The Astrophysical Journal, 979:88 (19pp), 2025 January 20 https://doi.org/10.3847/1538-4357/ad99db

Persistence and Burn-in in Solar Coronal Magnetic Field Simulations

Eric J. Hall[1], Karen A. Meyer[1], and Anthony R. Yeates[2]

[1] Division of Mathematics, School of Science and Engineering, University of Dundee, Dundee, DD1 4HN, UK

[2] Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK

Received 2024 September 17; revised 2024 November 12; accepted 2024 December 1; published 2025 January 20

Abstract

Simulations of solar phenomena play a vital role in space-weather prediction. A critical computational question for automating research workflows in the context of data-driven solar coronal magnetic field simulations is quantifying a simulation's burn-in time, after which a solar quantity has evolved away from an arbitrary initial condition to a physically more realistic state. A challenge to quantifying simulation burn-in is that the underlying solar processes and data, like many physical phenomena, are non-Markovian and exhibit long memory or persistence and, therefore, their analysis evades standard statistical approaches. In this work, we provide evidence of long memory in the nonperiodic variations of solar quantities (including over timescales significantly shorter than previously identified) and demonstrate that magnetofrictional simulations capture the memory structure present in magnetogram data. We also provide an algorithm for the quantitative assessment of simulation burn-in time that can be applied to nonstationary time series with long memory. Our approach is based on time-delayed mutual information, an information-theoretic quantity, and includes a small-sample bias correction.

Unified Astronomy Thesaurus concepts: Solar magnetic fields (1503); Solar physics (1476); Time series analysis (1916); Bootstrap (1906); Magnetohydrodynamical simulations (1966)

1. Introduction

Understanding the evolution of solar active regions is vital for space weather prediction. The effects of severe space weather are highlighted in the current UK Risk Register, with reasonable worst-case scenarios resulting in significant primary impacts that include disruptions to power systems, satellite navigation, telecommunications, aviation, and ground-based digital components; increased background radiation doses at high altitudes and in space; and increased risk of on-orbit collisions of tracked objects.[3] Simulations of the solar coronal magnetic field play an essential role in mitigating these risks by supporting decision-making; for example, by following the formation and eruption of magnetic flux ropes, which have been linked to filament eruptions and coronal mass ejections from the Sun (e.g., G. P. S. Gibb et al. 2014; S. L. Yardley et al. 2018), and by defining eruptivity criteria based on coronal magnetic field evolution (e.g., F. P. Zuccarello et al. 2015, 2018; P. Pagano et al. 2019; O. E. K. Rice & A. R. Yeates 2023) or as boundary conditions for space weather modeling (e.g., D. Rodkin et al. 2017; P. Pagano et al. 2018; S. Gonzi et al. 2021). Since it is currently difficult to measure the coronal magnetic field directly, assumptions must be made about the initial state of the corona in such simulations. In many cases, it is assumed that there is an initial ramp-up or burn-in phase, before a simulation evolves away from its (arbitrary) initial condition to a more realistic, self-consistent state. This assumed burn-in time varies from on the order of 1 hr, for small-scale, quiet-Sun (QS) simulations (K. A. Meyer 2013), to one or more days in active-region-scale simulations (P. Pagano et al. 2019) and to weeks to months in global-scale simulations (P. Bhowmik et al. 2022). Typically, the burn-in time is estimated by the user, based on prior experience and knowledge of the simulation method, rather than algorithmically.

