Space Weather

RESEARCH ARTICLE
10.1029/2024SW004164

Adapting Ensemble-Calibration Techniques to Probabilistic Solar-Wind Forecasting

N. O. Edward-Inatimi[1], M. J. Owens[1], L. Barnard[1], H. Turner[1], M. Marsh[2], S. Gonzi[2], M. Lang[3], and P. Riley[4]

[1]University of Reading, Reading, UK, [2]UK Met Office, Exeter, UK, [3]British Antarctic Survey, Cambridge, UK, [4]Predictive Science Inc, San Diego, CA, USA

Key Points:

- Ensemble methods capture model uncertainty. Solar-wind ensembles are generated through spatial perturbations to inner-boundary conditions
- Calibration aligns forecast probabilities with observed frequencies, to do so we find ideal perturbation scales for a solar-wind ensemble
- Calibrating the ensemble improved forecast performance; optimal perturbations might link to spatial uncertainty within the inner-boundary

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., et al. (2024). Adapting ensemble-calibration techniques to probabilistic solar-wind forecasting. *Space Weather*, *22*, e2024SW004164. https://doi.org/10.1029/2024SW004164

Received 13 SEP 2024
Accepted 21 NOV 2024

Author Contributions:

Conceptualization: N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, S. Gonzi, M. Lang
Data curation: P. Riley
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: M. J. Owens, L. Barnard, H. Turner, M. Lang
Software: N. O. Edward-Inatimi, M. J. Owens, L. Barnard, H. Turner, M. Lang
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, M. J. Owens, L. Barnard, H. Turner, M. Lang

Abstract Solar-wind forecasting is critical for predicting events which can affect Earth's technological systems. Typically, forecasts combine coronal model outputs with heliospheric models to predict near-Earth conditions. Ensemble forecasting generates sets of outputs to create probabilistic forecasts which quantify forecast uncertainty, vital for reliable/actionable forecasts. We adapt meteorological methods to create a calibrated solar-wind ensemble and probabilistic forecast for ambient solar wind, a prerequisite for accurate coronal mass ejection (CME) forecasting. Calibration is achieved by adjusting ensemble inputs/outputs to align the ensemble spread with observed event frequencies. We produce hindcasts in near-Earth space using coronal-model output over Solar Cycle 24, as input to Heliospheric Upwind eXtrapolation with time dependence (HUXt) solar-wind model. Making spatial perturbations to the coronal model output at 0.1 AU, we produce ensembles of inner-boundary conditions for HUXt, evaluating how forecast accuracy was impacted by the scales of perturbations applied. We found optimal spatial perturbations described by Gaussian distributions with variances of 20° latitude and 10° longitude; these might represent spatial uncertainty within the coronal model. This produced probabilistic forecasts better matching observed frequencies. Calibration improved forecast reliability, reducing the Brier score by 9% and forecast decisiveness increasing AUC ROC score by 2.5%. Improvements were subtle but systematic. Additionally, we explored statistical post-processing to correct over-confidence bias, improving forecast actionability. However, this method, applied post-run, does not affect the solar-wind state used to propagate CMEs. This work represents the first formal calibration of solar-wind ensembles, laying groundwork for comprehensive forecasting systems like a calibrated multi-model ensemble.

Plain Language Summary Current solar-wind forecasting methods combine coronal and heliospheric models to predict physical properties of the solar wind near the Earth. Ensemble forecasting combines many individual forecasts to provide probabilistic predictions and estimate forecast uncertainty. However, the uncertainty is poorly constrained for a naively generated ensemble forecast and needs to be calibrated to be reliable. We adapt established meteorological methods to create a calibrated solar-wind speed ensemble forecast, improving forecast accuracy by adjusting model inputs based on past solar activity. By applying spatial perturbations to model conditions, we find the optimal extent of spatial perturbations that improve forecast reliability, better matching observed event frequencies. This approach is a step toward more comprehensive and reliable solar-wind forecasting systems.

1. Introduction

Forecasting space weather with more than an hour lead time requires forecasting the near-Earth solar-wind conditions on the basis of solar observations, which has many inherent challenges. The solar wind—a continuous outflow of magnetized plasma from the solar corona—propagates radially outwards throughout the solar system (e.g., Cranmer, 2019, and references therein). The ambient solar-wind conditions—speed, density, and magnetic field orientation—underpin space-weather phenomena. In particular, ambient solar-wind conditions affect the dynamics and properties of Coronal Mass Ejections (CMEs) potentially directly affecting their geo-effectiveness (Hosteaux et al., 2019; Temmer et al., 2023). Although high-speed streams can drive minor geomagnetic storms at Earth, CMEs are responsible for the majority of major storms, which risk disrupting power grids, satellite communications, and navigation systems (Cannon, 2013; Schrijver, 2015).

Simulating and forecasting the solar wind at Earth is predominantly done with Hydrodynamic/Magnetohydrodynamic (MHD) numerical models. The boundary conditions for these models are typically the outputs from simulations of the corona. This can also be an MHD model of the corona such as Magnetohydrodynamic

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

Algorithm outside a Sphere (MAS). However, for forecasting MAS is too computationally expensive, so forecasts typically use the Wang-Sheeley-Arge (WSA) coronal model. In this study, we will use MAS due to the availability of a long duration archive. Current operational forecast schemes typically use coupled corona- and heliospheric models (Reiss et al., 2023). The UK Met Office uses the WSA coronal model (Arge & Pizzo, 2000; Sheeley, 2017) to estimate near-Sun (0.1 AU) solar-wind conditions using an empirical relationship between the coronal magnetic field configuration and the solar-wind speed (Wang & Sheeley, 1990). The WSA output is used as the inner-boundary conditions to the Enlil solar-wind model: a 3-D MHD numerical model (Odstrcil, 2003) that simulates the solar-wind flow between 0.1 AU, out past Earth orbit. The coronal model provides an inner-boundary which defines only the steady-state ambient solar-wind conditions. Thus it is necessary to add time-dependent perturbations at 0.1 AU which mimic CMEs. This is achieved using parameters derived from coronagraph observations (Millward et al., 2013; Xie et al., 2004; Zhao et al., 2002) to estimate Earth arrival times of CMEs. However, high-confidence in forecast CME arrival times is hard to achieve if there is a large amount of uncertainty within the ambient solar wind (Mays et al., 2015).

Ensemble methods are used to capture many different aspects of uncertainty from a model or forecast. Perturbed initial-condition ensembles are a method of ensemble forecasting which combines large numbers of deterministically run forecasts with adjusted initial conditions, perturbed away from an assumed ground truth (Wilks, 2019a). The goal of such an ensemble is to determine the sensitivity to the initial conditions, to quantify the uncertainty that is introduced as a result, and to produce a probabilistic forecast. This is a tried and tested method used in meteorological forecasts. Earth's weather systems are inherently chaotic, meaning numerical weather prediction forecasts with very small perturbations to initial conditions diverge rapidly (Zhu, 2005).

Similarly, there is a lot of uncertainty within the solar-wind forecast. However, as the solar wind is an outward supersonic, superalfvénic flow, the initial conditions are rapidly advected out of the model domain and are lost during model spin up; the solar wind state is primarily determined by the inner-boundary conditions (Lang et al., 2017). So whilst solar wind can be chaotic/highly turbulent, the turbulence occurs on length scales well below the grid size we are modelling. The precise impact of mesoscale processes (length scales 5–10,000 Mm at 1 AU, on the order of km scales in near-Sun space) on large-scale dynamics is ambiguous but current understandings generally agree that these processes probably do not have a significant impact on the large-scale flow (see Viall et al., 2021, and references therein). There is a lack of routine in-situ measurements of the solar-wind speed close to the Sun, thus coronal models must be used to provide the inner-boundary conditions to heliospheric models. However, solar-wind speeds obtained from coronal models still largely rely on weak empirical relationships (Riley et al., 2015; Wang & Sheeley, 1990). Hence, large uncertainty is present in the inner-boundary conditions used to model the solar wind. Whilst a coupled model like WSA-Enlil succeeds in many ways as a forecasting tool, a large drawback is that it is difficult to infer levels of forecast uncertainty from a single/handful of deterministic forecasts. Running a large ensemble regime with WSA-Enlil is computationally expensive, which ultimately limits the quality of ensemble statistics that can be achieved (Mays et al., 2015). One means to improve ensemble statistics is to use a more computationally efficient model. We here use Heliospheric Upwind eXtrapolation with Time dependence (HUXt) model to generate solar wind ensembles. HUXt is a reduced-physics solar-wind model developed by Owens et al. (2020); Barnard and Owens (2022). It is a time-dependent form of the 1-D upwind mapping scheme developed by Riley and Lionello (2011). Despite the gross simplifications employed, HUXt has been shown to emulate 3-D MHD model output, such as Enlil and EUHFORIA, at a fraction of the computational cost (Owens et al., 2020; Pomoell & Poedts, 2018).

Probabilistic solar-wind forecasting through the use of large ensembles has been examined by Owens and Riley (2017) and many of the methods we employ here build from those methodologies. The fundamental assumption is that we can construct a useful ensemble by applying spatial perturbations to the coronal-model output. There are, of course, many other sources of uncertainty, but explicitly incorporating them requires re-running the coronal model itself, with different boundary conditions or parameterisation schemes. Our simpler, pragmatic approach—the validity of which is tested here—has the benefit that the ensemble can be generated from a single coronal-model run. An ideal ensemble should provide a good indication of the levels of uncertainty within a forecast. In particular, the spread in ensemble results should be well-correlated with forecast uncertainty. For example, if 60 out of 100 ensemble members are predicting a high-speed solar-wind stream at 15:00 tomorrow then that outcome could be assigned a 60% probability of occurring. And, crucially, when looking at multiple forecasts, events assigned a 60% probability of occurring should be observed 60% of the time.

This is not the case, however, for a naively generated ensemble. For this to be true the ensemble needs to be calibrated.

Ensemble calibration involves adjusting the inputs and/or outputs of an ensemble to better align with the known uncertainties and observed frequencies of forecast events (Wilks, 2011). Calibration improves the reliability and interpretability of ensemble models, ensuring that predicted probabilities reflect real-world likelihoods. Reliability, in this context, describes the ability of a forecast model to accurately predict observed frequencies. Interpretability relates to a models consistency for the end user and how predictions can be used for practical decision-making. Thus, a calibrated ensemble provides a more trustworthy and informative prediction, which is important for accurate risk assessments. The process of ensemble calibration has been well documented by the European Centre for Medium-Range Weather Forecasts (ECMWF) who have developed many calibration techniques for weather and climate ensembles (Gneiting, 2014). While calibration can be achieved through consideration of both the ensemble input and output, in meteorological ensembles calibration is typically achieved through statistical post-processing of the output. that is, it does not change the model state, only the probabilities assigned to it. The application of these methods to the solar-wind forecast will diverge from this idea slightly. The majority of the uncertainty within the solar-wind forecast sits within the inner-boundary conditions. Hence, it makes sense to begin calibration efforts there. This has the additional benefit that the model representation of the ambient solar wind is changed in a physically consistent way, allowing CMEs to be propagated through the resulting ensemble states.

