University of Reading

Causal analysis of influence of the solar cycle and latitudinal solar-wind structure on co-rotation forecasts

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Chakraborty, N. ORCID: https://orcid.org/0000-0002-3134-1946, Turner, H. ORCID: https://orcid.org/0000-0002-4012-8004, Owens, M. ORCID: https://orcid.org/0000-0003-2061-2453 and Lang, M. ORCID: https://orcid.org/0000-0002-1904-3700 (2023) Causal analysis of influence of the solar cycle and latitudinal solar-wind structure on co-rotation forecasts. Solar Physics, 298. 142. ISSN 1573-093X doi: 10.1007/s11207-023-02232-4 Available at https://centaur.reading.ac.uk/114369/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1007/s11207-023-02232-4

Publisher: Springer

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Solar Physics (2023) 298:142
https://doi.org/10.1007/s11207-023-02232-4

RESEARCH

Causal Analysis of Influence of the Solar Cycle and Latitudinal Solar-Wind Structure on Co-Rotation Forecasts

Nachiketa Chakraborty[1,2] · Harriet Turner[2] · Mathew Owens[2] · Matthew Lang[2]

Received: 7 February 2023 / Accepted: 17 November 2023
© The Author(s) 2023

Abstract

Studying solar-wind conditions is central to forecasting the impact of space weather on Earth. Under the assumption that the structure of this wind is constant in time and co-rotates with the Sun, solar-wind and thereby space-weather forecasts have been made quite effectively. Such co-rotation forecasts are well studied with decades of observations from STEREO and near-Earth spacecraft. Forecast accuracy is primarily determined by three factors: i) the longitudinal separation of spacecraft from Earth determines the corotation time (and hence forecast lead time) [δt] over which the solar wind must be assumed to be constant, ii) the latitudinal separation (or offset) between Earth and spacecraft [$\delta\theta$]] determines the degree to which the same solar wind is being encountered at both locations, and iii) the solar cycle, via the sunspot number (SSN), acts as a proxy for both how fast the solar-wind structure is evolving and how much it varies in latitude. However, the precise dependencies factoring in uncertainties are a mixture of influences from each of these factors. Furthermore, for high-precision forecasts, it is important to understand what drives the forecast accuracy and its uncertainty. Here we present a causal inference approach based on information-theoretic measures to do this. Our framework can compute not only the direct (linear and nonlinear) dependencies of the forecast mean absolute error (MAE) on SSN, $\Delta\theta$, and Δt, but also how these individual variables combine to enhance or diminish the MAE. We provide an initial assessment of this with the potential of aiding data assimilation in the future.

✉ M. Owens
m.j.owens@reading.ac.uk

N. Chakraborty
ae0221@coventry.ac.uk

H. Turner
h.turner3@pgr.reading.ac.uk

M. Lang
matthew.lang@reading.ac.uk

1 School of Computing, Electronics and Mathematics, Coventry University, Coventry, UK

2 Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK

Published online: 13 December 2023

1. Introduction

Forecasting terrestrial space-weather impacts (e.g. Cannon et al., 2013) necessitates knowledge of the up-stream solar-wind conditions that will encounter the Earth's magnetosphere in the future. Advanced knowledge of the near-Earth solar wind is directly important for space-weather impacts, but also for understanding the medium through which transient events occurrences such as coronal mass ejections (CMEs) and solar energetic particles (SEPs) must propagate. Currently, direct (in-situ) solar-wind observations are only routinely available near the Sun–Earth line at the first Lagrange point L_1, giving less than 40 minutes' forecast lead time. Physics-based simulations of the whole Sun–Earth system can potentially provide forecast lead times of two to five days, but there remain many technical and scientific challenges to this approach (Luhmann et al., 2004; Toth et al., 2005; Merkin et al., 2007). A simple, yet robust, alternative forecast of near-Earth solar-wind conditions can be made using observations anywhere in the ecliptic plane by assuming that the structure of the solar wind is fixed in time and co-rotates with the Sun. For example, observations in near-Earth space can be used to predict conditions at the same location a whole solar (synodic) rotation ahead, approximately 27.27 days (Bartels, 1934; Owens et al., 2013; Kohutova et al., 2016). Of course, the structure of the corona and solar-wind does evolve over such time scales, particularly around solar maximum. From the L_5 Lagrange point, approximately 60° behind Earth in its orbit, the co-rotation time is approximately five days. This is sufficiently long that the forecast lead time is useful, but sufficiently short that the co-rotation approximation is generally appropriate (Simunac et al., 2009; Thomas et al., 2018). Partly for these reasons, *Vigil*, the upcoming operational space-weather monitor, will make routine observations at L_5 (Kraft, Puschmann, and Luntama, 2017).

Assessing and quantifying the factors influencing the accuracy of co-rotation forecasts is important directly for improving co-rotation forecasting, but also for effective data assimilation of the solar-wind observations into solar-wind models (Lang and Owens, 2019; Lang et al., 2021), as it informs the expected observational errors. Longitudinal separation between the observing spacecraft and the forecast point – and hence the forecast lead time – is obviously expected to increase forecast error, as the steady-state assumption becomes increasingly invalid. We may also expect that this effect would be more pronounced (and co-rotation forecasts generally less accurate) around sunspot maximum, when the corona is known to be more dynamic and the occurrence of time-dependent coronal mass ejections (CMEs) increases (Yashiro et al., 2004). However, for evidence that this effect is reduced near the ecliptic plane, the reader is referred to Owens et al. (2022). Similarly, it has been argued using simulation data that the co-rotation forecast error should increase with latitudinal separation of observing spacecraft from forecast position (Owens et al., 2019) and that this effect is maximised at sunspot minimum (Owens et al., 2020). This is the result of greater latitudinal ordering of the solar-wind – with slow wind at the Equator and fast wind at the Poles – at times when the solar dipole dominates and is rotationally aligned, which is primarily at solar minimum (McComas et al., 2003).

The OMNI dataset of near-Earth solar-wind observations (King and Papitashvili, 2005) allows us to assess co-rotation forecasts over nearly five complete solar cycles. As near-Earth observations are used to make near-Earth forecasts one solar rotation ahead, the forecast lead time is fixed at 27.27 days, and the latitudinal separation, caused by Earth's motion over a solar rotation, reaches a maximum value of around 3.5°. The twin spacecraft of the *Solar-Terrestrial Relations Observatory* (STEREO: Kaiser, 2005) provide a means to assess the performance of co-rotation forecasts over a larger parameter range. The spacecraft were launched into Earth-like orbits in late 2006, with STEREO-A moving ahead of Earth in

its orbit, and STEREO-B behind, separating from Earth at a rate of 22.5° per year. This allows the co-rotation forecast to be assessed for a full range of longitudinal separations – and hence forecast lead times between 0 and 27.27 days – and, due to the inclination of the ecliptic plane to the solar equator, latitudinal separations covering the range ±15°. More than a solar cycle of data is available (although the STEREO-B spacecraft was lost in 2014), allowing the effect of increasing solar activity to be estimated.

However, while uniquely valuable, assessing co-rotation forecasts with the STEREO dataset does present a number of challenges. Longitudinal and latitudinal separation from Earth are interdependent, as both are due to the same orbital geometry. Due to timing of launch and the orbital period, solar activity also varies approximately in step with the orbit; the spacecraft were launched just before sunspot minimum and reached maximum separation just after sunspot maximum. Thus it is difficult to isolate and quantify the individual sources of error in co-rotation forecasting (Turner et al., 2021). This kind of problem is perfectly suited for causal analysis.

Study of cause and effect is central to all branches of sciences, and there are questions in solar physics – such as factors affecting co-rotation forecasts – that can be cast in those terms. In non-interventional (or observational) systems like the Sun, causal discovery is the process of inferring mechanisms or models relating cause and effects from data. But even when the principal mechanisms are known from physics, causal frameworks can also be used as a diagnostic tool to determine how uncertainty in one or more variables influences another. This is very useful in making forecasts. Typically, establishing a causal relationship between variables entails determining their conditional dependency (Granger, 1969; Pearl, 2000). For random variables, both continuous and discrete, this is done via probabilistic measures. Conditional dependency has traditionally been established with Granger causality, and these measures are mostly derived from information theory, i.e. they are "Shannon based" (Schreiber, 2000; Kraskov, Stögbauer, and Grassberger, 2004; Williams and Beer, 2010). In addition, for time-series data, the temporal order of events is also critical to establishing causality. Time lags between different variables need to be carefully evaluated. Therefore the temporal resolution of time series must be sufficient for establishing the direction of information flow; missing data can lead to spurious correlations (Runge, 2018). Nonlinear correlations between multiple drivers can be very difficult to disentangle. We attempt here to address and demonstrate this with a framework (van Leeuwen et al., 2021) that uses a transformed information-theoretic measure that applies to both discrete and continuous variables. Typically, the current state-of-the-art causal estimates are point estimates: data are used to produce a single number to quantify the causal relationships. There is no robust uncertainty quantification. Addressing this in general is a work in progress (e.g. Runge, 2018; Heckerman, 2020). However, we will provide an elementary estimate of the distribution of the strength of causal relationships – the causal strength, *cs* from hereinafter.