The present work develops a systematic approach to quantifying the simulation burn-in for a particular family of data-driven coronal magnetic field simulations in solar physics, where the user is interested in the burn-in time following the initial condition or indeed the timescale after which the simulation state at an arbitrary time, t, has been "forgotten." The simulations considered here use a magnetofrictional relaxation method to evolve the coronal magnetic field through a series of nonlinear force-free equilibria in response to photospheric motions based on observed magnetogram data (e.g., A. A. van Ballegooijen 2000; D. H. Mackay 2011; K. A. Meyer 2013; D. J. Price et al. 2019; J. T. Hoeksema et al. 2020; K. Barczynski et al. 2022; A. R. Yeates & P. Bhowmik 2022). The method is computationally inexpensive compared to magnetohydrodynamic (MHD) simulations, which means that large numbers of simulations can be run relatively quickly, e.g., to investigate the coronal magnetic field evolution of many observed magnetogram series (S. L. Yardley et al. 2021), to explore the effects of different terms in the evolution equations (D. H. Mackay & A. R. Yeates 2021; D. H. Mackay & L. A. Upton 2022), or to generate training and validation sets for machine learning methods. Based on observed magnetogram data, we consider localized simulations of both the QS and solar active regions. We also consider cycle-length or longer global coronal simulations (A. R. Yeates 2014; V. Aslanyan et al. 2024), which would be prohibitively expensive with MHD methods. All these simulations are time-dependent, enabling us to follow the continuous evolution of scalar-valued quantities (i.e., time series) derived from the 3D magnetic field. Simulation-based quantities, such as the total magnetic energy and electric current, help one to understand the evolution of an active region (G. P. S. Gibb et al. 2014;

[3] National Risk Register: https://www.gov.uk/government/publications/national-risk-register-2023.

D. H. Mackay & A. R. Yeates 2021) and to identify possible signatures of eruptive behavior (P. Pagano et al. 2019; P. Bhowmik et al. 2022; O. E. K. Rice & A. R. Yeates 2022). We also consider quantities that can be calculated directly from observational magnetogram data, such as the total absolute flux through the photosphere.

Our algorithmic approach is informed by the presence of persistence or long memory, which we identify in all time series of the solar quantities considered. Time series with long memory are characterized by slowly decaying or persistent autocorrelations; that is, observed values at distant time points are correlated (J. Beran et al. 2013). This global statistical dependence indicates self-similar activity over the timescale of interest and is an important feature of the data-generating process for predictive modeling. Persistence is ubiquitous in complex systems across the physical sciences (e.g., see the surveys by S. Panchev & M. Tsekov 2007 and S. Salcedo-Sanz & J. Del Ser 2022). Quantifying persistence is vital for various tasks, including forecasting (M. G. Ogurtsov 2004) and predictive modeling (F. Maddanu & T. Proietti 2022), analyzing sensitivity, and correcting statistical estimators. Fundamentally, this knowledge elucidates the underlying system's dynamics, thereby contributing scientific insights about observations and simulations of the Sun.

The analysis of persistence is of longstanding interest in solar physics (B. B. Mandelbrot & J. R. Wallis 1969a; A. Ruzmaikin et al. 1994; R. W. Komm 1995; F. Lepreti et al. 2000; M. G. Ogurtsov 2004), with a recent resurgence from considering the effects of the solar cycle on climate change (K. Rypdal & M. Rypdal 2012; M. Rypdal & K. Rypdal 2012). These studies find evidence of long memory in solar activity using various proxy data sets spanning different timescales, from as short as 20 days up to a few thousand years. The seminal work of B. B. Mandelbrot & J. R. Wallis (1969a) identifies long memory in solar activity based on mean monthly sunspot numbers over the period 1749–1948 using a rescaled-range (Hurst) analysis. A. Ruzmaikin et al. (1994) and F. Lepreti et al. (2021) use cosmogenic radionuclide data as a proxy for solar activity and find evidence of long memory over longer timescales (periods between 100 and 3000 yr), while R. W. Komm (1995), using Mount Wilson differential rotation measurements, and F. Lepreti et al. (2000), using the daily averaged intensity of optical flares, find evidence of long memory in data over shorter timescales (20 days to several years). Related work, in M. Adams et al. (1997), studies the complexity of synthetic and observational magnetograms for solar active regions to identify possible signatures for flare activity using fractal dimension, a measure of complexity connected to measures of persistence. A physical mechanism for persistence in solar records, connected to the random variations in the solar dynamo, is proposed in A. Ruzmaikin et al. (1994), and observational evidence consistent with this mechanism is identified in A. A. Ruzmaikin et al. (2000).