This paper primarily focuses on describing a calibration procedure applied to a HUXt solar-wind ensemble supplied by inner-boundaries from the MAS coronal model. This will be achieved through adjustment of the perturbations to the inner-boundary conditions. This will be shown to improve the distribution of the ensemble output spread (i.e., the range of solar wind speeds predicted in near-Earth space) as well as provide a systematic increase in probabilistic forecast accuracy. Statistical post-processing techniques will also be investigated to tackle existing biases within the HUXt ensemble. Cost/Loss analysis will reveal how the calibrated probabilistic forecast provides a more actionable forecast in high cost/loss regimes where false alarms are most critical.

2. Data and Models

2.1. OMNI Solar Wind Data

To calibrate and verify the ensembles we use solar-wind speed data from the OMNI database. The OMNI database is set of inter-calibrated near-Earth solar-wind observations (King & Papitashvili, 2005). OMNI data are provided through NASA's Space Physics Data Facility (SPDF). We here use 1-hr resolution data.

2.2. HUXt Forecasts

All hindcasts are generated using HUXt driven by the output of the MAS coronal model (Linker et al., 1999; Riley et al., 2012). Forty years of MAS output are available at Carrington-rotation resolution via https://www.predsci.com/mhdweb/home.php. The MAS code solves the set of resistive MHD equations in spherical coordinates. The precise details of the model are well outlined in Riley et al. (2001); Riley and Luhmann (2012); Mikić et al. (2018) and references therein. In brief however, the model is driven by the observed photospheric magnetic field. The archived solution we used were driven by MDI (MAS solutions dated before April 2010) and HMI observations (all remaining solutions used) from the SOHO and SDO spacecraft respectively to construct a boundary condition for the radial magnetic field at 1 $R_{\odot}$. MAS is run from 1 to 30 $R_{\odot}$ to model the large-scale field structure of the solar corona. An important note is that a single map exists per solar rotation, and so, although MAS is time-dependent, it is run until a dynamic steady-state is achieved. This is a reasonable approximation when structure at the Sun is not appreciably changing from one rotation to the next. This approximation becomes less reasonable in the lead up to and during solar maximum. MAS models the coronal magnetic field structure reasonably well, but fails to generate solutions with sufficient variation in solar-wind speeds. So whilst the solar-wind flow is explicitly determined by the MAS model, a better match with observations is obtained in a similar manner to the WSA model: an empirical relations between coronal hole distance and magnetic field line expansion is used to derive an estimate of solar-wind speed at 30 $R_{\odot}$ (Riley et al., 2001, 2015, 2021). We use medium resolution MAS output which produces maps on a $1^{\circ} \times 1^{\circ}$ heliolatitude, Carrington longitude grid. High-resolution MAS runs are also available but the accessible archive is much more limited. To estimate the near-Earth conditions for a single HUXt run, we extract the solar wind speed along the sub-Earth path (the

10.1029/2024SW004164

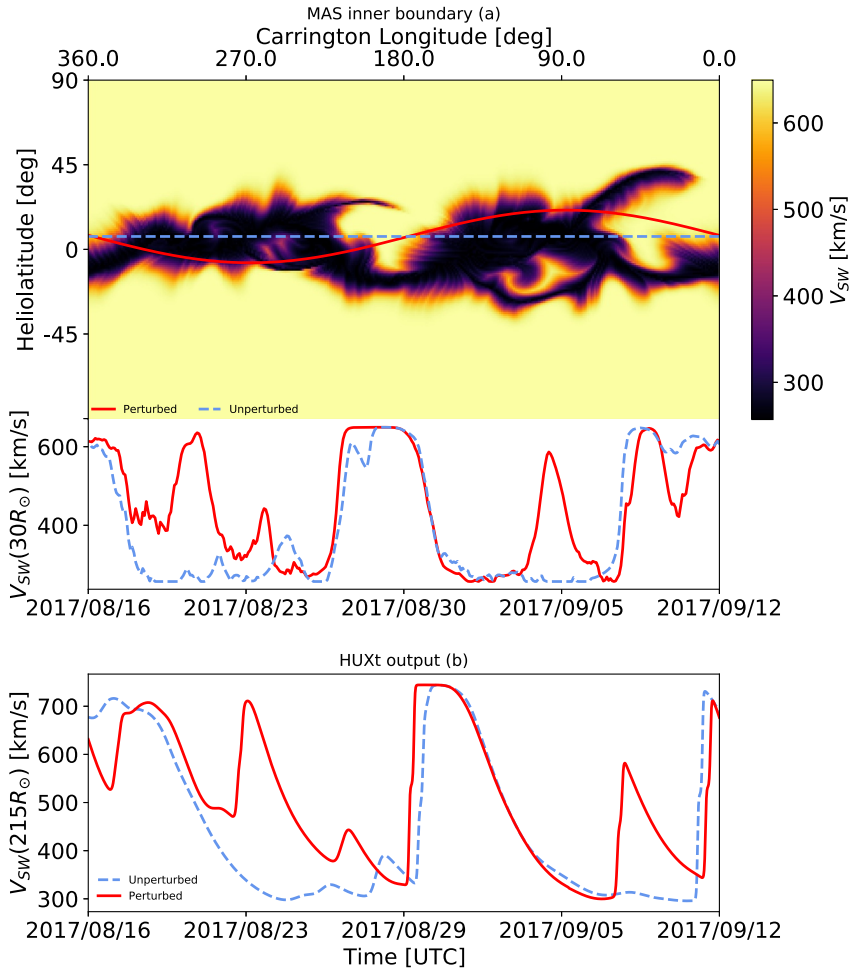

Figure 1. Example of a solar-wind forecast using the MAS and HUXt models (top) The MAS solar-wind speed at 30 $R_\odot$ as a function of Carrington longitude and latitude for Carrington rotation 2,194 2017/08/16–2017/09/12. Carrington longitude axis has been reversed to make the correspondence with the time-axis more intuitive. The sub-Earth path is shown as the blue dashed line. A perturbed Earth position, equivalent to a rotation of the MAS solution, is shown by the solid red line (middle) Solar-wind speed at 30 $R_\odot$ as a function of Carrington longitude along the sub-Earth path and the perturbed location, in the same format. This forms the input to the HUXt model (bottom) The HUXt forecast at 215 $R_\odot$ (at Earth) obtained from the two inner boundaries in the top panel.

projection of ecliptic plane onto the 30 $R_\odot$ solar-wind source surface) to use as the HUXt inner-boundary conditions. HUXt then models the bulk flow out to Earth through solving Burgers equation using an upwind numerical scheme (Barnard & Owens, 2022; Owens et al., 2020). This results in a time series forecast of wind speed in near-Earth space. This process is demonstrated in Figure 1. Whilst we will use the term 'forecast' throughout this paper, all analysis was technically performed on 'hindcasts' to allow use of large verification data sets.

2.3. Generating Ensemble Forecasts

HUXt ensembles are generated through spatially perturbing the inner-boundary. With this framework, we adapt and expand on the methods outlined in Owens and Riley (2017). To generate an ensemble of inner-boundary conditions, the MAS solution at 30 $R_\odot$ is spatially perturbed, in the manner detailed below. Each set of inner-boundary conditions is then individually propagated out to Earth using HUXt and this forms the ensemble output. The process for producing the perturbed inner-boundary conditions is outlined in Figure 1. Using the speed estimates along the sub-earth path as a basis for the 'true' inner-boundary conditions, this path is spatially (sinusoidally) perturbed through this equation:

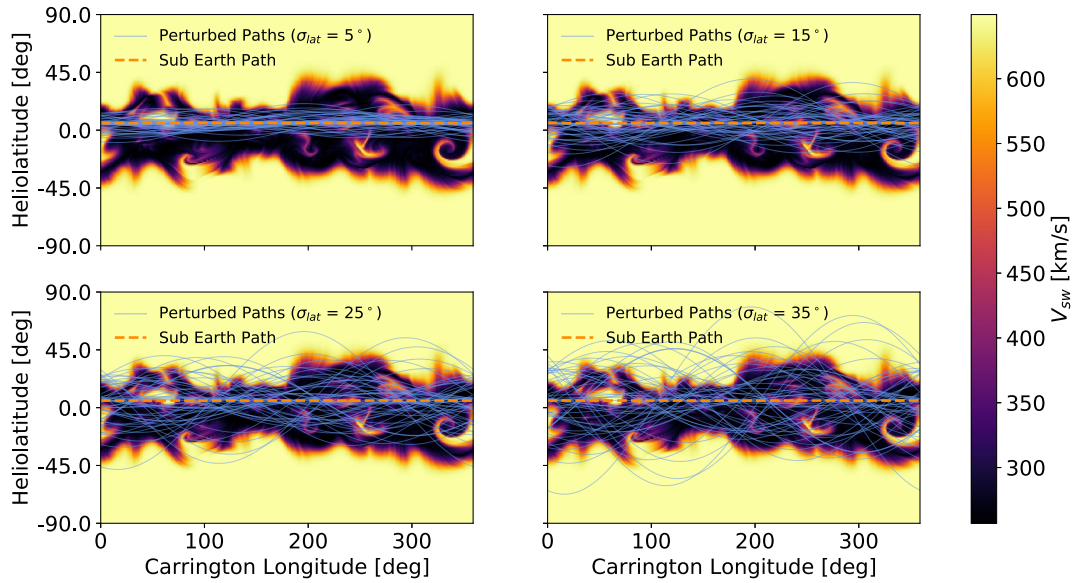

Figure 2. The effect of σ_{lat} on the sampling of the MAS model and hence the inner-boundary conditions used by HUXt. The color map shows solar-wind speed as a function of Carrington longitude and heliolatitude for a MAS solution at 30 $R_{\odot}$ during solar minimum (from observations during Carrington rotation 2,239, 25/12/2020–01/22/2021). Panels show increasing σ_{lat}, which makes larger perturbations away from then sub-Earth path more likely. To preserve continuity across longitude, the wave number is kept fixed at 1.

$$\theta(\phi) = \theta_E + \theta_{MAX} \sin\left(n\phi + \phi_0\right), \tag{1}$$

where θ is heliolatitude of the perturbed path, ϕ is Carrington longitude, θ_E is the unperturbed heliolatitude of Earth, θ_{MAX} is the maximum perturbation, n is the wave number, and ϕ_0 defines the phase. For $n = 1$, which is used throughout this study, this sinusoidal perturbation is equivalent to exacting the sub-Earth point after a latitudinal rotation of the MAS map. The phase, ϕ_0, is equivalent to the longitude of the ascending node. As no longitude should be more likely than any other, the value of ϕ_0 for each ensemble member is randomly drawn from a uniform distribution between 0 and 2π. The value of θ_{MAX}, on the other hand, is expected to be distributed about 0. Thus it is randomly generated from a Gaussian distribution with a width defined by scale parameter $\sigma_{latitude}$, which governs the magnitude of the perturbations:

$$\theta_{MAX} = \frac{1}{\sqrt{2\pi\sigma_{lat}^2}} e^{-\frac{\mu^2}{2\sigma_{lat}^2}} \tag{2}$$

Figure 2 visualizes sets of perturbed inner-boundaries with increasing σ_{lat}. For this solar minimum period, the increased heliolatitude sampling resulting from increased σ_{lat} will generally lead to faster solar wind at the inner boundary. The sets of inner-boundaries are then given to HUXt to propagate out to 1 AU, Figure 3a shows an example of the resulting ensemble. It should be noted that CME perturbations were not added to the inner boundary for HUXt within this study. Hence, the labeled CME shocks were not captured by the ensemble.