Our goal in this work is to provide an initial assessment of the causal dependencies between the accuracy of a forecast, the *target* or *"effect" variable*, with the *driver* or *"cause" variables*. For reasons explained above, the driver variables are assumed to be solar activity (quantified by sunspot number), forecast lead time (which is primarily determined by longitudinal spacecraft separation for the OMNI data and STEREO observations), and latitudinal spacecraft separation. The typical approach would be to cross-correlate these variables, or rather the time series associated with them, pairwise. However, as these relationships can often be nonlinear and multivariate, we need more advanced estimators such as those based on information-theoretic measures such as mutual information and higher-order terms (Chakraborty and van Leeuwen, 2022). So the approach that we follow here is to start with the analog of pairwise correlation, but with the nonlinear estimator: mutual information.

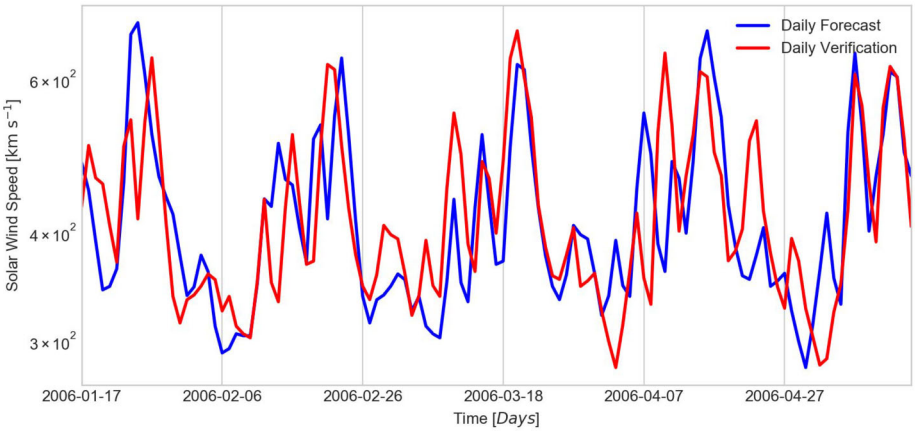

Figure 1 The forecast and the verification speeds for a small subset of the total dataset (between 16 January 2006 and 15 May 2006). *The blue curve* shows the forecast speed, and *the red one* the verification speed. Both properties are calculated at one-day resolution.

We then introduce a third variable, via conditional mutual information, to disentangle interdependencies amongst three driver/cause variables, in order that mediated or induced effects can be isolated. In principle, a full causal network (Runge, 2018; van Leeuwen et al., 2021) can be constructed using time-series observations. But this comes with computational and, in certain situations, interpretation challenges. Hence we leave this for future work.

We describe the solar-wind observations from OMNI and STEREO-A and -B spacecraft in Section 2. Next, we introduce the causal inference methods, demonstrating their application to the OMNI data in Section 3. We compute the distribution of causal relationships, first pairwise, quantified in terms of the mutual information using a nonlinear information-theoretic measure (Section 3.1), examine the time averaging effect on sunspot number (Section 3.2), followed by the conditional mutual information to separate influence of the third variable (Section 3.3). We use 27-day co-rotation forecasts (also called "recurrence" or "27-day persistence" forecasts) using only OMNI data first, as it eliminates the lead-time as a variable by design; this leaves us with testing two (instead of three) drivers: the solar activity encoded in the (smoothed) Sunspot Number (or SSN_{27}), and the latitudinal offset. By first learning dependencies in this simpler dataset, we then compare effects of this same subset of drivers in the STEREO datasets ignoring at first the lead time (Section 3.4). Following this, in Section 5, we study induced or mediated dependencies with lead time included by using the STEREO datasets. Finally, we interpret the results and conclude whilst looking at future opportunities to improve forecasts in Section 6.

2. Observations

Two primary data sets are used in this study. Firstly, the OMNI dataset of near-Earth solar-wind conditions (King and Papitashvili, 2005). Data are available from omniweb.gsfc.nasa.gov/. Prior to 1995, data coverage varies significantly, so the period of study is limited to 1995 to present. Secondly, the STEREO dataset, which is available from stereo-ssc.nascom.nasa.gov/data.shtml. STEREO-A data are used from the whole mission,

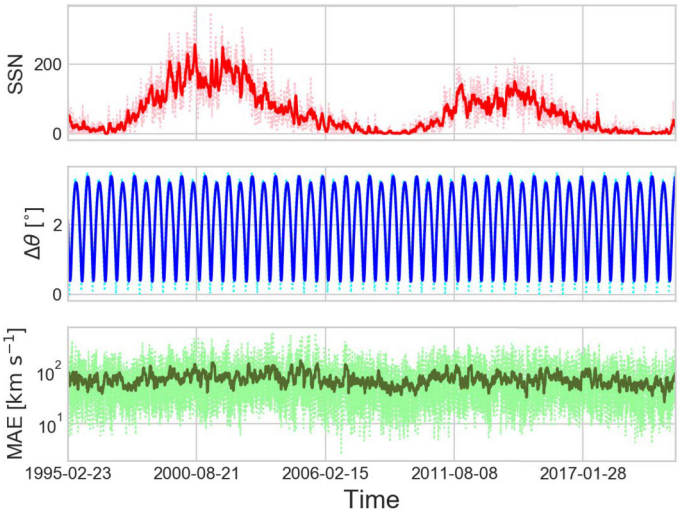

Figure 2 Subset of data (2006) showing speed(s) with MAE. A summary of the co-rotation forecast of solar-wind speed obtained by using OMNI near-Earth data to forecast near-Earth conditions 27.27 days ahead. *Top*: Sunspot number. *Middle*: The absolute value of the latitudinal separation between observation and forecast location [$\Delta\theta$]. *Bottom*: The mean absolute error or MAE in the solar-wind speed co-rotation forecast. All properties are calculated at one-day resolution (*dotted lighter curves*), then averaged over 27 days (*solid darker curves*).

2007 – present, whereas STEREO-B data are only available until 2014. All data are averaged to one-day resolution to remove the effect of small-scale stochastic structures, such as waves and turbulence (Verscharen, Klein, and Maruca, 2019).

Solar-wind speed co-rotation forecasts are produced by ballistically mapping data at the observed solar-wind speed from the observation radial distance to 1 AU, then applying a co-rotation delay consistent with the longitude separation. By far the dominant factor is the longitudinal separation. Further details can be found in Turner et al. (2021). For each forecast, we compute the mean absolute error (MAE) between the forecast and observed solar-wind speed. For reference, we have plotted the wind forecast speed and the verification speed for a small subset of the data in Figure 1. The curves show the daily values (red for verification and blue for forecast), which are noisy as expected. We can see even from this small (four-month-long) sample that the difference between the two – daily verification and the forecast – is compatible with the mean absolute error plotted in Figures 2 and 3.

For solar-cycle context, we use the daily sunspot number (SN) provided by Sunspot Index and Long-term Solar Observations (SILSO: Clette and Lefèvre, 2016) and available from www.sidc.be/silso/. While we use version 2.0 of the SILSO record, the time period considered in this study is not subject to any of the calibration issues or corrections that are necessary for the early data (Clette et al., 2023).

Figure 2 shows a summary of OMNI data used to make a 27.27-day lead time forecast of near-Earth conditions. By eye, some correlation can be seen between the MAE and SN, e.g. there are few intervals of MAE above 250 km s^{-1} during the solar minima of 1996 – 97, 2009 – 10, or 2019 – 20. Conversely, there is no immediately obvious relation between MAE and the absolute latitudinal separation between observation and forecast location: $\Delta\theta$. However, the $\Delta\theta$-variation here is very small, arising from Earth's latitudinal orbital motion over a 27.27-day interval and reaching a maximum magnitude of around 3.5°.