From a simulation standpoint, the existence of long memory suggests specific tools to automate decisions regarding the quality of simulated data. We propose, in Section 4, an algorithm for quantitatively determining when outputs from a magnetofrictional simulation have reasonably “forgotten” the simulation's initial potential field state. This algorithm is based on information-theoretic measures of dependence, as opposed to variance-based measures, to remain applicable in the presence of long memory and nonstationarity. Such algorithms can also be used to automate other research workflows, such as data subsetting for training/testing physics-informed neural networks and other domain-aware statistical surrogate models, an area of growing interest in the solar physics community (see, e.g., A. Asensio Ramos et al. 2017, 2023; S. Rahman et al. 2023). Moreover, the burn-in algorithm might apply to other nonequilibrium systems (arising from long memory or other sources of nonstationarity) where quantifying the transitions between states is of interest.

The statistical methods utilized in our persistence analysis rely both on standard time-series methods, which are well established in the solar physics community (P. Song & C. T. Russell 1999), and more specialized methods for long memory (J. Beran 1994; J. Beran et al. 2013). The typical approach to quantifying persistence is to propose a candidate long-memory model for the data-generating process and then to fit model parameters that capture the strength of the long-range dependence. The most common candidate in the geophysical community is the fractional Gaussian noise (FGN) model, obtained by fractional differencing of a Brownian motion and then discretizing. The FGN model has a memory parameter H, often referred to as the Hurst exponent, which can be estimated, e.g., using rescaled-range analysis (B. B. Mandelbrot & J. R. Wallis 1969b; A. W. Lo 1991),[4] or detrended fluctuation analysis (DFA; C. K. Peng et al. 1994).[5] An alternative class of long-memory models, more common in statistics and econometrics, involve the fractionally integrated process of C. W. J. Granger & R. Joyeux (1980) and J. R. M. Hosking (1981), which is obtained by fractional integration of a discretized Brownian motion. The parameter of interest is the fractional integration order d, which can be estimated using log-periodogram regression methods (J. Geweke & S. Porter-Hudak 1983; P. M. Robinson 1995a) and likelihood methods (H. R. Kuensch 1987; P. M. Robinson 1995a; K. Shimotsu & P. C. B. Phillips 2005; K. Shimotsu 2010). The two memory parameters are related by the linear relationship $H = d + 0.5$ (see J. Geweke & S. Porter-Hudak 1983),[6] which has been shown to hold empirically in stationary regimes (L. Ding et al. 2021). Here, we favor the fractionally integrated assumption, as our data are available at regular intervals, the model is easily extended to explicitly include short-term memory (a primary objection to the FGN model), and robust estimators are available in nonstationary regimes.

The remainder of the paper is outlined as follows. In the next section, we review the magnetofrictional simulations and quantities that will be analyzed. In Section 3, we recall long-memory models and associated inference methods. We utilize these models and methods to quantify the persistence in time series of solar quantities for the QS, solar active regions, and a global simulation. Based on the presence of long memory, we develop an algorithmic approach to quantifying simulation burn-in based on the time-delayed mutual information (TDMI) in Section 4. The key burn-in time calculation is provided in Algorithm 2, together with a small-sample bias correction in Algorithm 1. Additional statistical background and a list of the softwares utilized are provided in Appendices A–D.

[4] The form presented in B. B. Mandelbrot & J. R. Wallis (1969b) is a corrected version of the rescaled-range analysis initially introduced by H. E. Hurst (1951).

[5] Although we caution that DFA introduces uncontrolled bias and is inappropriate for nonstationary processes (R. M. Bryce & K. B. Sprague 2012).

[6] Theoretically, the spectral density of an FGN process with $H \in (0, 1)$ is related to that of a fractionally integrated process by $H = d + 0.5$, with $d \in (-0.5, 0.5)$.