The inner-boundary conditions can also be perturbed in longitude. The ambient solar wind is treated as being "steady-state", so a longitudinal perturbation—a rotation of the inner-boundary with respect to the target observer—is analogous to a time shift on the forecast output (Note that this would not be the case with time-dependent inner-boundary conditions, such as when CMEs are added at the inner boundary.) The result of this longitudinal perturbation is visualized in Figure 3b, where it introduces a more pronounced temporal spread. The magnitude of the longitudinal perturbation is drawn from a Gaussian distribution much like Equation 2, using a second scale parameter, σ_{lon}. Introducing the longitudinal perturbation is beneficial because often a high-speed stream can be correctly characterized by HUXt but not arrive at Earth at the correct time (as verified by OMNI). The longitudinal perturbation allows this timing uncertainty to be better captured by the ensemble and

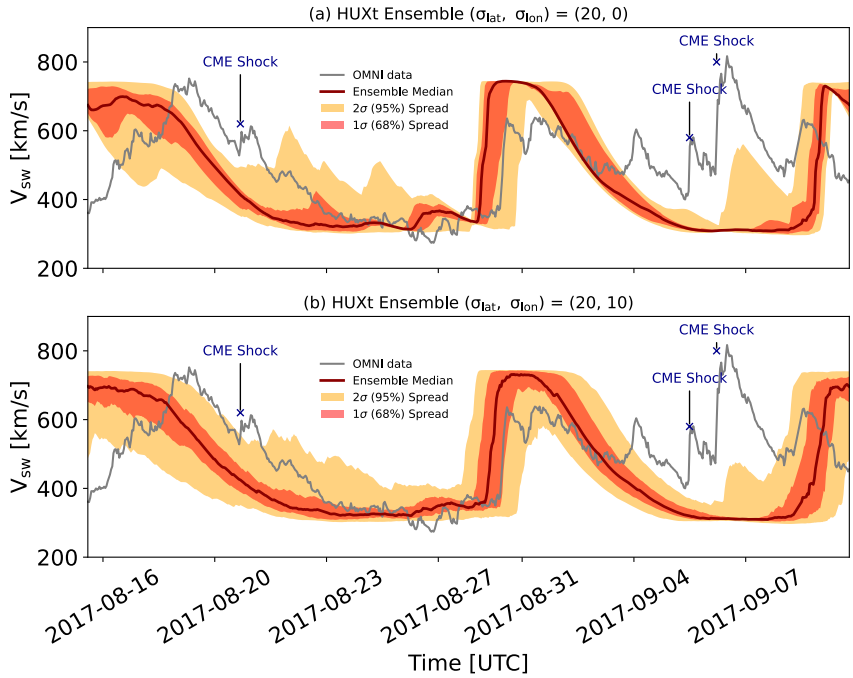

Figure 3. An example of the impact of the perturbation parameters on the resulting 100-member ensemble forecast time series at Earth. Time series are shown for forecast produced over Carrington Rotation 2,247 (from 16/08/2017 and 07/09/2017). The observed solar-wind speed from the OMNI data set is shown in gray. The times of three CME-driven shocks are noted with a cross and labeled. The solid red line shows ensemble median of the MAS/HUXt forecast, with 1σ and 2σ spreads as orange and yellow shaded regions, respectively. The top panel (a) shows $\sigma_{lat} = 20°$ and $\sigma_{lon} = 0°$ (i.e., no longitudinal perturbation). The bottom panel (b) shows $\sigma_{lat} = 20°$ and $\sigma_{lon} = 10°$.

lead to a more representative forecast. The implementation of the longitudinal perturbation was not considered by Owens and Riley (2017). With this addition, latitude and longitude are treated identically in terms of spatial perturbations to the MAS solution.

The sinusoidal perturbation scheme is mathematically equivalent to a rotation of the MAS model output itself. Generating a perturbed-boundary ensemble this way is based on an assumption that the primary source of uncertainty within the coronal model comes from a rotational/positional error. The validity of this assumption is difficult to directly test and future studies will incorporate additional aspects of the inner-boundary uncertainty. Other studies have worked to quantify the spatial and temporal uncertainty within solar-wind models, primarily from the synoptic maps used as input to coronal models (Bertello et al., 2014; Kennis et al., 2024; Perri et al., 2023). But this approach nevertheless provides a pragmatic way to efficiently generate large ensembles for uncertainty quantification. The size of these perturbations should ideally capture the size of the uncertainty within the initial conditions. However, the magnitude of this uncertainty is poorly constrained a priori. So we examine the impact of the two scale parameters, σ_{lat} and σ_{lon}, on the ensemble using rank histograms, calibration curves, paired with forecast evaluation to find the σ_{lat} and σ_{lon} values which produced the most calibrated ensemble.

3. Methods

3.1. Rank Histogram Analysis

Rank histograms are a method of inspecting the ensemble distribution with respect to the truth that is, a verification data set of observations. For each forecast time, the rank of the observation within the ensemble is determined. For example, if seven ensemble members are overpredicting the wind speed from a set of 10 members, this ensemble instance is assigned a rank of 7. The rank histogram displays the occurrence of ranks over a large number of forecasts (Hamill & Colucci, 1998). Figures 4a and 4e show how a perfectly calibrated ensemble should produce a completely uniform (or flat) rank histogram, since there should be no bias toward over/under predicting between any given ensemble member, which means that observations will consistently fall

10.1029/2024SW004164

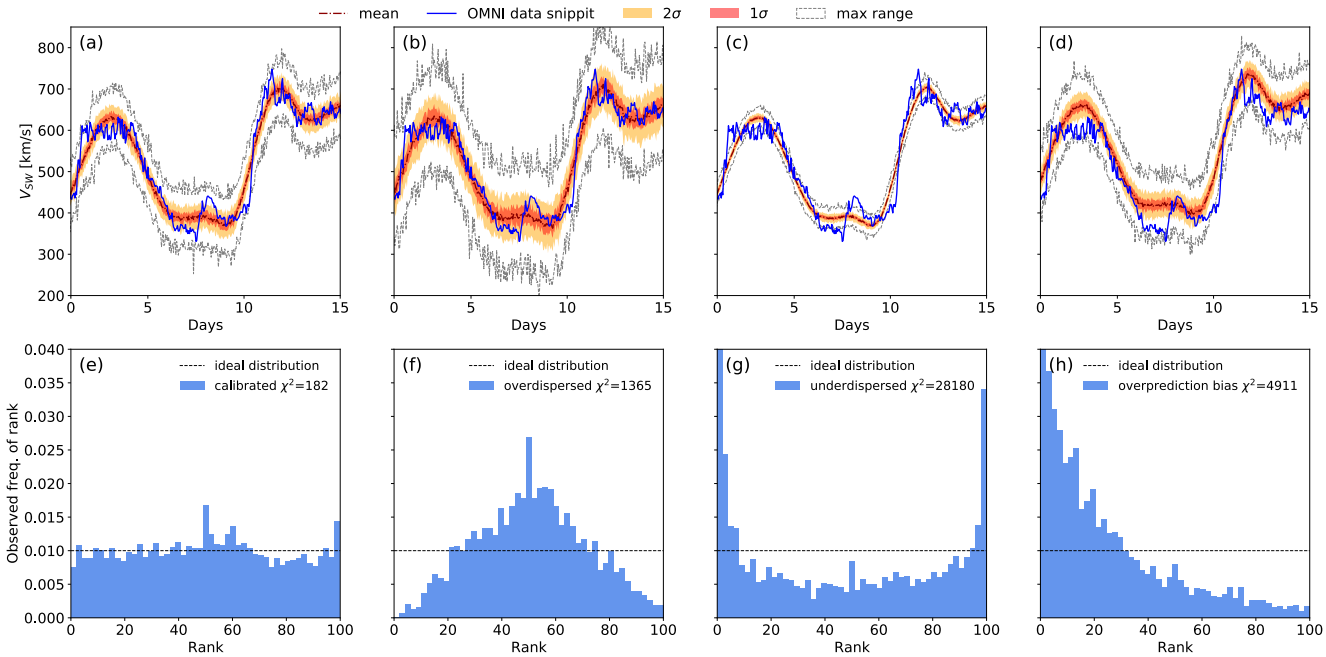

Figure 4. Examples of ensemble spreads and rank histograms generated from synthetic 100-member ensembles derived from a 6 months snippet of OMNI data which demonstrate various ensemble behaviors across panels a–h. (a) 15 days portion of a calibrated ensemble with corresponding flat rank histogram (e) showing little bias toward any ranks, (b) overdispersed ensemble with resulting (f) bowl shaped histogram, (c) an under dispersed ensemble with (g) U shaped histogram, and finally (d) an overbiased ensemble with (h) highly skewed rank histogram toward lower ranks.

within the ensemble but equally at all ranks (given a large enough sample population). Any deviation from a uniform distribution can be used to infer biases within the ensemble (Hamill, 2001). An overspread ensemble will be biased toward middling ranks (Figures 4b and 4f) whereas an underdispersed ensemble will produce a U-shape distribution because ranks are biased high/low as the limited spread fails to capture observed variability (Figures 4c and 4g). Figures 4d and 4h highlight what would happen when the ensemble members are consistently biased high. The nature of the perturbation scheme means that increased perturbation size means sampling the higher latitude solar wind more. At solar minimum, this directly translates to faster solar wind. That means more of the ensemble members overpredict solar wind speed at Earth. This would produce this trend toward low ranks.

In order to quantify how close a given rank histogram is to the ideal (flat) case, we use the Wilks (2019c) augmented form of the typical χ^2 test:

$$\chi^2 = n(m+1)\sum_{k=1}^{m+1}\left(\frac{n_k}{n} - \frac{1}{m+1}\right)^2, \tag{3}$$

where n is the total number of ensemble members, n_k is the total number of ensemble members with rank k, m is the rank index such that $k = 1, 2, \ldots, (m+1)$. This is constructed similarly to a traditional χ^2 metric such that n_k/n term represents the observed rank population and $1/(m+1)$ represents the expected (or rather ideal) population. This means it behaves in a similar way to the traditional metric which compares observed frequencies to expected frequencies. The main difference is that expected frequencies are always uniform (i.e., each bin is expected to have the same count if the ranks are uniformly distributed). In traditional χ^2 tests, expected frequencies can vary depending on the null hypothesis being tested. A smaller χ^2 indicates a flatter distribution, with $\chi^2 = 0$ indicating a perfectly flat histogram. Examples of χ^2 can also be seen in Figure 4. Parameterizing the histogram distribution in terms of χ^2 is useful for testing large numbers of ensemble sets generated using different perturbations. An example of the rank histogram evaluated on a portion of the HUXt ensemble can be seen in Figure 5b.