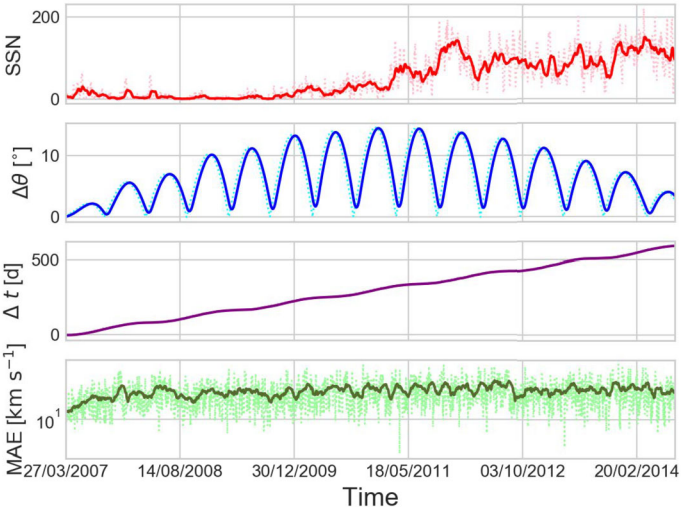

Figure 3 A summary of the co-rotation forecast of solar-wind speed obtained by using STEREO-B observations to forecast conditions at the STEREO-A spacecraft. *Top*: Sunspot number. *Second row*: The absolute value of the latitudinal separation between observation and forecast location [$\Delta\theta$]. *Third row*: Forecast lead time (directly proportional to longitudinal separation, and to a much lesser extent, radial separation of spacecraft). *Bottom*: The mean absolute error in the solar-wind speed co-rotation forecast. All properties are calculated at one-day resolution (*dotted lighter curves*), then averaged over 27 days (*solid darker curves*).

Figure 3 shows the summary of STEREO-B observations used to forecast solar-wind speed at STEREO-A. As the spacecraft separate in longitude, the forecast lead time [Δt] increases almost linearly. The maximum value of $\Delta\theta$ grows as the spacecraft increase their absolute longitudinal separation until mid-2010, then declines as the spacecraft move closer together (behind the Sun, from Earth's point of view). There is a somewhat linear growth in MAE from 2007 to 2012, although without further analysis it is not possible to say whether this is the result of sunspot number (post smoothing as we will see) Δt, or the amplitude of $\Delta\theta$, increasing through this time, or some combination of those variables.

3. Methods: Causal Dependencies of Co-Rotation Forecasts

We wish to study the principle drivers of the error in the co-rotation forecasts. To do that, we perform a causal analysis on the mean absolute error (MAE) as the target/effect variable and the sunspot number SN, latitudinal offset [$|\Delta\theta|$: °], and the forecast lead time [Δt: days] as the principal driver/cause variables. With this setup, we can use a nonlinear measure of dependency to compute the causal relationships between these variables. There are a number of choices for such measures: those based on information theory as (conditional) mutual information (Kraskov, Stögbauer, and Grassberger, 2004), transfer entropy (Schreiber, 2000), directed information transfer (Amblard and Michel, 2009), etc. We chose the mutual information (and its conditional variants) as it is well studied (e.g. Runge, 2015; van Leeuwen et al., 2021), and there are robust estimators available, along with an analytical result for Gaussian variables. The mutual information $I(x; y_{1:N})$ between a target process x and a possible driver process y, or a whole range of driver processes denoted in our general formalism (van Leeuwen et al., 2021) by $y_{1:N}$ (or sometimes *y, z, w,* etc.) is defined via the

Shannon entropy $H(\cdot)$ as

$$I(x; y_{1:N}) = H(x) - H(x|y_{1:N}) \tag{1}$$

$$= \int p(x, y_{1:N}) \log \left[\frac{p(x, y_{1:N})}{p(x)\ p(y_{1:N})} \right] \mathrm{d}x\, \mathrm{d}y_{1:N}. \tag{2}$$

Mathematically, the mutual information $I(x; y_{1:N})$ is a positive-definite quantity. It can be thought of as the reduction in entropy (or uncertainty) in the target (here x) in the presence of information content from the driver variables (here $y_{1:N}$).

3.1. Mutual Information: Pairwise Dependency Between Latitudinal Offset, Sunspot Number, and MAE

With the goal of disentangling causal influences of drivers in co-rotation forecasts, we begin with OMNI data used to make a forecast at Earth. In this case, co-rotation forecasts have a fixed lead time of 27.27 days, and forecast error [MAE] inherently has two primary drivers, the activity modulation of the Sun – approximated by the sunspot number (SN) – and latitudinal offset $\Delta\theta$ between the observation and forecast position (i.e. between Earth's location 27.27 days apart). This provides a relatively simple causal network to explore with our framework.

We compute the mutual information (MI) between pairs of the target and one of the drivers, e.g. $I(\mathrm{MAE}; |\Delta\theta|)$. Given the length of the observation time series, we can empirically estimate the distribution of these quantities as histograms. The mutual information serves as the measure of causal dependency between pairs of one of the drivers and the target variable. Once again, we refer to Figure 6 for a visualisation graphically via a Venn diagram described in Section 4.1. In this figure, the example is given for variables x and y representing target MAE, and driver either $\Delta\theta$ or SN. In other words, we determine the reduction in entropy (or random uncertainty) in MAE due to $\Delta\theta$ or SN. Note that we break the symmetry between the two variables (target vs. driver), with the driver (cause) as lagging in time with respect to the target (effect). The quantities (or rather their distributions) represented by these information diagrams are estimated in Figure 4.

As a positive-definite quantity with no upper limit, MI can take very large values. Thus it is useful to normalise this measure, which is possible in a number of ways. One option is to normalise it with the total entropy or uncertainty in the variable x, giving the causal strength $cs(x; y_{1:N}) = \frac{I(x; y_{1:N})}{H(x)}$ or simply $cs(x; z) = \frac{I(x;z)}{H(x)}$ for two variables: target x and driver z. There is a challenge here; the entropy we use is for continuous variables, also known as the differential entropy, which can acquire negative values. In practice, we do not encounter this here in our applications. However, to mitigate this effect – and for general interpretation – we will ultimately use relative causal strengths to the total over all the drivers combined; in these relative causal strengths, we ignore the contribution of noise or unmodeled drivers to merely focus on interpreting selected drivers.

3.2. Influence of Sunspot Number: Timescale Matters

The measured or observed quantity for solar activity is the daily sunspot number. These observations display large variability as seen in Figures 2 and 3. As we will demonstrate here, the stochasticity has an impact on the causal association with the forecast-accuracy term, MAE. Figure 4 shows the corresponding distribution of causal strengths of the pairs of MAE with 27-day smoothed (right) and daily unsmoothed (left) SN and the latitudinal

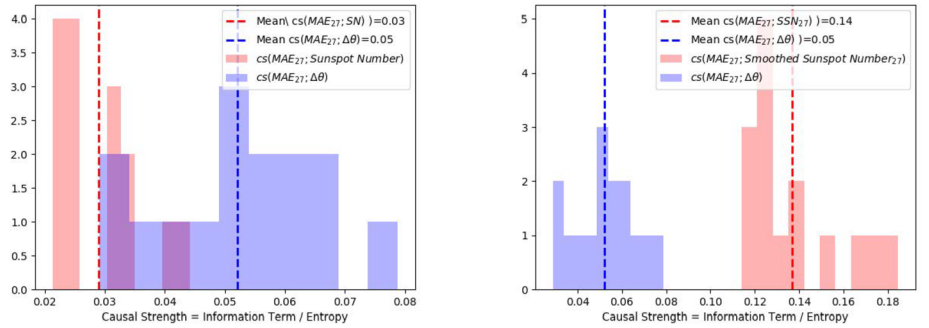

Figure 4 Histograms of causal strengths [cs] of drivers – sunspot number (SN) and latitudinal offset $\Delta\theta$ – on the target, co-rotation MAE obtained from 27.27-day forecasts using OMNI data. *Dashed lines* represent mean cs-values used in causal diagrams. We take the 27-day smoothed MAE as the target in both cases. *Left:* SN and $\Delta\theta$ are used at daily resolution; cs computed from daily SN. *Right:* cs computed from 27-day smoothed SN. cs for SN shows a greater dependence on $|\Delta\theta|$ than on SN, whereas the 27-day smoothed SN clearly shows the greater association of solar cycle with MAE by suppressing the stochasticity and emphasising the solar activity.

offset $\Delta\theta$. The unsmoothed daily sunspot number (SN) has lower cs [= MI/H] than the latitudinal offset $\Delta\theta$. However, upon performing a rolling mean on the daily SN to yield 27-day smoothed averaged SN or the SSN_{27}, the hierarchy reverses. As shown in Figure 4, we see that the total causal strength of SN, $cs(SSN_{27} \rightarrow MAE_{27})$, goes from 0.03 to 0.14 upon averaging, compared to $cs(\Delta\theta \rightarrow MAE_{27})$ with a mean value of 0.05. As expected, the stage of solar cycle and overall time variability of the Sun is better represented by the smoothed SN, which has a significant influence on MAE. That the daily SN has a significant stochastic component is also confirmed by / evident from the entropy estimates. The entropy is reduced upon smoothing or averaging SN and is lower than that of $\Delta\theta$ by a factor of a few (for example ≈ 2 for STB–STA); however, this is not a significant effect.