2. Magnetofrictional Simulations

Magnetofrictional methods have been used extensively in data-driven simulations of the Sun's coronal magnetic field, from shorter-term, localized simulations over hours or days, following regions of the QS (e.g., K. A. Meyer 2013; K. Barczynski et al. 2022; L. R. Bellot Rubio & M. C. M. Cheung 2022) and active regions (e.g., D. H. Mackay 2011; P. Pagano et al. 2019; S. L. Yardley et al. 2021), to long-term simulations of the global corona over months or years (e.g., A. R. Yeates 2008; P. Bhowmik et al. 2022; D. H. Mackay & L. A. Upton 2022; V. Aslanyan et al. 2024). There are several advantages to magnetofrictional methods over other types of solar physics simulations. (i) Observed photospheric magnetograms can be used directly as a lower-boundary condition to drive the evolution of the coronal magnetic field.[7] (ii) Magnetofrictional methods produce a continuous time evolution of the coronal magnetic field in response to photospheric motions, allowing for a memory of connectivity and the buildup of electric currents and free magnetic energy. This is in contrast to models such as potential field or nonlinear force-free field extrapolation methods (see, e.g., the models compared in C. J. Schrijver 2006), which produce time-independent extrapolations of the coronal magnetic field from each magnetogram. (iii) Magnetofrictional simulations are much less computationally expensive to run than MHD simulations, allowing for global coronal simulations over whole solar cycles (A. R. Yeates 2014, 2024) and explorations of the parameter space (O. E. K. Rice & A. R. Yeates 2022, 2023).

The idea of the magnetofrictional method is that the coronal magnetic field, $\boldsymbol{B}$, evolves through a series of quasi-static, nonpotential equilibria in response to a changing (photospheric) lower boundary. In this paper, we will consider two different implementations of the magnetofrictional method. The first is a model in a Cartesian coordinate system, where the lower-boundary evolution is determined by tracking a region of the solar photosphere over time in observed magnetogram data and remapping it to Cartesian coordinates. The second is a global model in a spherical coordinate system, where the lower-boundary evolution is given by a surface flux transport (SFT) model that incorporates active regions determined from observed magnetograms. An SFT model typically includes a source term for newly emerging magnetic flux regions, an advective velocity incorporating the observed large-scale flows of differential rotation and meridional flow, and a diffusive term approximating the effect of supergranulation on magnetic flux dispersal (see, e.g., N. R. Sheeley 2005; A. R. Yeates et al. 2023 for reviews of SFT models). The photospheric lower-boundary evolution for both implementations will be discussed in Section 2.1 and the coronal evolution in Section 2.2.

2.1. Photospheric Observations and Evolution

Four simulations are considered in this paper, covering different spatial and temporal scales. These are summarized in Table 1 (along with a toy model, discussed in Section 2.3). The first three simulations are driven by observed magnetogram regions that have been extracted from full-disk observations of the Sun and remapped to a Cartesian coordinate system. The first simulation is driven by magnetogram observations of a QS region, taken by the Helioseismic and Magnetic Imager (HMI) Instrument (J. Schou et al. 2012) on board the Solar Dynamics Observatory (W. D. Pesnell et al. 2012). A region of size 512 × 512 pixels was extracted from full-disk line-of-sight magnetogram observations between 22:03:05 UTC on 2022 March 16 and 02:57:05 UTC on 2022 March 17. The region was cut out and derotated using the Joint Science Operations Center (JSOC) export tool.[8] The pixel size of HMI magnetograms at this time was approximately 364 km on the solar photosphere and the cadence of the data is 45 s. The QS magnetogram series consists of 393 magnetograms and covers a period of just under 5 hr.

Standard cleaning procedures were applied to prepare the magnetograms for use in simulations (see, e.g., K. A. Meyer 2013; G. P. S. Gibb et al. 2014). To reduce noise and remove 5 minute oscillations, the magnetograms were smoothed in time by averaging using a Gaussian kernel, with $\tau = 2$ the number of frames over which the weighting falls by $1/e$ (see the Appendix of G. P. S. Gibb et al. 2014). The remaining noise in the data set was estimated by fitting a Gaussian to a histogram of pixel values, giving $\sigma_B = 6.9$ G as the Gaussian half-width at half-maximum. Pixels of magnitude less than $2\sigma_B$ were set to zero. An example of a cleaned QS magnetogram can be seen in Figure 1(a).