10.1029/2024SW004164

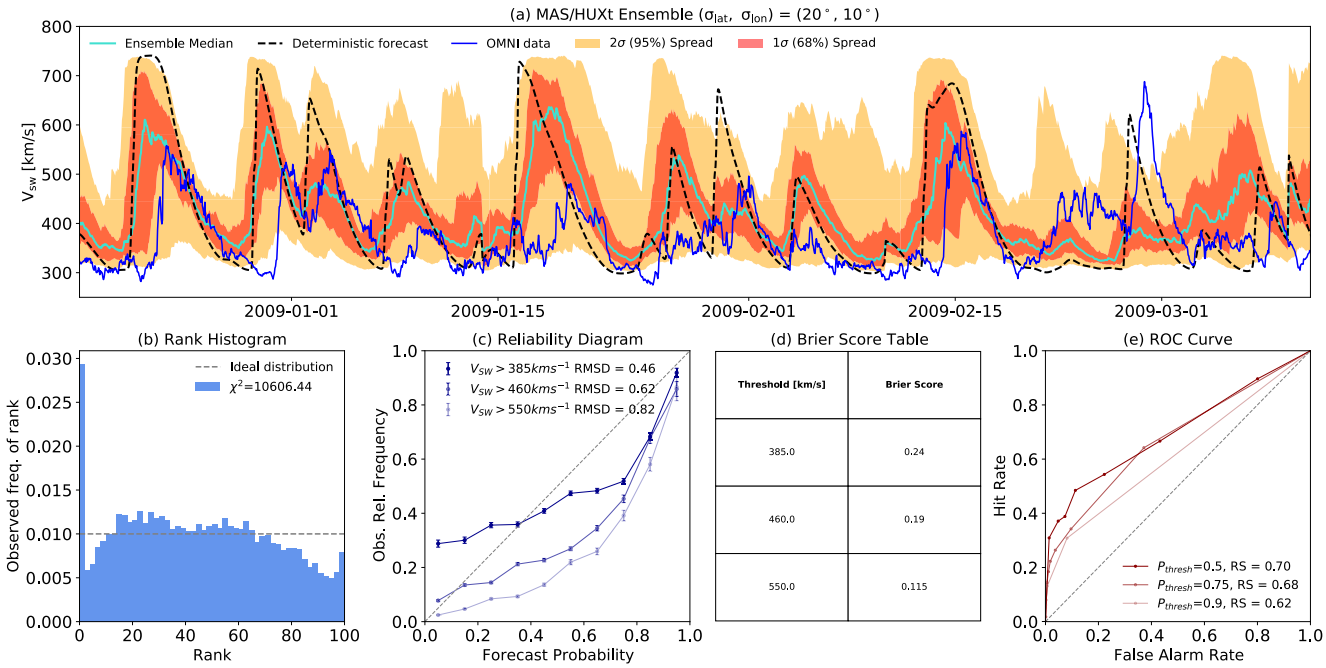

Figure 5. A demonstration of how each ensemble was evaluated (*Top*/(a)) A 3-month portion of OMNI data (blue), deterministic forecast (black dashed line) and a portion of the solar-minimum HUXt ensemble spread (oranges) with ensemble median (turquoise) generated using ensemble perturbation scale parameters $(\sigma_{lat}, \sigma_{lon}) = (20°, 10°)$ (bottom) From left to right: a rank histogram with χ^2 (b); reliability diagrams at a range of wind speed event thresholds (c); Brier scores calculated at a range of solar-wind speed thresholds (d); and ROC Curves at a series of probability thresholds (0.5, 0.75, 0.9) (e). The event thresholds were set as the median (50^{th}), 75^{th}, and 90^{th} percentiles, equating to values of 385, 460, and 550 km/s respectively. Evaluations completed over solar-minimum data set.

3.2. Reliability Diagrams

The reliability diagram (also called a calibration curve) is a tool used in meteorological forecasting to examine how well calibrated a model/forecast is. The diagram is used to infer biases toward over/under confident predictions and can also distinguish between forecasts which are well calibrated but differ in their decisiveness; a model's ability to assign more extreme probabilities to events with high confidence (Dawid, 1982; DeGroot & Fienberg, 1983). Reliability diagrams plot the forecast probability against the observed relative frequency. A perfectly calibrated forecast would produce a line along the center diagonal, where the forecast probability perfectly matches the frequency of times the forecast event was observed. Example diagrams derived from a portion of the HUXt ensemble output can be seen in Figure 5c. The use of this metric alongside the rank histograms can provide a clear picture of the HUXt ensemble output and how well calibrated (or otherwise) the forecast is. We evaluate the calibration curve using the root mean square deviation (RMSD) from the perfect $(y = x)$ calibration line. A lower RMSD indicates a better calibration curve and hence, more calibrated ensemble. However, the calibration curve involves binning the probabilistic forecast into discrete probability bins. As such, each bin population is subject to a sampling error, especially for rarer events. To improve the interpretability of the reliability diagrams, error bars were approximated using a Poisson sampling error such that:

$$f_{err,i} = \frac{f_i}{\sqrt{N_i}} \tag{4}$$

where f_{err} is the error in observed frequency, f is the observed frequency, and N is the number of samples within i^{th} probability bin.

3.3. Forecast Evaluation

Two metrics were used to judge the effectiveness of the calibration in terms of improving the forecast. The Brier score [BS] is a performance metric used for probabilistic forecasts. After defining an event threshold, $V_{threshold}$,

the BS can be evaluated by comparing forecast probability of exceeding $V_{threshold}$ with observations. The probabilities are generated by the fraction of ensemble members above $V_{threshold}$. Thus, BS is equivalent to the mean square error of the forecast and is a useful single value indicator of probabilistic accuracy (Brier, 1950). BS is evaluated as follows:

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(f_t - o_t\right)^2 \tag{5}$$

where BS is the Brier Score as before, N is the number of forecast instances t, f_t is the forecast probability, o_t labels the actual outcome of a forecast event (0 for non-event or 1 for an event). The score takes a value between 0 and 1, where 0 would indicate a perfect forecast. BS is primarily an indication of forecast reliability (Wilks, 2019b). As stated earlier, reliability is a forecaster/model's ability to predict accurate probabilities of events. Hence, it is closely linked to calibration.

The second metric used was the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the true-positive rate versus the false-positive rate, at a variety of solar-wind speed event thresholds (Mason, 1982). The shape of the ROC curve reveals how the ensemble performs compared to climatology, which produces a straight line from the origin to 1. The ROC score is the evaluated area under the ROC curve. This produces a score which indicates a better forecast as it approaches 1 (Wilks, 2019b). The ROC curve is useful for assessing how well a model can distinguish between classes—in this case events and non-events—independent of the distribution of forecast probabilities. It gives insight into how sensitivity (the true positive rate) and specificity (false positive rate) change at different event thresholds. Overall, the ROC curve describes the forecast resolution; the ability to distinguish events and non-events beyond the climatological probabilities (Wilks, 2001).

In summary, the ROC curve provides information about the model's ability to discriminate between event and non-event classes, while the BS measures the model's predictive accuracy within those classes (Wilks, 2001). Essentially, the ROC curve measures the ordering of classes; whether a forecast is accurate when it says there is more likely to be an event than not. Whereas the BS measures whether the forecast probability of an event is accurate. Hence, the ROC curve and BS used in conjunction provide a comprehensive picture of the forecast performance. Figures 5d and 5e provide examples of the evaluation metrics as applied to a portion the HUXt ensemble output.

4. Results

4.1. Calibration Metrics

Having established the perturbation scheme, a series of 100-member ensembles were run across Solar Cycle 24. The 100-member ensembles represent a large number of samples for the statistical analysis, additionally there is no set consensus on what defines a large N-member ensemble. Mays et al. (2015) use 36 ensemble members for WSA-Enlil + Cone forecasts. Furthermore, in weather and climate contexts, typically ensembles of size ≥30 are deemed large (Leutbecher, 2019; Milinski et al., 2020). This series of forecasts were split into subsets at solar minimum, maximum, and transitional periods based on the solar activity index (SAI)—a normalized count of sunspot number which represents solar activity (e.g., Owens et al., 2022). Solar minimum contained model solutions during periods of low solar activity when $SAI < 1/3$ (2008-Dec to 2010-Sep and 2016-Aug to 2019-Nov), solar maximum contained model solutions at times when activity was high when $SAI > 2/3$ (2011-Jun to 2014-Mar). The solutions from the periods approaching/leaving solar maximum (when $1/3 \leq SAI \leq 2/3$) were contained within the intermediate transitionary subset. The scale parameters $\sigma_{\mathrm{lat}}, \sigma_{\mathrm{lon}}$ were tested from 0° to 40° in increasing increments of 5°, resulting in 9 × 9 ensembles to cover all permutations (hence, 81 ensembles were evaluated). Using an ensemble from the solar minimum subset, Figure 5 demonstrates how each ensemble was evaluated using: a rank histogram with corresponding χ^2 value (b), reliability diagram (c), ROC curve with area under curve ROC score (d), and Brier scores (e). The processed results were visualized as a series of color maps of the relevant metric to identify the best-calibrated ensembles and find trends in forecast performance relating to the calibration.

To further emphasize the practicalities of using HUXt for generating the ensembles as described: the data set of forecasts consists of approximately 325K simulation days, requiring ∼$1e3$ core hours to generate using HUXt. To

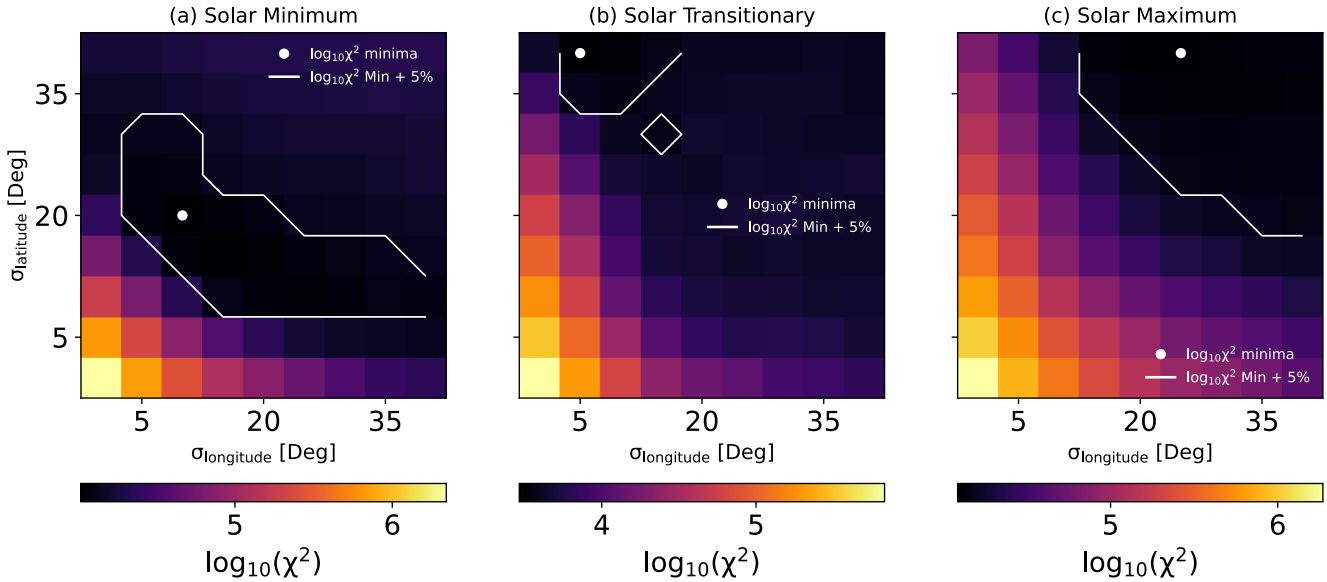

Figure 6. Summary of ensemble calibration with different levels of boundary-condition perturbation. The color map shows $\log_{10}\chi^2$ of rank histograms compared with the ideal case for a range of values of σ_{lat} and σ_{lon}. Panels, from left to right, show solar minimum (a), transitionary (b) and solar maximum (c) periods. The minimum χ^2 (i.e., best-calibrated ensemble) has been marked with a white dot, the white contours bound the minimum +5% of this value.

generate the same data set using a full 3D MHD model such as EUHFORIA or Enlil would have required upwards of $\sim 1e5$ core hours. That would be impractical for the parameter space exploration done for this study. Though the results garnered through the use of HUXt should be directly applicable to 3-D MHD models.