3.3. Conditional and Interaction Information: Higher-Order Terms

In the presence of multiple causes or drivers (say y and z), the aforementioned causal-strength term $cs(x; z)$ will come to represent the fractional reduction in uncertainty in the target due to the driver: z. There is a similar term for y. To further disentangle and isolate the influence of each driver, we also compute the conditional mutual information (CMI), e.g. $I(x; z|y)$. For two drivers y and z (and a single target x), conditional mutual information $I(x; z|y)$ given in Equation 3, "conditions out" the effect of one driver (y), leaving the direct influence of the other one (z). This can be visualised in terms of Venn diagrams in Figure 7. It is the difference between the intersection of x- and z-circles (black stripes) and that of x-, z-, and y-circles (orange spots). In our application to co-rotation forecasts, the example used for illustration has x as MAE_{27} and the y and z as the drivers $\Delta\theta$ and SSN_{27}, respectively. We will keep the same normalisation with entropy for all information terms so that they can be combined or compared.

The conditional mutual information can be defined in terms of the conditional entropies as

$$I(x; y|z) = H(x|z) - H(x|y, z). \tag{3}$$

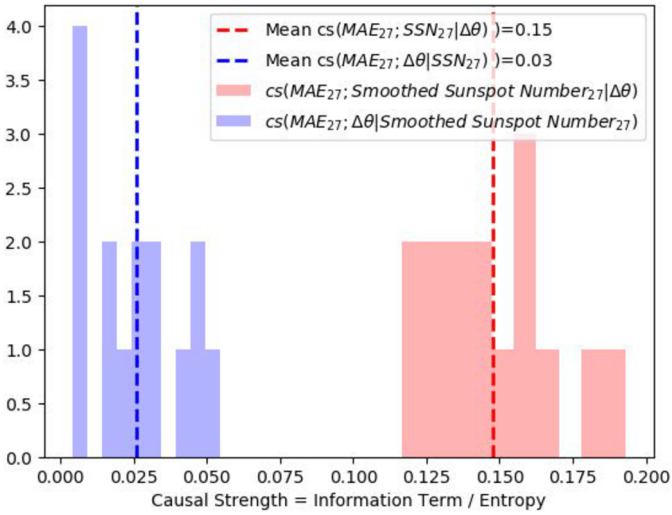

Figure 5 The distribution of conditional causal strengths with OMNI data. *cs* here is associated with conditional mutual information normalised by the entropy H for the combinations of MAE_{27} with Sunspot Number SSN_{27} (27-day rolling averages for both) and Latitudinal Offset for the full dataset, namely $cs(MAE_{27}; SSN_{27}|\Delta\Theta)$ and $cs(MAE_{27}; \Delta\Theta|SSN_{27})$.

This equation for the conditional mutual information of x with respect to y and z represents the difference in entropy of x "conditioning out" z alone ($H(x|z)$) and entropy of x "conditioning out" y and z together $[H(x|y, z)]$. This leaves us with the direct influence of driver y on target x, excluding any indirect influence mediated by or shared with z. The distributions of such conditional information terms for the triplet (MAE_{27},$\Delta\theta$,SSN_{27}) are estimated in Figure 5; these provide the so-called direct causal influence contribution of SSN_{27} and $\Delta\theta$ on MAE_{27}. These are symbolised by the black arrows in the causal-summary diagram in Figure 8. The causal-summary diagram, as the name suggests, provides a summary of the information flow from (and therefore the causal influence of) the driver variables; in this case the latitudinal offset $\Delta\theta$ and the smoothed sunspot number SSN_{27}. Now the interaction information can be written in terms of the mutual and conditional mutual information as

$$\begin{aligned} I(x; y; z) &= I(x; y) - I(x; y|z) \\ &= I(x; z) - I(x; z|y). \end{aligned} \tag{4}$$

This equation for the interaction information of x with y and z gives the difference between the mutual information shared between x and y $[I(x; y)]$ and information shared between them, upon conditioning out z $[I(x; y|z)]$. This is the interaction information shared between the three variables x, y, and z and is symmetric in all three variables. If we fix one as the target with the other two as drivers, as we do for our application, then the expression for interaction information is symmetric in the two drivers as demonstrated by the two equivalent expressions for $I(x; y; z)$ in Equation 4. So it does not matter which driver we condition on. We will exploit this later on as estimates from actual measurements may not converge to the same value as Equation 4. So we can take the average of the two symmetric expressions to represent the interaction information between one target and two drivers. This is seen later in the observational estimates.

This quantity can be interpreted as the information shared between x and y, less the information shared between them when z is known. If the interaction information is non-negative, or $I(x; y) \geq I(x; y|z)$, it implies that the dependency of x on z partially or entirely (equality) constitutes the dependency on y (Ghassami and Kiyavash, 2017). If the interaction information is negative, or $I(x; y) < I(x; y|z)$, then each one of the variables induces and increases correlation between the other two. The case of zero interaction information implies that the mutual and conditional mutual informations are equal. This means that the entire information shared between the first two variables (here x and y) is the direct component of causal influence having conditioned out the third (here z).

In the previous subsection, we ascertained that the smoothed sunspot number or SSN_{27} is more appropriate as a proxy for the solar activity in evaluating its causal influence on the average co-rotation forecast accuracy MAE_{27}. Now we wish to disentangle the direct and indirect effects of both SSN_{27} and the latitudinal offset $\Delta\theta$ on MAE_{27}. Their joint effect, or one mediating through the other, is naturally a higher-order effect, and hence we need the higher-order information terms. We compute the higher-order information-theoretic quantities, namely conditional mutual information and interaction information between drivers SSN_{27} and $\Delta\theta$ and the target MAE_{27} still using only the OMNI dataset. For $MAE_{27}(x)$, $\Delta\theta(y)$, and $SSN_{27}(z)$, the interaction information corresponds to the common part with orange circles in the Venn diagram in Figure 7. This is therefore the information shared across all three variables in general.

Formally, causality necessitates there be a time lag between the cause and effect such that the former precedes the latter. Indeed, there is a time lag between the forecast accuracy of a future step MAE_{27}^{n+1} and the drivers $\Delta\theta^n$ and SSN_{27}^n. This is innate/intrinsic to the way the time-series observations are done. However, in this particular application, we are considering daily variations, and the drivers – latitudinal separation, longitudinal separation, and 27-day smoothed sunspot number – vary over much longer timescales. Thus a single time-step between n and $n + 1$ makes negligible difference to the computed information components. However, the notation involving target at $n + 1$ and drivers at n is maintained to demonstrate the general principle.

3.4. Consistency Across Datasets

We next test this relative influence of SSN_{27} and $\Delta\theta$ on MAE_{27} across the available datasets, namely STB–STA, STB–OMNI, and OMNI-STA pairings. This is shown in Figure 9. In each case, we find the interaction information $I(MAE_{27}^{n+1}; \Delta\theta^n; SSN_{27}^n)$ to be positive. This is an indication that SSN_{27} partially constitutes the dependency of MAE_{27} on $\Delta\theta$ and vice versa, but it is not very significant. Across these three datasets (as well as OMNI–OMNI recurrence forecasts), we found that the direct causal strengths of latitudinal offset $I(MAE_{27}^{n+1}; \Delta\theta^n|SSN_{27}^n)$ are around 60 – 70% of the direct causal strength $I(MAE_{27}^{n+1}; SSN_{27}^n|\Delta\theta^n)$ of SSN_{27}. Furthermore, estimates of the interaction information, given by $I(MAE_{27}^{n+1}; SSN_{27}^n) - I(MAE_{27}^{n+1}; SSN_{27}^n|\Delta\theta^n)$, are merely $\approx 7\%$ of the direct causal influence, as was also shown in the OMNI dataset in Figure 8. This suggests that to a good approximation, the causal influence of the solar activity is decoupled from that of the latitudinal offset. This will aid us in considering the causal influence of lead time in turns with these two variables, simplifying the causal network.

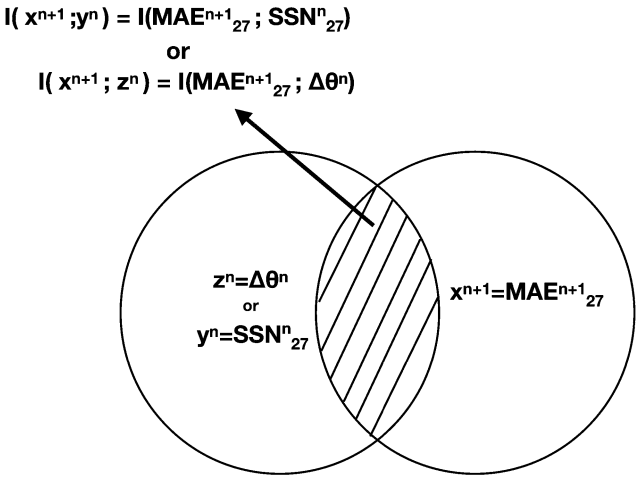

Figure 6 Illustration of the mutual information via Venn diagrams of the conditional entropies $H(x|y)$ or $H(x|z)$. One circle represents the entropy content in the target – forecast accuracy (at time $n+1$), and the second represents that of one driver – either SSN or $\Delta\theta$. The intersection represents the reduction of entropy in the target by knowledge of the driver.