The second simulation is of NOAA Active Region 10977 (henceforth, AR10977), which has been studied extensively using magnetofrictional simulations (e.g., D. H. Mackay 2011; G. P. S. Gibb et al. 2014; D. H. Mackay & A. R. Yeates 2021). A 127 × 127 pixel region was extracted from magnetogram data observed by the Michelson Doppler Imager (MDI) instrument (P. H. Scherrer et al. 1995) on board the Solar and Heliospheric Observatory (SOHO; V. Domingo et al. 1995), between 12:51:01 UTC on 2007 December 2 and 22:23:01 UTC on 2007 December 10, consisting of 121 magnetograms. The cadence of MDI's magnetograms is 96 minutes and the pixel size is approximately 1386 km. Similar cleaning was applied to the magnetograms, with the additional steps of the magnetograms being corrected for flux imbalance, so that the coronal simulation can be carried out in a closed domain (see, e.g., D. H. Mackay 2011 or G. P. S. Gibb et al. 2014 for full details), and each frame being interpolated to a 256 × 256 pixel grid, to run the coronal simulation at a higher resolution than the original data (new pixel size ≈692 km), to reduce numerical diffusion. An example cleaned magnetogram from AR10977 can be seen in Figure 1(b).

The third simulation is of NOAA AR11680, extracted from HMI magnetogram observations between 12:59:18 UTC on 2013 February 25 and 12:59:18 UTC on 2013 March 2. The observations used were the 720 s HMI magnetograms from JSOC. The region extracted is 512 × 384 pixels in size and consists of 601 frames. The pixel size of the HMI magnetograms at this time was approximately 362 km. Cleaning was applied as described for the QS simulation above. An example cleaned magnetogram from AR11680 can be seen in Figure 1(c).

The fourth simulation models the global solar corona, so cannot directly use magnetogram observations, since these only show the Earth-facing side of the Sun. Instead, an SFT model simulates the evolution of the Sun's global photospheric magnetic field. The initial condition for the SFT is HMI radial-component pole-filled Carrington map 2097

[7] Subject to standard data preparation methods, such as noise subtraction and flux imbalance correction; see Section 2.1.

[8] http://jsoc.stanford.edu

Table 1
Grid, Time Step (For Data Assimilation, i.e., Observed Magnetograms in Cartesian Cases or Active Region Insertion in the Global Case), and Boundary Condition Information for Each Simulation

Simulation	Grid Cells	Boundary Conditions	Magnetogram Frames	Time Step	Spatial Resolution
QS	512 × 512 × 256	Periodic in *x*- and *y*-directions, open top.	393	45 s	364 km
AR10977	256 × 256 × 256	Closed in *x*- and *y*-directions, closed top.	121	96 minutes	692 km
AR11680	512 × 384 × 384	Closed in *x*- and *y*-directions, open top.	601	12 minutes	362 km
Solar Cycle 24	360 × 180 × 60	Periodic in longitudinal direction, closed in latitudinal direction, open at source surface $R_{ss} = 2.5\ R_\odot$.	⋯	1 day[a]	1° (≈1930 km) at $r = R_\odot$, $\theta = 90°$.
Toy model	256 × 256 × 256	Closed in *x*- and *y*-directions, open top.	201	N/A	N/A

Note.
[a] In the global simulation, 1 day is the frequency at which newly emerging active regions may be assimilated. The full 3D coronal magnetic field is output once per 27 days, while quantities such as the total magnetic energy are output with a mean frequency of once per 83.9 s.

(X. Sun 2018), from 2010 May to June. Figure 2 shows this initial map alongside an example map of the photospheric magnetic field from later in the simulation. Newly emerging active regions are determined automatically from the HMI/SHARP database (M. G. Bobra et al. 2014), for inclusion in the SFT simulation. See A. R. Yeates & P. Bhowmik (2022) for full details of the SFT method and flux emergence procedure used (for simplicity here, we neglect any additional twist/helicity in the emerging regions). The simulation covers Solar Cycle 24, over a period of 10 yr, with 1072 emerging regions (run T0 from A. R. Yeates 2024). The quantities in this paper are calculated every 10 simulation time steps, giving an average cadence of 83.9 s. On the other hand, emerging active regions are assimilated only at a cadence of 24 hr, with each region emerging by applying a steady electric field over 24 hr. Therefore, we consider time series downsampled to a cadence of approximately 24 hr in the present analysis.