Figure 6 shows the distribution of χ^2 values for the rank histograms across the perturbation parameter space. The lower the χ^2 value, the more-evenly distributed—and hence better calibrated—the ensemble is with respect to the observed solar-wind speed. The minimum χ^2 values and the minimum $\chi^2 + 5\%$ contours are highlighted. This was done to identify the approximate ranges of σ_{lat} and σ_{lon} which produced the best-distributed ensembles. At solar minimum, there is a bounded minimum in the perturbation scale parameters. During intermediate and solar maximum conditions, this is not the case. This is discussed further in Section 5.

Figure 7a shows the reliability diagrams for the best ensemble at solar minimum $(\sigma_{lat}, \sigma_{lon}) = (20\,^\circ, 10^\circ)$, for a range of speed thresholds. These speed thresholds are the median $\left(50^{th}\right)$, 75^{th}, and 90^{th} percentiles, equating to values of 385, 460 and, 550 km/s respectively. The remaining three panels show three other combinations of

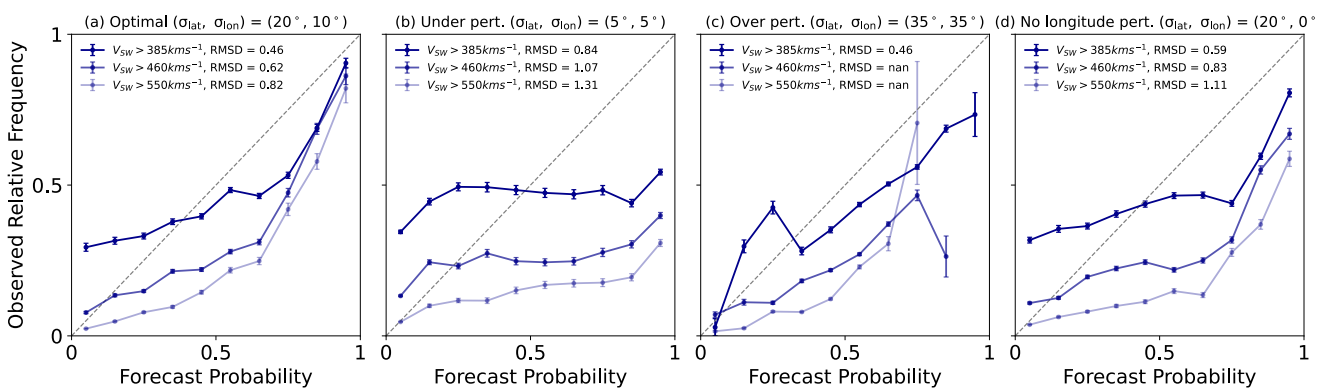

Figure 7. Reliability diagrams for ensembles generated using different combinations of perturbation scale parameters, σ_{lat} and σ_{lon}, over the solar-minimum data set. (a) An ensemble generated using the best parameters $(\sigma_{lat}, \sigma_{lon}) = (20\,^\circ, 10^\circ)$ from minimum χ^2 in Figure 6a. (b) An under-perturbed ensemble $(5\,^\circ, 5^\circ)$, (c) an over-perturbed ensemble $(35\,^\circ, 35^\circ)$, and (d) an ensemble with no longitudinal perturbation $(20\,^\circ, 0^\circ)$. RMSD indicates distance from optimal calibration line $(y = x)$. Diagrams evaluated at the median, 75^{th} and 90^{th} percentile solar-wind speed event thresholds of 385, 460, and 550 km/s respectively (dark to light blue). The plotted error bars are the approximate sampling errors calculated by assuming Poisson counting statistics holds for each forecast probability bin.

perturbation scale parameters, for comparison. These are, from left to right, an under-perturbed ensemble $(\sigma_{lat}, \sigma_{lon}) = (5°, 5°)$ (b), an over-perturbed ensemble $(\sigma_{lat}, \sigma_{lon}) = (35°, 35°)$ (c), and an ensemble with no longitudinal perturbation $(\sigma_{lat}, \sigma_{lon}) = (20°, 0°)$ (d), as per Owens and Riley (2017). From the reliability diagrams the most calibrated curves are closer to the $y = x$ line. However, reliability diagrams for all four ensembles show trends toward overconfident forecasts at high forecast probabilities.

4.2. Forecast Improvement

For reasons outlined in the discussion, the remaining analysis was completed on only the solar minimum data set. As expected, the reliability diagrams in Figure 7a also suggest that a more calibrated ensemble was achieved by using the scale parameters which result in the most-optimal rank histograms (best parameters were $(\sigma_{lat}, \sigma_{lon}) = (20°, 10°)$). However, the calibration is only useful if it can produce a clear improvement in the forecast uncertainty quantification. As demonstrated in Figures 5d and 5e, this was evaluated using ROC score and BS. Figure 8 displays the BS and ROC scores for the complete set of ensembles over the solar minimum subset. The minima and contours from the χ^2 grids in Figure 6a were added in Figure 8 to enable comparison of ensemble calibration and improvements in the forecast. Considerable agreement is seen between the regions of parameter space that improve calibration and the measures of forecast improvement.

To further quantify this relationship, we split the data set into two populations of ensembles: a 'calibrated' set, being those 21 sets within the minimum +5% contour of rank histograms in Figure 8, and an 'uncalibrated set', being those 60 sets outside. Figure 9 compares the distributions of Brier and ROC scores for the calibrated and uncalibrated ensembles using a histogram and cumulative distribution function (CDF). The scores from the calibrated-ensemble population clearly show a consistent shift toward lower BS (improved reliability) than the uncalibrated ensembles (Figure 9a). Figure 9b shows the CDFs generated from both populations. The shift toward lower BS indicates a systematic improvement in the forecast due to the calibration. Figures 9c and 9d show the same population breakdown but evaluated using the ROC score. Again, this shows shift toward higher scores among the calibrated population of ensembles. There was an average shift of 0.025 (9%) toward lower BS. An average increase of 0.01 (2.5%) was found in the ROC scores. Average improvement in the metrics used (particularly for ROC score) is relatively small, but changes in score distribution indicate the shift is systematic. The improvement is most clearly seen in CDFs from Figures 9b and 9d.

4.3. Statistical Post-Processing Using Platt Scaling

So far only the solar-wind ensemble input has been considered as a means to perform calibration. Meteorological models are typically very computationally expensive and so hard limits exist on ensemble size. As a result, calibration is typically achieved through statistical post-processing of outputs. The majority of the uncertainty within the solar-wind ensemble comes from the inner-boundary conditions which justifies the focus on the input spread. However, we show here that the forecast can be further improved through employing some post-processing techniques.

Platt-scaling (Platt, 2000) is a simple method of post-processing in which the outputs of a classification model are transformed into calibrated probabilities using logistic regression. Logistic regression models/predicts the probability of a binary outcome based on predictor variables (the forecast probabilities from the ensemble, in this application). At present, the ensemble forecast is used categorically, predicting probabilities for two classes of events defined as being speeds above/below $V_{threshold}$. Thus, it can be treated as a classification model and hence calibrated through this method. Using a gradient-descent algorithm, weights can be trained to account for biases within the forecast probabilities. The raw probabilities, $\mathbf{p}$, are scaled using the following function:

$$\tilde{\mathbf{p}} = \mathbf{f}(\theta, \mathbf{p}) = \frac{1}{1 + e^{-(\theta \mathbf{p} + b)}} \tag{6}$$

where $\tilde{\mathbf{p}}$ are the resulting Platt-scaled probabilities, θ are the weights which are optimized through gradient descent, b is a fitting constant, and $\mathbf{f}(\theta, \mathbf{p})$ is the activation function used to transform the raw probabilities using said weights. Logistic regression means in this case $\mathbf{f}(\theta, \mathbf{p})$ takes the form of a sigmoid function. The Platt-scaled probabilities produced the calibration curves in Figure 10 which show a consistent shift toward a better calibration, particularly at higher event thresholds.

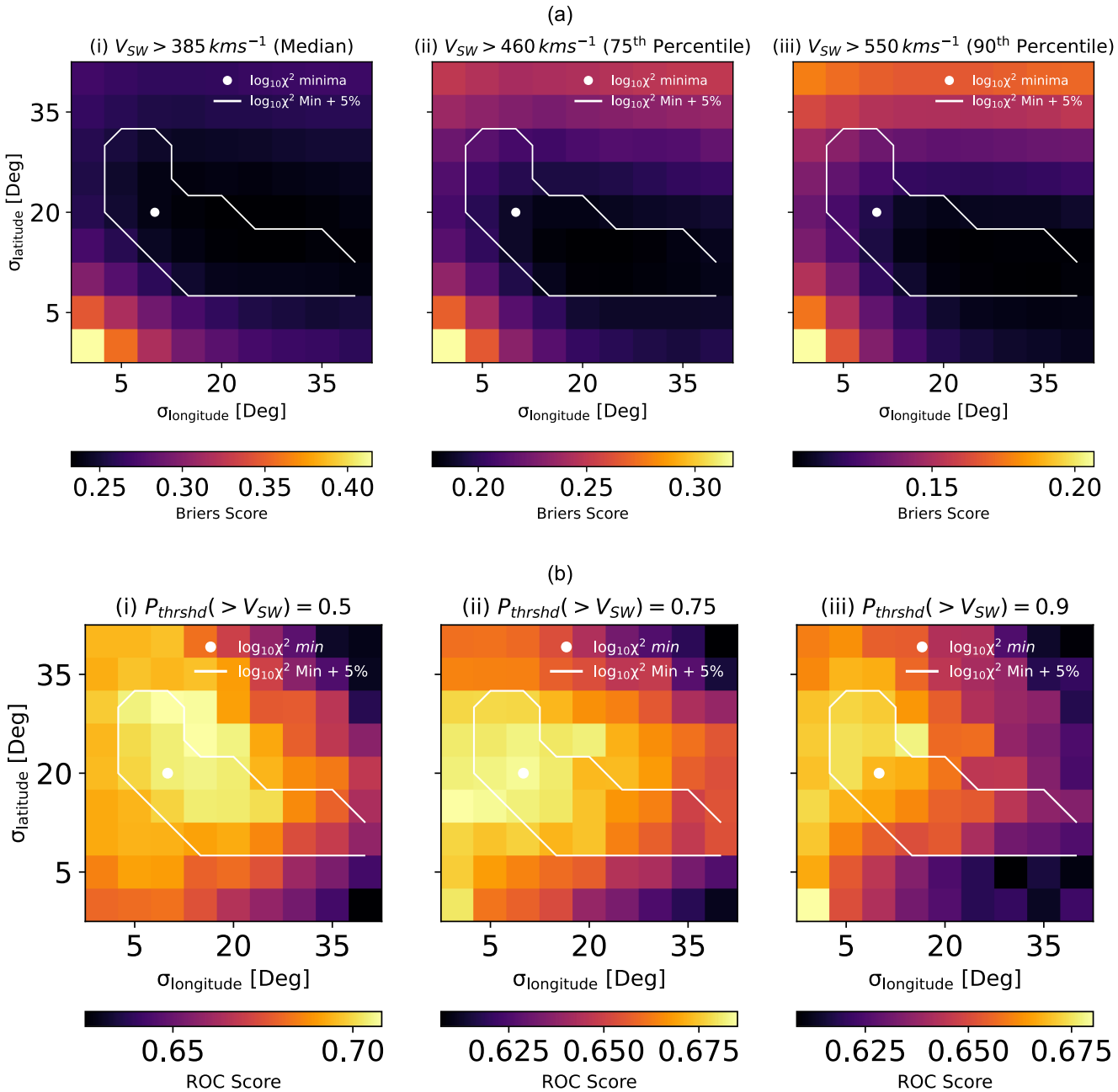

Figure 8. Color maps of forecast metrics (a) Brier score and (b) ROC score as a function of perturbation scale parameters $(\sigma_{lat}, \sigma_{lon})$ for the solar-minimum data set. Brier score in (a) was calculated for events defined by solar-wind speed thresholds V_{thresh} at the median, 75^{th}, and 90^{th} percentiles of the observed solar-wind speed during solar minimum period (labeled left to right). A lower Brier score indicates an improved forecast. ROC score was evaluated using increasing probability threshold 0.5, 0.75, and 0.9 defining when the forecast had identified an event as per V_{thresh}. A higher ROC score indicates an improved forecast. Overlaid on the both color-map sets as a white point is the best-calibrated ensemble, as determined by χ^2, with the minimum +5% contour of this value as a white line.