4. Symbolic Representation and Visualisation

4.1. Symbolic Representation: Venn Diagram Visualisation

The information "shared" between the driver variable(s) and the target variable is shown graphically as an Information Venn Diagram in Figure 6 (and 7 for higher-order terms that we will discuss later). The circles represent the conditional entropy $H(x|y)$ of the individual variables, and the intersection (shaded region with lines) represents the reduction in entropy of variable due to the presence of the other, which is the mutual information $I(x;y)$ defined in Equation 1. For our application, one of these variables is the target, and the other a driver; hence the superscripts $n+1$ and n showing different time indices. The labels also show the specific case at hand with the solar-wind variables MAE, SSN, $\Delta\theta$, but we will elaborate on these in upcoming subsections. The drivers that causally influence the target would reduce the entropy, and the extent of this reduction is viewed as the extent of causal influence. On the other hand, if a driver does not have a causal influence, then it does not reduce the entropy, and the mutual information of the target with that driver is zero. Graphically this would mean a separation of the two circles with zero overlap. There are limitations to a formal interpretation of all situations in terms of Venn diagrams – this will become clear for higher-order terms such as interaction information described in Section 3.3 (Ghassami and Kiyavash, 2017). Hence these Venn diagrams serve as a visualisation to build up our intuition rather than be a formal representation.

4.2. Symbolic Representation: Causal Summary Diagrams

The causal information flow between the variables is summarised in Figure 8. The nodes (or ovals) represent the variables, and the arrows represent the flow of information to the target variable, MAE_{27} one time step in the future ($n+1$). Black arrows represent the influence of single driver conditioning out influence of the other drivers. The red segments ending in an

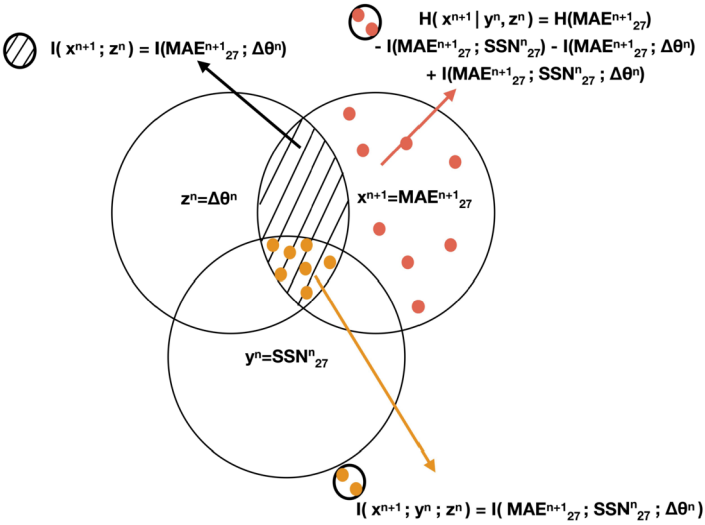

Figure 7 The different information components for three variables. The intersections represent information shared between variables. The *black striped region* is the mutual information shared between target MAE and driver $\Delta\theta$. The *orange dots* denote interaction information: information shared between MAE_{27} and both $\Delta\theta$ and SSN_{27}. The *red circles* show the entropy or uncertainty in MAE_{27} that is not explained or shared by either SSN_{27} or $\Delta\theta$ or their interaction.

arrowhead on the target represents the joint causal influence of the drivers. The confluence of the segments out of the drivers into a point symbolises this joint or combined effect; the arrowhead as usual points to the information flow into the target. This represents the component of influence that is driven by the combination of drivers together, distinct from their individual, direct influences on the target, shown by black arrows. This combined or joint effect could be a positive one showing a redundancy in the driver or that one driver partially or entirely captures the influence due to another. It could be negative, suggesting that one driver induces an influence from the other driver. These can be mathematically quantified in terms of the interaction information. Here for 27-day co-rotation forecasts using only OMNI data, the only drivers are $\Delta\theta$ and SSN_{27}, now considered simultaneously. (As we will see in an upcoming section, the lead time Δt – related to the longitudinal separation – will have a role to play for STEREO data.) We find that the direct influences of SSN_{27} and $\Delta\theta$ (black arrows) are more important than the joint influence (red arrow) on MAE_{27}. In general, we can compute the joint influence due to multiple drivers starting from pairs (the red arrows) to the joint influence of all n drivers simultaneously. However, to use the full general mathematical framework in van Leeuwen et al. (2021) is computationally expensive and complex. It is also not essential in our work here to get the higher-order dependencies. We compute the causal strengths (defined earlier) from the mutual and conditional mutual information terms in accordance with van Leeuwen et al. (2021). The black arrows are given by

$$(\Delta\theta^n \to MAE_{27}^{n+1})_{1\text{link}} = I(MAE_{27}^{n+1}; \Delta\theta^n | SSN_{27}^n), \tag{5}$$

$$(SSN_{27}^n \to MAE_{27}^{n+1})_{1\text{link}} = I(MAE_{27}^{n+1}; SSN_{27}^n | \Delta\theta^n). \tag{6}$$

The red arrow symbolising the joint influence of $\Delta\theta$ and SSN_{27} represents and is related to the interaction information shown in Figure 7. Graphically this represents the intersection

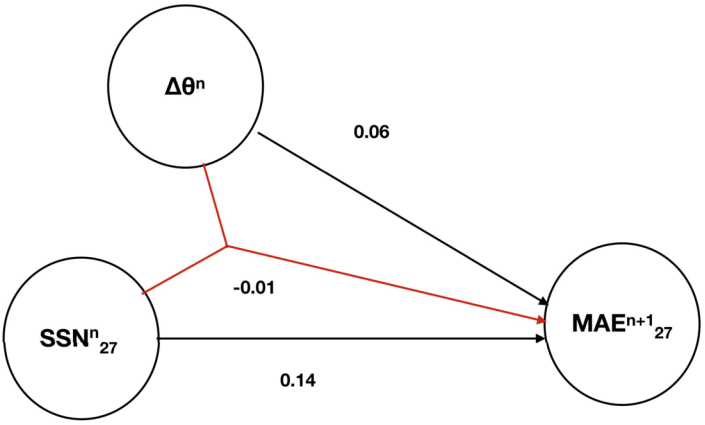

Figure 8 Summary of mean MAE_{27} dependence on latitudinal offset $\Delta\theta$ and SSN_{27} individually (*black*) and in combination (*red*) for OMNI–OMNI. This is done in terms of the causal strength ($cs = \frac{\text{Information Term}}{\text{Entropy}}$) values defined earlier – the numbers attached to the arrows. The *superscripts* merely indicate the formal need for the causes SSN^n_{27}, $\Delta\theta^n$ to precede the effect MAE^{n+1}_{27} – in practice for this application, a single time-step makes negligible difference.

of the information component common to each of the three variables in our triplet, i.e. two drivers (latitudinal offset and smoothed sunspot number) and target MAE_{27}. This is therefore symmetric and is, theoretically, independent of the variable that it is conditioned on. However, when estimating from measured quantities, this symmetry, indicated both graphically in Figure 7 and in Equation 4, is not strictly adhered to. Hence we can express the joint influence indicated by the red arrow as the average of the two equivalent ways of estimating it as

$$\begin{aligned}(\Delta\theta^n \rightarrow MAE^{n+1}_{27})_{\text{2link}} + (SSN^n_{27} \rightarrow MAE^{n+1}_{27})_{\text{2link}} \\ = \quad 1/2\,[\; I(MAE^{n+1}_{27}; \Delta\theta^n) - I(MAE^{n+1}_{27}; \Delta\theta^n | SSN^n_{27}) \\ + \quad I(MAE^{n+1}_{27}; SSN^n_{27}) - I(MAE^{n+1}_{27}; SSN^n_{27} | \Delta\theta^n)]. \end{aligned} \tag{7}$$

5. STEREO: Effect of Lead Time

As explained in the previous section, the OMNI (27-day) co-rotation forecast dataset allows us to focus on the causal influence of $\Delta\theta$ and SSN_{27} as proxy of the solar activity on MAE_{27}. Having learnt that the interaction information between these three driver variables is small, we can assume their influence to be largely independent. We will now proceed to pair $\Delta\theta$ and SSN_{27} by turns, with the lead time Δt. This will give us the direct and interaction terms for each case, analogously to the causal network in Figure 8.