2.2. Coronal Models

The general equation for the evolution of the coronal magnetic field in both the Cartesian and spherical implementations of the magnetofrictional method is

$$\frac{\partial \boldsymbol{A}}{\partial t} = \boldsymbol{v} \times \boldsymbol{B} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{A}$ is the magnetic vector potential, $\boldsymbol{B} = \nabla \times \boldsymbol{A}$, and $\boldsymbol{N}$ is a nonideal term representing unresolved smaller-scale turbulent motions (A. A. van Ballegooijen 2000). We set $\boldsymbol{N} = \boldsymbol{0}$ in the localized Cartesian model. In the global spherical model, $\boldsymbol{N}$ is modeled by fourth-order hyperdiffusion, which smooths gradients in α, while preserving the magnetic helicity density $\boldsymbol{A} \cdot \boldsymbol{B}$ (A. A. van Ballegooijen & S. R. Cranmer 2008):

$$\boldsymbol{N} = \frac{\boldsymbol{B}}{|\boldsymbol{B}|^2} \nabla \cdot (\eta_h |\boldsymbol{B}|^2 \nabla \alpha), \tag{2}$$

where $\alpha = \boldsymbol{j} \cdot \boldsymbol{B}/|\boldsymbol{B}|^2$ is the twist of the magnetic field with respect to the corresponding potential magnetic field extrapolation, $\boldsymbol{j} = \nabla \times \boldsymbol{B}$ is the electric current density, and $\eta_h = 10^{11}\,\mathrm{km}^4\,\mathrm{s}^{-1}$.

The magnetofrictional velocity is assumed to be proportional to the Lorentz force:

$$\boldsymbol{v} = \frac{\boldsymbol{j} \times \boldsymbol{B}}{\nu B^2} + \boldsymbol{v}_{\mathrm{out}}, \tag{3}$$

where ν is the friction coefficient and $B = |\boldsymbol{B}|$. In the Cartesian model, $\boldsymbol{v}_{\mathrm{out}} = \boldsymbol{0}$ and the friction coefficient is constant, such that $1/\nu = 3000\,\mathrm{km}^2\,\mathrm{s}^{-1}$. In the global model, $\boldsymbol{v}_{\mathrm{out}} = v_0 (r/R_\odot)^{11.5} \boldsymbol{e}_r$, where $v_0 = 100\,\mathrm{km\,s}^{-1}$, $\boldsymbol{e}_r$ is the unit vector in the radial direction, and $R_\odot$ is the solar radius. This represents the effect of the solar wind on the upper corona. In the global model, the friction coefficient varies with magnetic field strength, colatitude, θ, and radius, r, and is given by $\nu = \nu_0 |\boldsymbol{B}|^2/(r^2 \sin^2\theta)$, with $\nu_0 = 2.8 \times 10^5$ s. On the photosphere, the magnetofrictional velocity is set to zero in both models. See K. Barczynski et al. (2022) for further details of the Cartesian model and A. R. Yeates & P. Bhowmik (2022) and A. R. Yeates (2024) for further details of the the global model.

The initial condition for each model is a potential field extrapolation from the first frame of the corresponding photospheric data. In theory, the initial condition could also be defined using a linear or nonlinear force-free extrapolation method (see, e.g., T. Wiegelmann & T. Sakurai 2021 for a review of such methods). In practice, the twist parameter α must be estimated for these extrapolation methods, either from vector magnetograms or observations of coronal structures. Such approaches are subject to their own uncertainties, e.g., in the disambiguation of horizontal magnetic fields (T. R. Metcalf et al. 2006), or due to spatial downsampling (J. K. Thalmann et al. 2022), and are typically more successful in strong-field regions and/or where coherent coronal structures can be observed. In the present study, we are interested in how long it takes for each simulation to "forget" its initial condition, not the form of the initial condition, so the simpler potential field extrapolation is sufficient for our needs.