4.4. Cost/Loss Analysis

As seen within the ROC score results (Figures 8b, 9c, and 9d) the ensemble skill is very sensitive to the probability threshold used to define an actionable forecast. Cost/Loss analysis is a technique developed by Murphy (1985) which can examine the impact of this threshold on the value of a forecast. The analysis assumes that for every forecast event, an action is taken to prevent incurring some kind of loss, L, such as damage to the power grid or loss of a communications satellite. But taking action itself incurs a smaller cost, C. The ratio of C to L is analogous to the probability threshold at which action should be taken. When C/L is small, action should be taken even at

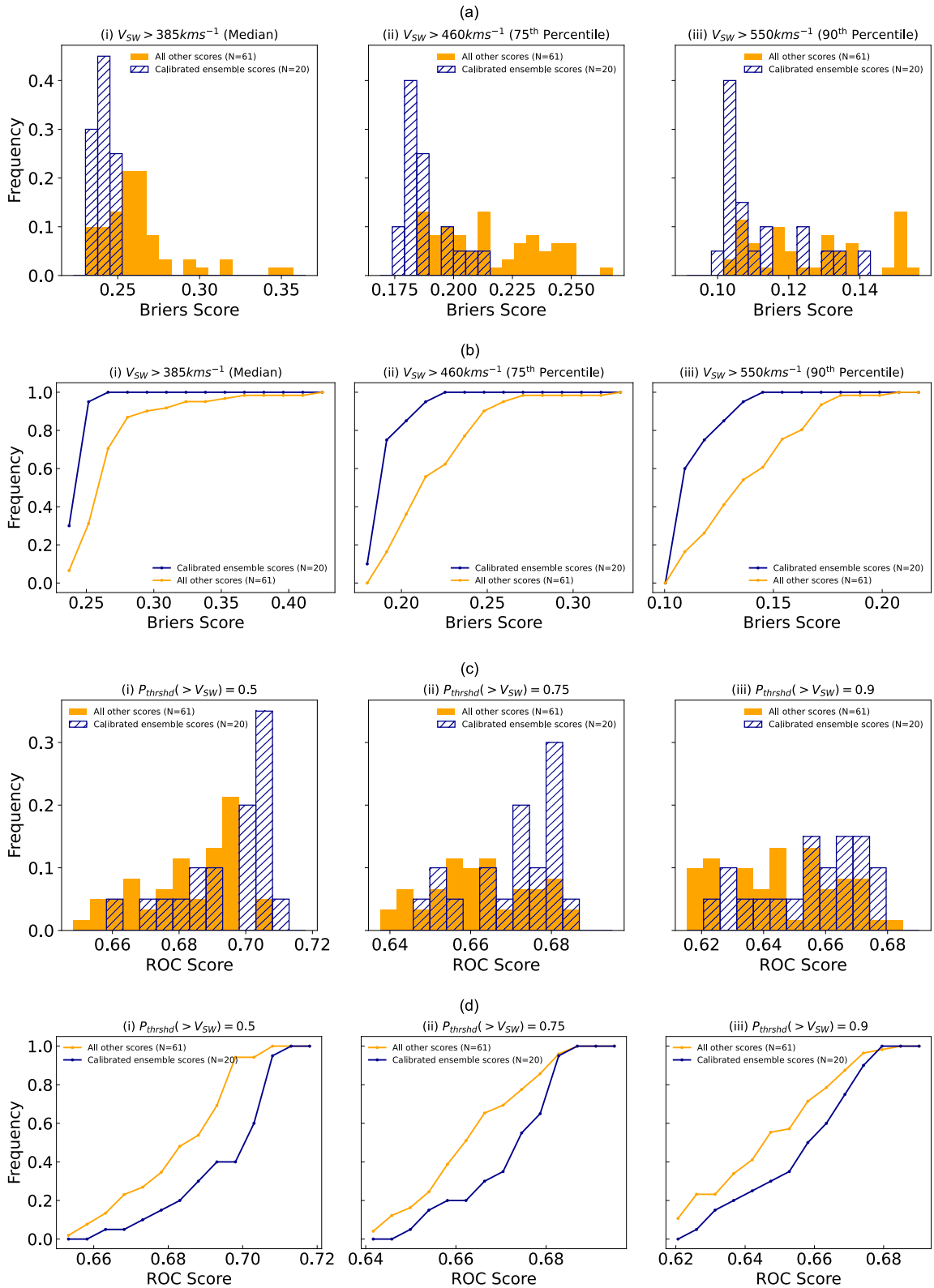

Figure 9. Comparisons of calibrated (blue hashed shading) and uncalibrated (orange) ensembles over the solar-minimum interval. Panels compare Brier score (a), (b) and ROC score (c), (d) histogram and CDF distributions. Brier score is evaluated at increasing event thresholds (median, 75^{th}, 90^{th} percentile wind speed thresholds). ROC score is evaluated at increasing probability thresholds (0.5, 0.75, 0.9).

10.1029/2024SW004164

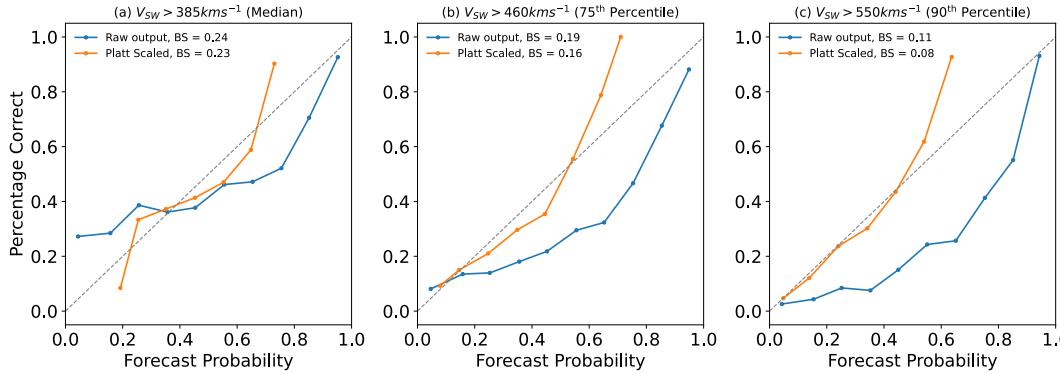

Figure 10. Comparisons of the reliability diagrams derived from the raw ensemble output (blue) versus the Platt-scaled probabilities (orange). Processing and evaluation was done over solar-minimum data set using ensemble generated with most-calibrated parameters $(\sigma_{lat}, \sigma_{long}) = (20°, 10°)$. Evaluated at the median, 75^{th}, and 90^{th} wind speed percentile action thresholds (columns left to right). The Brier score has also been evaluated and listed.

low probabilities, as the cost of false alarms is relatively low. The converse is true at high C/L. A more detailed discussion in the context of solar wind forecasting is presented in Owens and Riley (2017). The effect of the cost/loss (C/L) ratio can be examined by looking at Potential Economic Value (PEV), a percentage metric comparing forecast performance against a theoretical perfect forecast which incurs no losses or false positive costs.

PEV was evaluated as a function of C/L for the probabilistic forecast and Platt-scaled probabilistic forecast, as well as a deterministic HUXt forecast (i.e., using the unperturbed MAS output) for comparison. Unlike the deterministic, the probabilistic forecast maintains some value at very high and very low C/L ratios. In the high C/L regime, PEV is very sensitive to false positives, so it is promising that the calibrated probabilistic forecast can maintain some level of PEV at high C/L. However, the impact of the bias toward over-confident forecasts can be very clearly seen at the high C/L range, where the PEV drops due to false positives becoming the dominant factor. Value begins to return as the bias decreases, as seen most clearly in Figures 11b and 11c. The Platt-scaled forecasts do noticeably better in the high C/L regime where PEV is very sensitive to false positives due to the reduced over-confidence bias.

5. Discussion

5.1. Calibration Trends

The χ^2 color map in Figure 6a shows a reasonably defined locus of well-distributed ensembles at solar minimum. That is, the minimum appears bounded, as expected if optimum calibration is achieved. The best parameters at solar maximum and intermediate activity levels (Figures 6b and 6c) are unbounded, in that unrealistically large

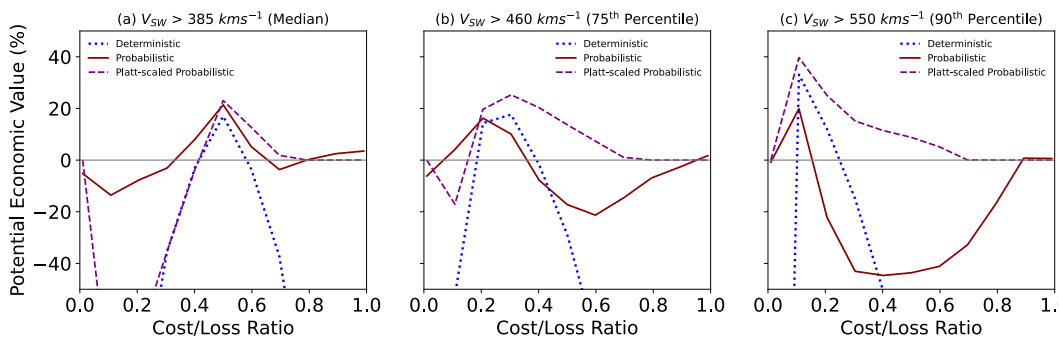

Figure 11. Cost/Loss analysis comparing the Potential Economic Value (PEV) against cost/loss ratio of the raw (red) and Platt-scaled (purple dashed) probabilistic forecasts with a deterministic forecast (blue dashed). PEV was evaluated over the solar-minimum data set using ensemble generated with most-calibrated parameters $(\sigma_{latitude}, \sigma_{longitude}) = (20°, 10°)$. Evaluated at the median, 75^{th}, and 90^{th} wind speed percentile action thresholds (left to right).