5.1. Conditional Causal Influence: Lead Time

Analogously to the causal analysis of the OMNI time series, we estimate causal linkages using the STA–OMNI co-rotation time series. We begin with the time series of Δt and

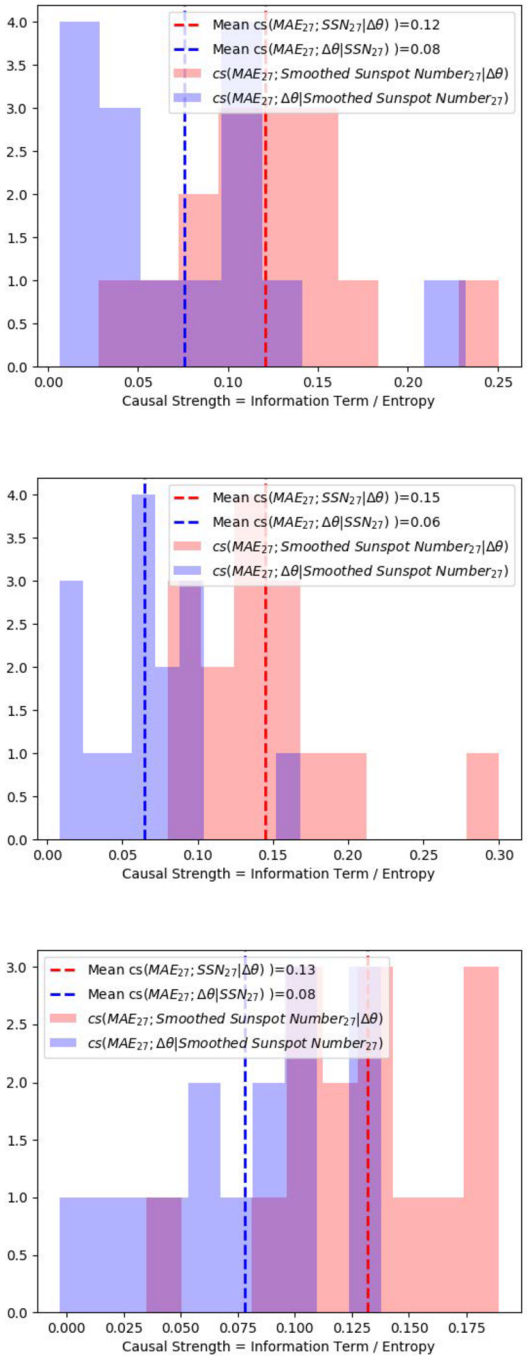

Figure 9 MAE_{27} dependence on principle drivers for co-rotation forecasts; latitudinal offset $\Delta\theta$ conditioned on smoothed sunspot number or SSN_{27} and vice versa in pairs of *Top:* STB–STA, *Middle:* STB–OMNI, and *Bottom:* OMNI–STA.

$\Delta\theta$ as drivers along with the MAE_{27} as the target. We compute causal-strength terms corresponding to mutual information terms $I(MAE_{27}^{n+1}; \Delta\theta^n)$ and $I(MAE_{27}^{n+1}; \Delta t^n)$ and the conditional information terms $I(MAE_{27}^{n+1}; \Delta\theta^n|\Delta t^n)$ and $I(MAE_{27}^{n+1}; \Delta t^n|\Delta\theta^n)$.

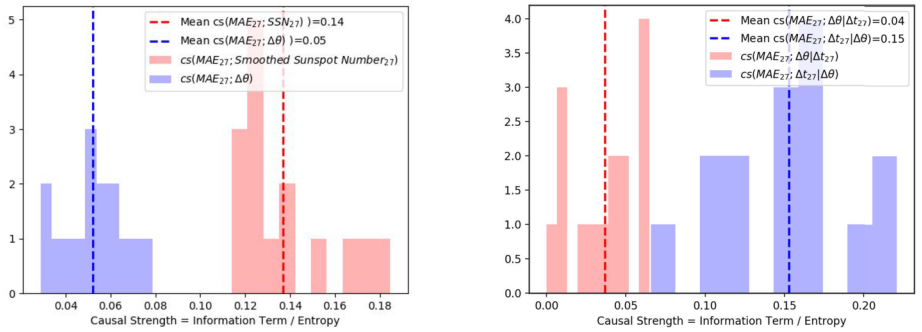

Figure 10 Histogram of causal strengths [cs] of drivers – the lead time, Δt_{27}, and $\Delta\theta$ – with the target, MAE_{27} using STA–OMNI data. *Left:* the distributions of pairwise strengths [MI/H]. *Right:* the conditional causal strengths [CMI/H].

$I(MAE_{27}^{n+1}; \Delta\theta^n; \Delta t^n)$ is in fact the influence of $\Delta\theta$ on MAE_{27} that is shared by (or mediated through / induced by) Δt, in general. From Figure 10, it is evident from the histograms of $cs(MAE_{27}^{n+1}; \Delta\theta^n)$ and $cs(MAE_{27}^{n+1}; \Delta\theta^n|\Delta t^n)$ (or, equivalently, the corresponding information terms) that the interaction-information $I(MAE_{27}^{n+1}; \Delta\theta^n; \Delta t^n)$ is positive. Using the mean values of each term in the information triplet, direct (i.e. conditional mutual) and mutual information for (MAE_{27},$\Delta\theta$,Δt), we get an interaction information term $\approx 20\%$. The direct influence of Δt on MAE_{27} is $\approx 50\%$, and the direct influence of $\Delta\theta$ is the remaining $\approx 30\%$. These numbers are the relative proportions of influence of the two drivers: considered $\Delta\theta$ and Δt, without considering SSN_{27}. The "missing" 20% is likely to have a significant contribution from the other driver, SSN_{27}. However, we also have noise, which includes the statistical fluctuations. Also, the averaged sunspot number is used as a proxy for the temporally evolving nature of the solar wind; it is not an exact proxy.

Histograms with 15 bins, as before, are used to estimate the distribution of these causal-strength terms and shown in Figure 10. It is clear that the lead time Δt has an influence on the MAE_{27}. Upon conditioning on $\Delta\theta$, we find a sizable direct influence of Δt on MAE_{27}, around 1.5 times greater than the direct influence of $\Delta\theta$; this is shown by the black arrows in the summary diagram in Figure 11.

Next, we look at the driver pair of lead time and (smoothed or rolling 27-day average) sunspot number [Δt, SSN_{27}]. We again estimate the pairwise (mutual information) and direct dependencies (conditional mutual information) of Δt and SSN_{27} on MAE_{27}. Once again, the two drivers SSN_{27} and Δt have shared dependencies. This is summarised in Figure 11.

The black arrows are given by

$$(\Delta t^n \to MAE_{27}^{n+1})_{1\text{link}} = I(MAE_{27}^{n+1}; \Delta t^n|SSN_{27}^n) \tag{8}$$

and

$$(SSN_{27}^n \to MAE^{n+1})_{1\text{link}} = I(MAE^{n+1}; SSN_{27}^n|\Delta t^n). \tag{9}$$

The red arrows symbolising the joint influence can be written symmetrically as the sum of

$$\begin{aligned}&(\Delta t^n \to MAE_{27}^{n+1})_{2\text{link}} + (SSN_{27}^n \to MAE_{27}^{n+1})_{2\text{link}}\\ &= \quad 1/2\,[A\ I(MAE_{27}^{n+1}; \Delta t^n) - I(MAE_{27}^{n+1}; \Delta t^n|SSN_{27}^n)\end{aligned}$$