2.3. Toy Model

We created a "toy model" for comparison with the simulations described above. A series of 201 synthetic "magnetograms" was generated, each of size 256 × 256 pixels. For each synthetic magnetogram, a 16 × 16 pixel grid of squares of side 16 was randomly generated, with uniformly distributed values of −1, 0, or 1. Figure 1(d) shows the first synthetic magnetogram in the series. Since each synthetic magnetogram is independent of every other in the series, we expect any time series of quantities derived from this model not to display evidence of persistence. We compute a potential field extrapolation from each of the synthetic magnetograms, but it would not be meaningful to run a magnetofrictional simulation

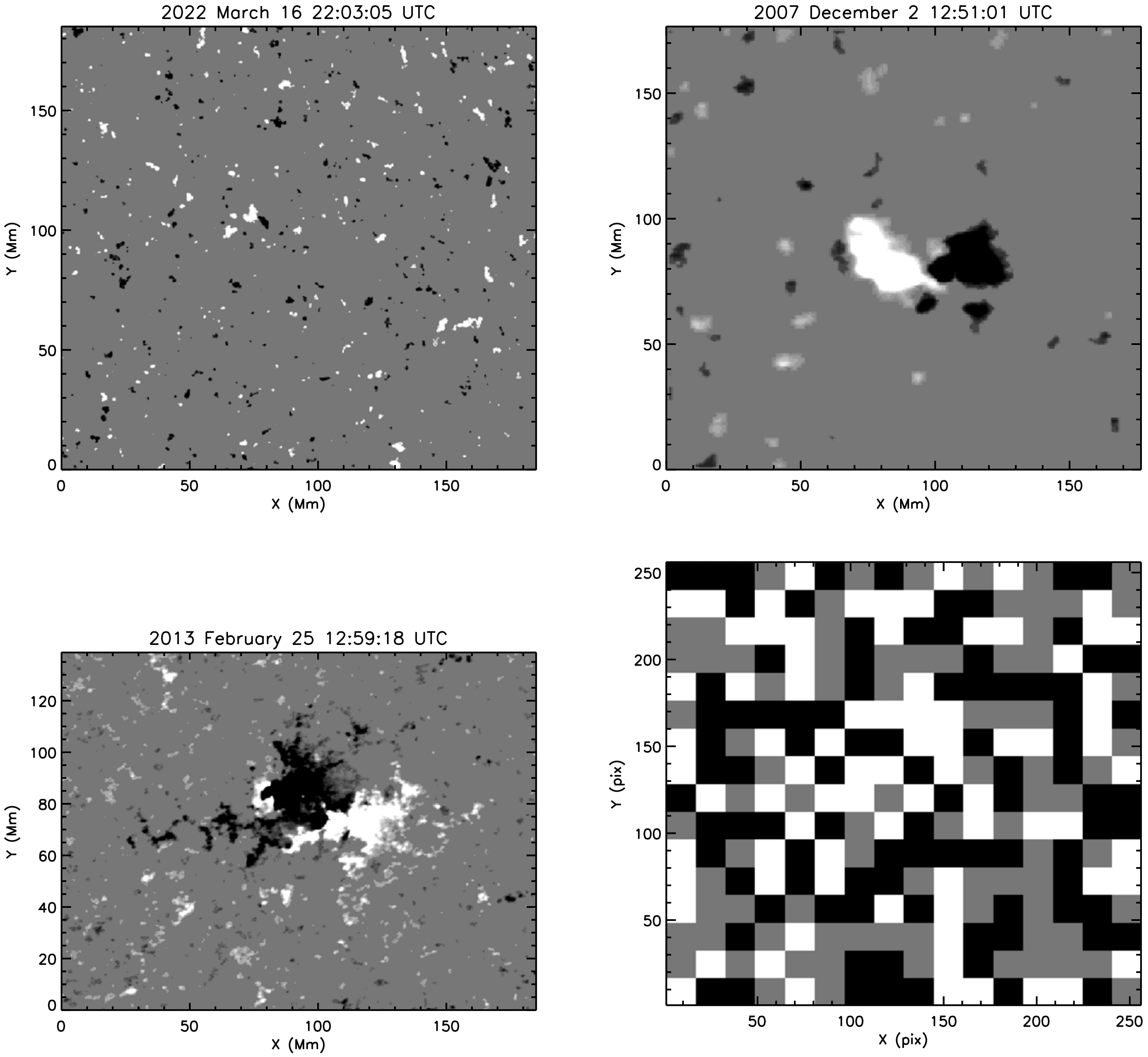

Figure 1. Initial cleaned magnetogram for each localized simulation: (a) QS; (b) AR10977; and (c) AR11680. The black and white regions indicate the negative and positive magnetic field, respectively. The QS magnetogram is saturated at ±30 G and the active region magnetograms at ±100 G. (d) Shows a "toy model," described in Section 2.3, saturated at ±1 unit.

with this synthetic series, due to the independence of each magnetogram from the next.