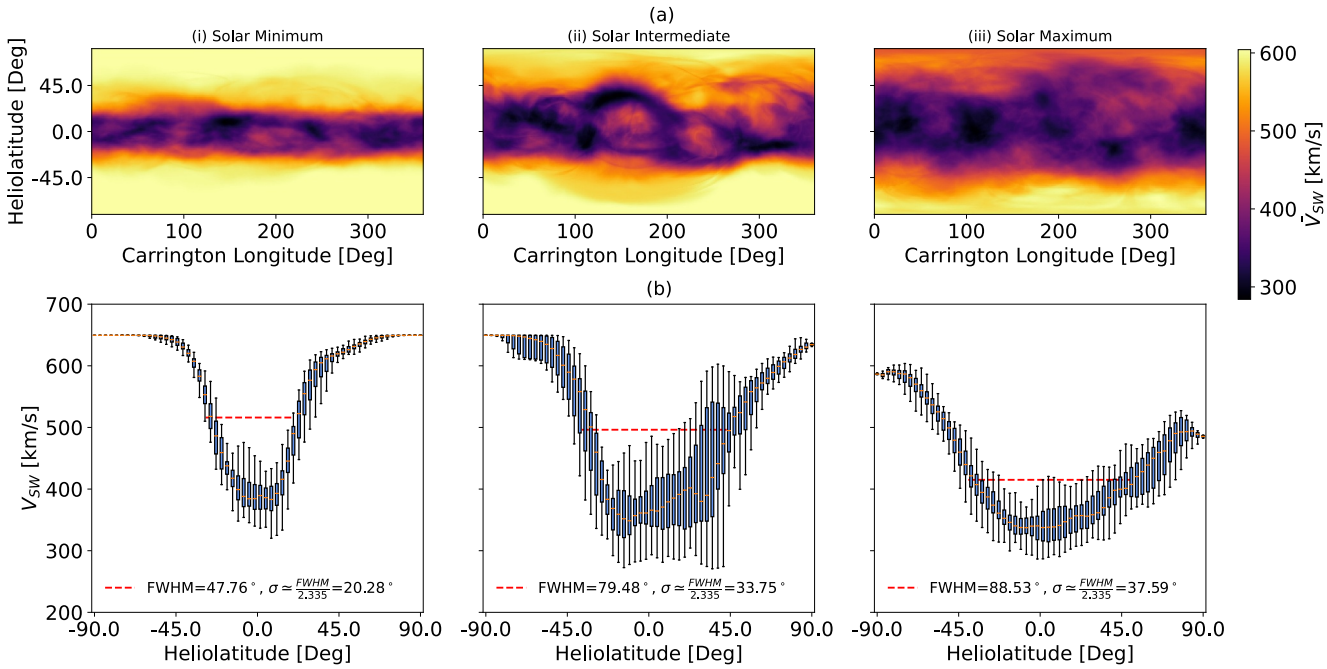

Figure 12. (a)Mean solar wind speeds at 30 $R_{\odot}$ from the MAS model during solar minimum, intermediate, and maximum data subsets. (b): The corresponding box-plots of zonal-mean solar wind speed. The orange line shows the median wind speed, the box extends from the first quartile to the third quartile of the data, the whiskers extend from the box to the farthest data point lying within 1.5x the inter-quartile range. The red-dashed line marks the wind-speed distribution Full Width Half Maxima (FWMH). FWHM and calculated scale parameter σ are labeled.

spatial perturbations produce similar calibration results. This is likely occurs for two reasons. First, during (and approaching) solar maximum, the ambient solar wind no longer captures the majority of high speed events at Earth, as CMEs become more prevalent in the observations used to evaluate the ensembles. Second, the ambient solar wind at 30 $R_{\odot}$ shows less variability at solar maximum than solar minimum (e.g., see Figure 12). This means that large spatial perturbations do not result in adequately large speed perturbations. And thus large spatial perturbations are unable to compensate for other (missing) sources of uncertainty. Indeed, Figure 6a shows that the locus of well-calibrated ensembles extends out into the more extreme regime of σ_{lon} parameters tested. Whilst this initially implies that more extreme perturbations in longitude can improve the ensemble distribution, later ROC score evaluations reveal that the extreme perturbations have a negative impact on forecast resolution (the ability for the ensemble to perform beyond climatology).

Looking at the trends in the reliability diagrams with increasing event threshold in Figure 7, the calibration appears to become worse as $V_{threshold}$ increases. This likely because a larger fraction of the evaluated high-speed events would have originated from CMEs present in the observations. Since CMEs were not included in the forecast model, this would naturally affect the calibration negatively, as it would adjust observed frequencies in ways not captured by the sets of initial inner-boundary conditions. However, we chose not to remove ICMEs from the verification data set as leaving them in more easily allows future comparison with models that do include CMEs and transient structures. The reliability diagrams also showed a growing trend toward overconfident forecasts with increasing action threshold. This is likely due to a bias present within the perturbed MAS boundaries where sampling from higher latitudes systematically increases solar-wind speed. This, compounded with a reduction in observed events at higher wind-speed thresholds, introduces the overconfidence.

BS reflects the overall behavior. If events of interest are very rare, such as at high speed thresholds, then the BS can be dominated by the majority class (no event) (Wilks, 2010). We find that at high $V_{thresholds}$, where events are rarer, a lower BS is seen (decreasing from left to right within Figure 8a). The reduction in BS is further amplified with large perturbation parameters, as the ensemble will forecast a greater number of events but with lower confidence (lower probabilities), as consensus between ensemble members is reduced. So whilst initial inspection of the BS indicates that large scale parameters improve the forecast, we can refer to the ROC score to provide a more balanced perspective on the trend. ROC score is independent of the distribution of forecast probabilities and

so Figure 8b reveals how the highly perturbed ensembles are not performing as well, due to a weaker forecast resolution. A significant longitudinal perturbation causes forecasted wind speeds, such as a high-speed stream, to be spread out in time across ensemble members, making the ensemble spread better capture the true arrival time/peak of the event. However, this reduces agreement on arrival times between members, lowering event probability scores and forecast decisiveness. In the limit of extremely large longitudinal perturbations, this will approach a climatological forecast as the predictions become analogous to average behavior over a long time baseline. Calibration balances this trade-off, and using BS and ROC scores we can visualize the impacts on forecast accuracy versus decisiveness.

5.2. Physical Interpretation of Optimal Parameters

Calibration seeks to ensure that the ensemble spread is a measure of the forecast uncertainty, in this case resulting from the solar-wind inner-boundary conditions. As such, the optimal spread can potentially be linked to our understanding of the physical origins of the solar wind. The perturbation scale parameters appear to match well with the angular extent of the slow-wind band at solar minimum. Within the coronal models, the broad size and shape of the slow wind band is well-characterized by MAS, whereas the finer details (e.g., coronal hole shapes, turbulent fast/slow mixing regions along the band edges, etc.) are less constrained/resolved (Riley et al., 2021; Riley & Luhmann, 2012). This relationship can be illustrated using the zonal-mean (i.e., longitudinally averaged)) solar-wind speeds. The top panels of Figure 12 shows mean solar-wind speeds from the MAS solutions at solar minimum, maximum and intermediate stages. The dominant structures during each period become clear, with solar minimum showing a well constrained slow-wind band, tightly constrained about the equator. During (and approaching) solar maximum, slow solar wind becomes more ubiquitous at a greater range of latitudes. The bottom row in Figure 12 shows the zonal means. During Solar minimum, the scale parameter σ computed from the Full Width Half Maximum (FWHM) of the zonal mean speed distribution matches very closely with the optimal σ_{lat} found through the rank histogram analysis of 20°. So the optimal σ_{lat} more often captures the dominant coronal model features across the years of hindcasts it was evaluated against. Furthermore, this highlights the changing structure of coronal model solutions approaching/during solar maximum as they become increasingly dominated by slow wind across all latitudes. Hence, the zonal distributions become a lot wider and the perturbation scheme breaks down, as even larger perturbations do not produce sufficient speed variability in the inner-boundary conditions. Figure 6 shows that at solar maximum, much larger perturbations are needed to improve the distribution. However, the need for such extreme perturbations creates a nonphysical representation of the possible initial conditions and pushes the forecast away from any meaningful representation of the coronal model solution. Additionally, the observations used to evaluate these eras also becomes increasingly dominated by CMEs over the ambient solar wind. The increased variability of wind speed across the solar cycle can be seen in the boxplot distributions in the lower panel of Figure 12.

5.3. Assumptions

We assume that the uncertainty takes the form of a positional/rotational error. This is a pragmatic assumption and works reasonably well for the current applications. However, it will be essential to more accurately characterize this uncertainty as it will likely put limits on the ensemble interpretability. It also appears to be inadequate at higher solar activity levels. Analysis were also limited to solar minimum for the majority of our study, in part due to our use of Carrington maps which rely on the steady state assumptions which break down during solar maximum, future work using daily updated WSA maps could be done to extend analysis into more active periods.

The MAS coronal model used for generating the 11 years of hindcasts is not a viable option for operational forecasting. MAS is primarily a research tool and was used in this paper due to its large archive of solutions. Models such as WSA, which can run much faster and have been tuned for forecasting applications, will likely require subtly different calibration parameters. Adapting the calibration for use with WSA would also give scope for further development and refinement of the calibration procedure, in particular providing a step toward calibrated multi-model ensembles. This would be a powerful forecasting tool constructed of a series of calibrated ensembles using different coronal model inputs; assembled through a weighted combination of the various outputs. Multi-model techniques have been extensively explored and utilized in weather and climate numerical weather prediction systems (Krishnamurti et al., 2016). This would be a logical application of the outlined calibration methods as much of the multi-model framework has already been developed through the Space

Weather Empirical Ensemble Package (SWEEP) collaboration (https://www.ralspace.stfc.ac.uk/Pages/SWIMMR.aspx).

In generating the probabilistic forecast, a direct relationship between the fraction of ensemble members $>V_{threshold}$ and event probability has been assumed. As the reliability diagrams revealed, the linear relationship between the number of ensemble members predicting a certain speed and the forecast probability did not hold for the solar-wind ensemble. Statistical post-processing with Platt-scaling started to address this issue. However, much more work can be done to explore the applications and benefits of post-processing. ECMWF employ more sophisticated methods such as Bayesian Model Averaging and Ensemble Model Output Statistics which could be applied to the solar-wind ensemble (Gneiting, 2014; Raftery et al., 2005). It should be noted that the benefits of post-processing at this stage are limited as the calibration, so far, has only been applied to the ambient-wind conditions. For a fully comprehensive solar-wind ensemble to be achieved, the CME ensemble will also need to be calibrated. CME forecasts typically involve modeling CMEs within a pre-existing ambient-wind solution (as defined by the inner-boundary). Post-processing of the ambient conditions before considering the impact of CMEs within the model would likely be a detriment to the forecast, as the probabilities representing the ambient conditions would no longer represent the 'best guess' of ambient conditions as realised by the model. Post-processing is an augmentation of the predicted probabilities and has no impact on the model state. Hence, ambient wind probabilities would quickly diverge when compared with a model which includes CME perturbations to the inner-boundary. A fully calibrated CME ensemble would likely be the most effective stage for post-processing to be applied. However, the methods demonstrated in this paper attempt to show the viability of post-processing within solar-wind forecasting contexts.