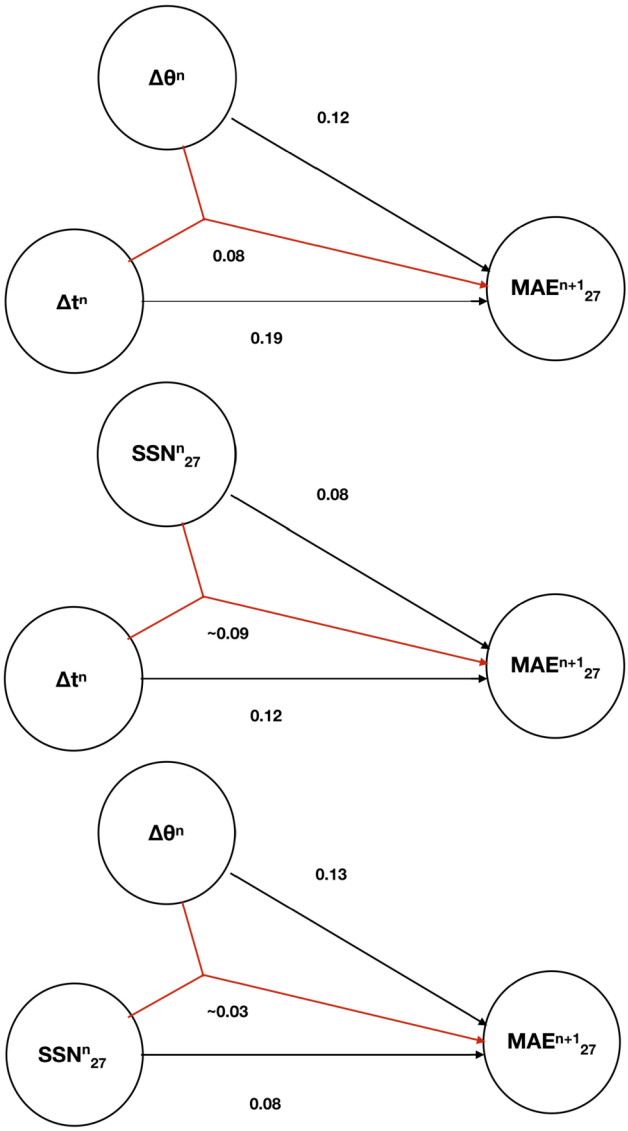

Figure 11 Summary of the mean causal dependence of forecast accuracy MAE_{27} on latitudinal offset $\Delta\theta$, lead time Δt, and average/smoothed sunspot number SSN_{27} individually (*black*) and in pairwise combinations (*red*) for STA–OMNI.

$$+ \quad I(MAE_{27}^{n+1}; SSN_{27}^{n}) - I(MAE_{27}^{n+1}; SSN_{27}^{n} | \Delta t^{n})]. \tag{10}$$

The quantitative analysis is summarised in the causal-summary diagram in Figure 11. The direct causal influence of Δt and SSN_{27} are $\approx 41\%$ and $\approx 28\%$, respectively, in relative terms. The joint influence is $\approx 31\%$.

5.2. Dependencies of the Driver Triplet

Here we put aside the target variable MAE_{27} and apply the causal measures to explore the dependencies between the drivers themselves. The goal is to directly probe the statistical (in)dependence of the drivers without any lag, i.e. we do not seek causal-information flow

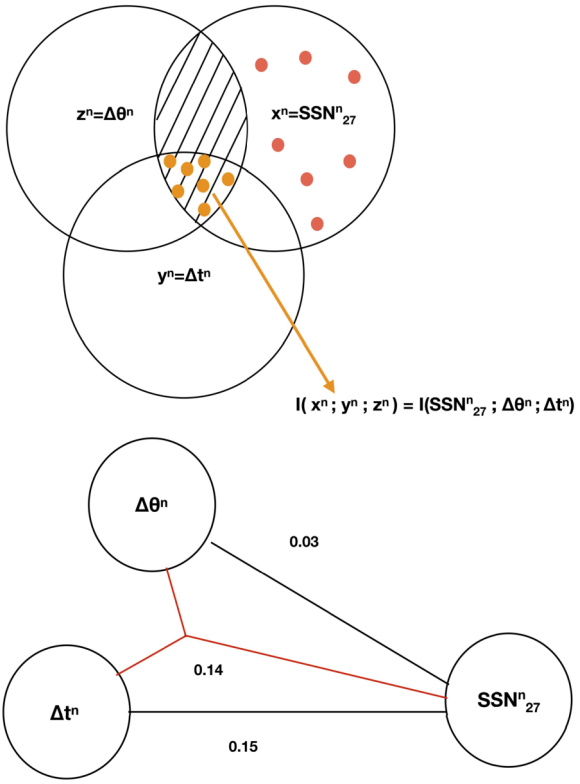

Figure 12 Summary of the mean (in)dependence of the smoothed sunspot number SSN_{27} on the latitudinal offset $\Delta\theta$ and co-rotation time Δt individually (*black*) and in combination (*red*) for STB–OMNI. Here we test the interdependence of the drivers and not a relation between causes preceding the effect symbolically shown by the absence of the arrow heads; hence each is at n.

in time, from one variable to another. Instead, we look simply for information overlap that tells us the how the drivers relate to each other. Hence the three variables are SSN_{27}^n, $\Delta\theta^n$, and Δt^n. These dependencies between drivers are symbolised by line segments instead of arrows. So with no natural target variable, we treat SSN_{27} effectively as the target. The reason for this is that a priori we might expect a stronger dependence between $\Delta\theta^n$ and Δt^n, and hence we wish to see how these two affect SSN_{27}. Our approach is reflected in the causal diagram shown in Figure 12. We find that the direct effect (black line segment) of Δt on SSN_{27} is greater than that of $\Delta\theta$ by over an order of magnitude. This is because solar activity does depend upon the phase and hence the lead time across cycles. Although the joint effect (red segment) of Δt and $\Delta\theta$ on SSN_{27} is lower than the direct effect of Δt, it is still non-trivial. As the corresponding interaction information term is positive, it suggests that Δt mediates the dependency on $\Delta\theta$.

6. Conclusions

In this article, we probe what drives the accuracy of co-rotation forecasts of the solar-wind observations. As we do not have means to make interventional experiments, we apply causal inference methods to the available observations. The causal drivers of relevance are the latitudinal separation or offset $\Delta\theta$ between the observing and forecast locations, the forecast lead time Δt, and the solar activity, as measured by the 27-day rolling average of the sunspot number SSN_{27}. The target variable is the forecast accuracy measured in terms of the mean

absolute error between observation and forecast. We use information-theoretic measures to estimate the strength of the causal influence of the drivers on the target. These do not merely estimate correlations between pairs of variables, but can disentangle the influence of a third variable via conditioning the information content shared between the three of them. Depending on the different information terms or components, we can estimate the influence of the individual drivers as well as their joint effect, induced effects, and redundancy due to correlation amongst drivers.

We draw the following conclusions:

i) The decomposition of information flow between the different drivers (or causes) and the target is effective in identifying cause and effect relationships driving dynamical systems in the presence of complex, nonlinear relationships between multiple variables. This approach is perfectly suited to trace what drives the co-rotation forecast accuracy or rather the uncertainty.

ii) The pairwise causal relationship between drivers – latitudinal offset $\Delta\theta$, forecast lead time Δt, and the average sunspot number SSN_{27} – and target MAE_{27} given by mutual information normalised to the target entropy confirms our understanding from previous results (e.g. Turner et al., 2021) that solar-activity levels measured via SSN_{27} and the lead time Δt have a big influence on the average forecast error MAE_{27} followed by the latitudinal offset $\Delta\theta$. Statistical noise levels impact the absolute values of the dependencies. The relative values of the causal strengths, however, teach us about the hierarchy of influence of the different driver combinations.

iii) Exploiting higher-order measures such as interaction information, we can probe deeper into causal influence of multiple drivers in conjunction with one another. A non-zero interaction information has two possible cases with corresponding interpretations. Negative values show an induced effect (i.e. one driver showing a coupling to the target induced only due to the presence and influence of another driver), whereas positive values show a redundant / shared influence. We can be quantitative about the relative importance of joint causal (positive or negative) influence and therefore potentially impact forecasting. Also qualitatively, knowledge of whether drivers act independently of one another can help design future data experiments. For instance, from Figure 9, it is clear from the consistency of MAE_{27} dependence on SSN_{27} and Latitudinal Offset – from OMNI alone and OMNI with STEREO – that solar activity has a stronger, direct influence on MAE_{27} than latitudinal offset, independent of lead time. Hence the appropriate weight can be given to each of these drivers in an assimilation forecast. On the other hand, from Figure 12, while solar activity and latitudinal offset have a weaker direct association (black line between them) relative to its stronger, direct coupling with lead time, they do have an indirect coupling. The association of sunspot number with latitudinal offset is owed predominantly to the lead time. This implies that the phase of the solar cycle is important.

iv) The interaction-information terms, such as $I(MAE_{27}^{n+1}; \Delta\theta^n; \Delta t^n)$ and $I(MAE_{27}^{n+1}; SSN_{27}^n; \Delta\theta^n)$, or, equivalently, the corresponding causal-strength terms I/H quantify the information content shared between the three variables. From these terms, we can learn that the SSN_{27} and $\Delta\theta$ share very little information. That is, their influence on MAE_{27} is predominantly independent of one another. On the other hand, for the STEREO dataset, $\Delta\theta$ and Δt share a non-trivial fraction ($\approx 20\%$) of the their total information content (or influence on MAE_{27}). In this case, the positive sign of the interaction information indicates that $\Delta\theta$ partially contributes to the influence of Δt on MAE_{27}, and vice versa. These effects could not be revealed by standard correlation analysis.

