/HUXt: HUXT V4.2.0 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.12772120

Owens, M. J., Barnard, L., & Arge, C. N. (2024). The importance of boundary evolution for solar-wind modelling. *Scientific Reports*, *14*(1), 28975. https://doi.org/10.1038/s41598-024-80162-2

Owens, M. J., Lang, M., Barnard, L., Riley, P., Ben-Nun, M., Scott, C. J., et al. (2020). A computationally efficient, time-dependent model of the solar wind for use as a surrogate to three-dimensional numerical magnetohydrodynamic simulations. *Solar Physics*, *295*(3), 43. https://doi.org/10.1007/s11207-020-01605-3

Owens, M. J., & Riley, P. (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. https://doi.org/10.1002/2017SW001679

Perri, B., Kuźma, B., Brchnelova, M., Baratashvili, T., Zhang, F., Leitner, P., et al. (2023). Coconut, a novel fast-converging MHD model for solar corona simulations. II. Assessing the impact of the input magnetic map on space-weather forecasting at minimum of activity. *The Astrophysical Journal*, *943*(2), 124. https://doi.org/10.3847/1538-4357/ac9799

Petrie, G. J. D. (2015). Solar magnetism in the polar regions. *Living Reviews in Solar Physics*, *12*(1), 5. https://doi.org/10.1007/lrsp-2015-5

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

Reiss, M. A., MacNeice, P. J., Muglach, K., Arge, C. N., Möstl, C., Riley, P., et al. (2020). Forecasting the ambient solar wind with numerical models. II. An adaptive prediction system for specifying solar wind speed near the Sun. *The Astrophysical Journal*, *891*(2), 165. https://doi.org/10.3847/1538-4357/ab78a0

Richardson, I., & Cane, H. (2024). Near-Earth interplanetary coronal mass ejections since January 1996. *Harvard Dataverse*. https://doi.org/10.7910/DVN/C2MHTH

Riley, P., Linker, J. A., & Arge, C. N. (2015). On the role played by magnetic expansion factor in the prediction of solar wind speed. *Space Weather*, *13*(3), 154–169. https://doi.org/10.1002/2014SW001144

Riley, P., & Ben-Nun, M. (2021). On the sources and sizes of uncertainty in predicting the arrival time of interplanetary coronal mass ejections using global MHD models. *Space Weather*, *19*(6), e2021SW002775. https://doi.org/10.1029/2021SW002775

Riley, P., Ben-Nun, M., Linker, J. A., Mikic, Z., Svalgaard, L., Harvey, J., et al. (2014). A multi-observatory inter-comparison of line-of-sight synoptic solar magnetograms. *Solar Physics*, *289*(3), 769–792. https://doi.org/10.1007/s11207-013-0353-1

Riley, P., Linker, J. A., & Mikić, Z. (2001). An empirically-driven global MHD model of the solar corona and inner heliosphere. *Journal of Geophysical Research*, *106*(A8), 15889–15901. https://doi.org/10.1029/2000JA000121

10.1029/2025SW004614

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

Riley, P., Linker, J. A., Mikić, Z., Odstrcil, D., Pizzo, V. J., & Webb, D. F. (2002). Evidence of posteruption reconnection associated with coronal mass ejections in the solar wind. *The Astrophysical Journal*, *578*(2), 972–978. https://doi.org/10.1086/342608

Riley, P., & Lionello, R. (2011). Mapping solar wind streams from the Sun to 1 AU: A comparison of techniques. *Solar Physics*, *270*(2), 575–592. https://doi.org/10.1007/s11207-011-9766-x

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C., et al. (2023). Improver: The new probabilistic postprocessing system at the met office. *Bulletin of the American Meteorological Society*, *104*(3), E680–E697. https://doi.org/10.1175/BAMS-D-21-0273.1

Sheeley, N. R. (2017). Origin of the Wang-Sheeley-Arge solar wind model. *History of Geo- and Space Sciences*, *8*(1), 21–28. https://doi.org/10.5194/hgss-8-21-2017

Toth, Z., Talagrand, O., & Zhu, Y. (2006). The attributes of forecast systems: A general framework for the evaluation and calibration of weather forecasts. In T. Palmer & R. Hagedorn (Eds.), *Predictability of weather and climate* (pp. 584–595). Cambridge University Press.

Wang, Y. M., & Sheeley, J. N. R. (1990). Solar wind speed and coronal flux-tube expansion. *The Astrophysical Journal*, *355*, 726. https://doi.org/10.1086/168805

Wilks, D. S. (2010). Sampling distributions of the brier score and brier skill score under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, *136*(653), 2109–2118. https://doi.org/10.1002/qj.709

Wilks, D. S. (2019a). Chapter 8 - Ensemble forecasting. In D. S. Wilks (Ed.), *Statistical methods in the atmospheric sciences* (4th ed., pp. 313–367). Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00008-0

Wilks, D. S. (2019b). Indices of rank histogram flatness and their sampling properties. *Monthly Weather Review*, *147*(2), 763–769. https://doi.org/10.1175/MWR-D-18-0369.1

Zamo, M., & Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, *50*(2), 209–234. https://doi.org/10.1007/s11004-017-9709-7