5.4. Cautionary Note

We wish to emphasize that caution should be taken in the blanket application of the optimal calibration parameters. As discussed, the calibration can be affected by the phase of the solar cycle. However, calibration is likely further affected by the seasonal characteristics of a given solar phase. Solar cycle 24 had an unusually broad slow-wind band and we believe this has been reflected within the optimal parameters found by the calibration (Figure 12). Solar cycles 22 and 23 had comparatively narrow slow-wind bands at solar minimum (Owens et al., 2017). As discussed previously, a consequence of the perturbation scheme is that a rotation of the MAS solution more often results in an increase in solar-wind speed as the sub-Earth path is perturbed to sample more fast wind at higher/lower latitudes. Hence, generating an ensemble using the 'optimal' parameters from a low-activity period where the current sheet was very narrow (i.e., a thin slow-wind band) during cycle 23 results in a saturated ensemble spread. The confidence intervals cover the whole extent of forecastable ambient solar-wind speeds. In this case the link between forecast accuracy to the size of the perturbations breaks down and it is hard to interpret the ensemble output. We also expect that the optimal parameters are model dependent. Models such as WSA produce spatial structures which can be very different to MAS and sets the inner-boundary height at 21.5 $R_{\odot}$. Applying the optimal parameters naively to the WSA solutions would likely result in an ensemble which does not reflect the true scale of uncertainty within the inner boundary from WSA. This is not to invalidate the calibration, however it highlights that careful consideration of the perturbation scheme and coronal model is important for the application of calibration into solar-wind forecasting schemes. In an operational context this procedure would be employed on archived forecasts or hindcasts with relevant verification data to find an optimal set of perturbation parameters much like how we demonstrate. The optimal parameters can then be used within the operational set-up. As we try to show, the calibration provides a more reliable forecast which operationally means a more trustworthy and interpretable forecast.

6. Conclusion

We have generated HUXt solar-wind ensembles through spatially perturbing the inner-boundary conditions derived from the MAS coronal model. Using an 11-year data set of hindcasts, we have demonstrated that calibration—through adjusting the scale of the perturbations—shows a clear improvement in ensemble distribution and forecast reliability as measured by a number of different forecast metrics. Rank histograms and reliability diagrams revealed an optimal set of perturbation parameters $(\sigma_{lat}, \sigma_{long}) = (20\,°, 10°)$ which control the extent to which the inner-boundary is spatially perturbed. Brier and ROC scores revealed that calibration improved forecast reliability and resolution, observed through an average 9% shift toward lower BS and a 2.5% increase in ROC score. A small-but-systematic increase in forecast performance between ensembles with more

optimal perturbation parameters against the remaining evaluated ensembles. The optimal parameters were linked to the average angular extent of the slow-wind band at solar minimum. We theorize this is representative of the uncertainty within coronal models where the broad structure of the slow-wind band can be well described but finer details still harbor large amounts of uncertainty which are now being more accurately captured by the ensemble.

Cost/loss analysis highlighted how the probabilistic forecast can provide a more actionable forecast in high cost/loss regimes, where false alarms are expensive and deterministic models struggle. Statistical post-processing through Platt-scaling was then employed to correct for the overconfidence bias within the HUXt ensemble. This showed a noticeable improvement across majority of cost/loss regimes. In both scientific and operational contexts, calibration serves to more accurately constrain the levels of uncertainty within the models. Understanding and accurately characterizing the ambient wind remains a critical component in the modeling and forecasting of space weather as it is strongly linked with predictions of severe space-weather phenomena such as CMEs. But for a comprehensive solar-wind forecast, the CME ensemble remains to be calibrated in this context.

We emphasize that the calibration is likely very sensitive to the model set-up. However, as the methodology described is largely generic it could easily be adapted to other ensemble arrangements such as WSA which is often used operationally. An important aim of calibration is to introduce more evidence-based judgment when deciding perturbation scales; it allows existing models to be used in a more optimal fashion. The spread adjustment of inner-boundary conditions is something that could also be applied to new perturbation schemes. This work marks a beginning for applying formal calibration techniques to solar-wind ensembles.

Acronyms

BS	Brier Score
CDF	Cumulative Distribution Function
CME	Coronal Mass Ejection
ECMWF	Eurppean Center for Medium-Range Weather Forecasts
HUXt	Heliospheric Upwind eXtrapolation with Time dependence
MAS	Magnetohydrodynamic Algorithm outside a Sphere
MHD	Magnetohydrodynamics
EUHFORIA	EUropean Heliospheric FORecasting Information Asset
OMNI	No special abbreviation, just “variety”
PEV	Potential Economic Value
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
WSA	Wang-Sheeley-Arge

Data Availability Statement

Verification data was sourced from OMNI solar-wind observations which can be found: https://omniweb.gsfc.nasa.gov/form/dx1.html.

HUXt is an open-source solar-wind model available at: Owens and Barnard (2024).

The archive of MAS coronal model solutions used within this paper can be found here: https://www.predsci.com/mhdweb/home.php.

Notebooks written to develop and explore the calibration can be found at: Edward-Inatimi (2024).

Acknowledgments
We would like to thank the providers of the data for this project. Nathaniel Edward-Inatimi is funded through SCENARIO Grant NE/S007261/1. Harriet Turner is funded through NERC Grant NE/Y001052/1. This work was part-funded by Science and Technology Facilities Council (STFC) grant number ST/V000497/1.

References

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*. https://doi.org/10.3389/fphy.2022.1005621

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

Cannon, P. S. (2013). Extreme space weather—A report published by the UK royal academy of engineering. *Space Weather*, *11*(4), 138–139. https://doi.org/10.1002/swe.20032

Cranmer, S. R. (2019). Solar-wind origin. In *Oxford research encyclopedia of physics*. https://doi.org/10.1093/acrefore/9780190871994.013.18

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, *77*(379), 605–610. https://doi.org/10.1080/01621459.1982.10477856

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(1/2), 12–22. https://doi.org/10.2307/2987588

Edward-Inatimi, N. O. (2024). University-of-Reading-Space- science/Ensemble_Calibration: Code for publication [software]. *Zenodo*. https://doi.org/10.5281/zenodo.14008183

Gneiting, T. (2014). Calibration of medium-range weather forecasts (No. 719). *ECMWF*. https://doi.org/10.21957/8xna7glta

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, *129*(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2

Hamill, T. M., & Colucci, S. J. (1998). Evaluation of eta–rsm ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, *126*(3), 711–724. https://doi.org/10.1175/1520-0493(1998)126⟨0711:EOEREP⟩2.0.CO;2

Hosteaux, S., Chané, E., & Poedts, S. (2019). Effect of the solar wind density on the evolution of normal and inverse coronal mass ejections. *Aqua*, *632*, A89. https://doi.org/10.1051/0004-6361/201935894

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). https://doi.org/10.1029/2004JA010649

Krishnamurti, T. N., Kumar, V., Simon, A., Bhardwaj, A., Ghosh, T., & Ross, R. (2016). A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Reviews of Geophysics*, *54*(2), 336–377. https://doi.org/10.1002/2015RG000513

Lang, M., Browne, P., van Leeuwen, P. J., & Owens, M. J. (2017). Data assimilation in the solar wind: Challenges and first results. *Space Weather*, *15*(11), 1490–1510. https://doi.org/10.1002/2017SW001681

Leutbecher, M. (2019). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 107–128. https://doi.org/10.1002/qj.3387

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

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, *30*(4), 291–303.

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

Mikić, Z., Downs, C., Linker, J. A., Caplan, R. M., Mackay, D. H., Upton, L. A., et al. (2018). Predicting the corona for the 21 August 2017 total solar eclipse. *Nature Astronomy*, *2*(11), 913–921. https://doi.org/10.1038/s41550-018-0562-5

Milinski, S., Maher, N., & Olonscheck, D. (2020). How large does a large ensemble need to be? *Earth System Dynamics*, *11*(4), 885–901. https://doi.org/10.5194/esd-11-885-2020

Millward, G., Biesecker, D., Pizzo, V., & de Koning, C. A. (2013). An operational software tool for the analysis of coronagraph images: Determining cme parameters for input into the wsa-enlil heliospheric model. *Space Weather*, *11*(2), 57–68. https://doi.org/10.1002/swe.20024

Murphy, A. H. (1985). Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation. *Monthly Weather Review*, *113*(3), 362–369. https://doi.org/10.1175/1520-0493(1985)113⟨0362:DMATVO⟩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

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., Chakraborty, N., Turner, H., Lang, M., Riley, P., Lockwood, M., et al. (2022). Rate of change of large-scale solar-wind structure. *Solar Physics*, *297*(7), 83. https://doi.org/10.1007/s11207-022-02006-4

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., Lockwood, M., & Riley, P. (2017). Global solar wind variations over the last four centuries. *Scientific Reports*, *7*(1), 41548. https://doi.org/10.1038/srep41548

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

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, *10*.

Pomoell, J., & Poedts, S. (2018). Euhforia: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, *8*, A35. https://doi.org/10.1051/swsc/2018020

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, *133*(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Reiss, M. A., Muglach, K., Mullinix, R., Kuznetsova, M. M., Wiegand, C., Temmer, M., et al. (2023). Unifying the validation of ambient solar wind models. *Advances in Space Research*, *72*(12), 5275–5286. https://doi.org/10.1016/j.asr.2022.05.026

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., Linker, J. A., Lionello, R., & Mikic, Z. (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. https://doi.org/10.1016/j.jastp.2011.12.013

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), 15889–15902. https://doi.org/10.1029/2000JA000121

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

Riley, P., Lionello, R., Caplan, R. M., Downs, C., Linker, J. A., Badman, S. T., & Stevens, M. L. (2021). Using Parker Solar Probe observations during the first four perihelia to constrain global magnetohydrodynamic models. *Aqua*, *650*, A19. https://doi.org/10.1051/0004-6361/202039815

Riley, P., & Luhmann, J. G. (2012). Interplanetary signatures of unipolar streamers and the origin of the slow solar wind. *Solar Physics*, *277*(2), 355–373. https://doi.org/10.1007/s11207-011-9909-0

Schrijver, C. J. (2015). Socio-economic hazards and impacts of space weather: The important range between mild and extreme. *Space Weather*, *13*(9), 524–528. https://doi.org/10.1002/2015SW001252

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

Temmer, M., Scolini, C., Richardson, I. G., Heinemann, S. G., Paouris, E., Vourlidas, A., et al. (2023). Cme propagation through the heliosphere: Status and future of observations and model development. *Advances in Space Research*. https://doi.org/10.1016/j.asr.2023.07.003

Viall, N. M., DeForest, C. E., & Kepko, L. (2021). Mesoscale structure in the solar wind. *Frontiers in Astronomy and Space Sciences*, *8*. https://doi.org/10.3389/fspas.2021.735034

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

Wilks, D. S. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, *8*(2), 209–219. https://doi.org/10.1017/s1350482701002092

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. (2011). *Statistical methods in the atmospheric sciences*. Academic press.

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). Chapter 9 - Forecast verification. In D. S. Wilks (Ed.), *Statistical methods in the atmospheric sciences* (4th ed., pp. 369–483). Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00009-2

Wilks, D. S. (2019c). 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

Xie, H., Ofman, L., & Lawrence, G. (2004). Cone model for halo CMEs: Application to space weather forecasting. *Journal of Geophysical Research (Space Physics)*, *109*(A3), A03109. https://doi.org/10.1029/2003JA010226

Zhao, X. P., Plunkett, S. P., & Liu, W. (2002). Determination of geometrical and kinematical properties of halo coronal mass ejections using the cone model. *Journal of Geophysical Research (Space Physics)*, *107*(A8), 1223. https://doi.org/10.1029/2001JA009143

Zhu, Y. (2005). Ensemble forecast: A new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, *22*(6), 781–788. https://doi.org/10.1007/BF02918678