A causal inference approach disentangles the drivers of the forecast accuracy in ways that standard statistical analysis or data assimilation cannot. Rather, the latter two can be improved through causal diagnostics. It allows us to not only disentangle individual sources of uncertainty in the forecast, but also calculate partial and complete redundancies in drivers of this uncertainty. This learning can potentially be applied to improve the solar-wind data-assimilation forecasts.

Acknowledgements We have benefited from sunspot data provided by the Royal Observatory of Belgium SILSO, RGO/SOON sunspot-latitude and -area data collated by David Hathaway and Lisa Upton, and OMNI data provided by NASA/SPDF. We benefited from useful discussions as part of the International Space Science Institute (ISSI, Bern) team "Magnetic open flux and solar-wind structuring in interplanetary space" (2019 – 2021) led by Manuela Temmer.

Author contributions H. Turner was responsible for the analysis of the spacecraft data. N. Chakraborty did the causal analyses and along with M. Owens determined the appropriate tests and methods that were needed. All authors contributed to the design of the study and the writing of the manuscript.

Funding This work was part-funded by Science and Technology Facilities Council (STFC) grant number ST/V000497/1.

Data Availability OMNI data are available from omniweb.gsfc81.nasa.gov/. Sunspot data are provided by the Royal Observatory of Belgium SILSO and available from www.sidc.be/silso/DATA/SN_m_tot_V2.0.csv. All analysis and visualisation code is packaged with all the required data here: github.com/University-of-Reading-Space-Science/SSNfromOSF.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Amblard, P., Michel, O.J.J.: 2009, Measuring information flow in networks of stochastic processes. CoRR. arXiv.

Bartels, J.: 1934, Twenty-seven day recurrences in terrestrial-magnetic and solar activity, 1923 – 1933. *Terr. Magnet. Atmosph. Elect. (J. Geophys. Res.)* **39**, 201. DOI.

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., Greene, G., Hapgood, M., Horne, R.B., Jackson, D.: 2013, *Extreme Space Weather: Impacts on Engineered Systems and Infrastructure*, Royal Academy of Engineering, London ISBN 1-903496-95-0.

Chakraborty, N., van Leeuwen, P.J.: 2022, Using mutual information to measure time lags from nonlinear processes in astronomy. *Phys. Rev. Res.* **4**, 013036. DOI.

Clette, F., Lefèvre, L.: 2016, The new sunspot number: assembling all corrections. *Solar Phys.* **291**, 2629. DOI.

Clette, F., Lefèvre, L., Chatzistergos, T., Hayakawa, H., Carrasco, V.M.S., Arlt, R., Cliver, E.W., Dudok de Wit, T., Friedli, T.K., Karachik, N., Kopp, G., Lockwood, M., Mathieu, S., Muñoz-Jaramillo, A., Owens, M., Pesnell, D., Pevtsov, A., Svalgaard, L., Usoskin, I.G., van Driel-Gesztelyi, L., Vaquero, J.M.: 2023, Recalibration of the sunspot-number: status report. *Solar Phys.* **298**, 44. DOI.

Ghassami, A., Kiyavash, N.: 2017, Interaction Information for Causal Inference: the Case of Directed Triangle. arXiv.
Granger, C.W.J.: 1969, Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424.
Heckerman, D.: 2020, A Tutorial on Learning with Bayesian Networks. arXiv.
Kaiser, M.L.: 2005, The STEREO mission: an overview. *Adv. Space Res.* **36**, 1483. DOI.
King, J.H., Papitashvili, N.E.: 2005, Solar wind spatial scales in and comparisons of hourly wind and ACE plasma and magnetic field data. *J. Geophys. Res.* **110**. DOI.
Kohutova, P., Bocquet, F.-X., Henley, E.M., Owens, M.J.: 2016, Improving solar wind persistence forecasts: removing transient space weather events, and using observations away from the Sun-Earth line. *Space Weather* **14**, 802. DOI.
Kraft, S., Puschmann, K.G., Luntama, J.P.: 2017, Remote sensing optical instrumentation for enhanced space weather monitoring from the L1 and L5 Lagrange points. In: Cugny, B., Karafolas, N., Sodnik, Z. (eds.) *International Conference on Space Optics — ICSO 2016* **10562**, SPIE, Bellingham, 115. DOI.
Kraskov, A., Stögbauer, H., Grassberger, P.: 2004, Estimating mutual information. *Phys. Rev. E* **69**. DOI.
Lang, M., Owens, M.J.: 2019, A variational approach to data assimilation in the solar wind. *Space Weather* **17**, 59. DOI.
Lang, M., Witherington, J., Turner, H., Owens, M.J., Riley, P.: 2021, Improving solar wind forecasting using data assimilation. *Space Weather* **19**, e2020SW002698. DOI.
Luhmann, J.G., Soloman, S.C., Linker, J.A., Lyon, J.G., Mikic, Z., Odstrcil, D., Wang, W., Wiltberger, M.: 2004, Coupled model simulation of a Sun-to-Earth space weather event. *J. Atmos. Solar-Terr. Phys.* **66**, 1243.
McComas, D.J., Elliott, H.A., Schwadron, N.A., Gosling, J.T., Skoug, R.M., Goldstein, B.E.: 2003, The three-dimensional solar wind around solar maximum. *Geophys. Res. Lett.* **30**. DOI.
Merkin, V.G., Owens, M.J., Spence, H.E., Hughes, W.J., Quinn, J.M.: 2007, Predicting magnetospheric dynamics with a coupled Sun-to-Earth model: challenges and first results. *Space Weather* **5**, 1. DOI.
Owens, M.J., Challen, R., Methven, J., Henley, E., Jackson, D.R.: 2013, A 27 day persistence model of near-Earth solar wind conditions: a long lead-time forecast and a benchmark for dynamical models. *Space Weather* **11**, 225. DOI.
Owens, M.J., Riley, P., Lang, M., Lockwood, M.: 2019, Near-Earth solar wind forecasting using corotation from L5: the error introduced by heliographic latitude offset. *Space Weather* **17**, 1105. DOI.
Owens, M.J., Lang, M., Riley, P., Lockwood, M., Lawless, A.S.: 2020, Quantifying the latitudinal representivity of in situ solar wind observations. *J. Space Weather Space Clim.* **10**, 8. DOI.
Owens, M.J., Chakraborty, N., Turner, H., Lang, M., Riley, P., Lockwood, M., Barnard, L.A., Chi, Y.: 2022, Rate of change of large-scale solar-wind structure. *Solar Phys.* **297**, 83. DOI.
Pearl, J.: 2000, *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge UK.
Runge, J.: 2015, Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E* **92**, 062829.
Runge, J.: 2018, Causal network reconstruction from time series: from theoretical assumptions to practical estimation. *Chaos Interdiscip. J. Nonlinear Sci.* **28**, 075310.
Schreiber, T.: 2000, Measuring information transfer. *Phys. Rev. Lett.* **85**, 461.
Simunac, K.D.C., Kistler, L.M., Galvin, A.B., Popecki, M.A., Farrugia, C.J.: 2009, In situ observations from STEREO/PLASTIC: a test for L5 space weather monitors. *Ann. Geophys.* **27**, 3805. DOI.
Thomas, S.R., Fazakerley, A., Wicks, R.T., Green, L.: 2018, Evaluating the skill of forecasts of the near-Earth solar wind using a space weather monitor at L5. *Space Weather* **16**, 814. DOI.
Toth, G., Sokolov, I.V., Gombosi, T.I., Chesney, D.R., Clauer, C.R., De Zeeuw, D.L., Hansen, K.C., Kane, K.J., Manchester, W.B., Oehmke, R.C., Powell, K.G., Ridley, A.J., Roussev, I.I., Stout, Q.F., Volberg, O., Wolf, R.A., Sazykin, S., Chan, A., Yu, B., Kóta, J.: 2005, Space weather modeling framework: a new tool for the space science community. *J. Geophys. Res.* **110**, A12226. DOI.
Turner, H., Owens, M.J., Lang, M.S., Gonzi, S.: 2021, The influence of spacecraft latitudinal offset on the accuracy of corotation forecasts. *Space Weather* **19**, e2021SW002802. DOI.
van Leeuwen, P.J., DeCaria, M., Chakaborty, N., Pulido, M.: 2021, A framework for causal discovery in non-intervenable systems. arXiv.
Verscharen, D., Klein, K.G., Maruca, B.A.: 2019, The multi-scale nature of the solar wind. *Liv. Rev. Solar Phys.* **16**, 5. DOI.
Williams, P.L., Beer, R.D.: 2010, Nonnegative decomposition of multivariate information. arXiv.
Yashiro, S., Gopalswamy, N., Michalek, G., St Cyr, O.C., Plunkett, S.P., Rich, N.B., Howard, R.A.: 2004, A catalog of white light coronal mass ejections observed by the SOHO spacecraft. *J. Geophys. Res.* **109**. DOI.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.