2.4. Solar Quantities

We are interested in several scalar-valued quantities calculated over time. These quantities can depend on the evolution of the 3D magnetic field together with the observational data (simulation-based) or on the observational data alone (observation-based).

1. Total flux. The total magnetic flux through the photosphere is given by

$$\Phi_{\rm tot} = \int_S B_j \, dS, \tag{4}$$

where $j = z$ in the Cartesian simulations and $j = r$ in the spherical simulation. S is the photospheric surface area. The simulations that are flux-balanced (AR10977 and the global simulation) have equal amounts of positive and negative magnetic flux through the photosphere, so their total flux is 0.

2. Total absolute flux and open flux. The total absolute flux through the photosphere is

$$\Phi_{\rm abs} = \int_S |B_j| \, dS, \tag{5}$$

with j as above. For the spherical simulation, we also compute the absolute flux through the outer boundary at $r = 2.5\,R_\odot$, given by

$$\Phi_{\rm open} = \int_{r=R_{ss}} |B_r| \, dS, \tag{6}$$

which is known as the "open flux."

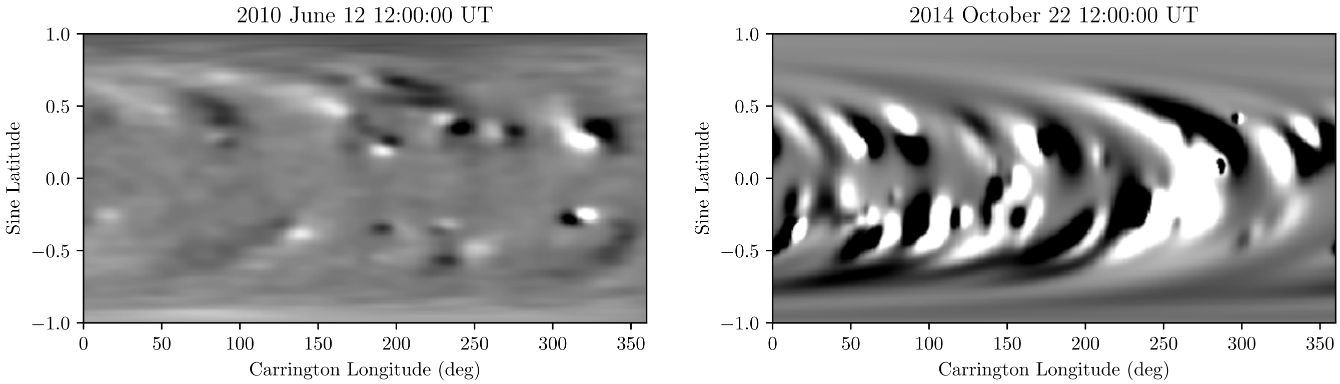

Figure 2. Example magnetograms from the global spherical simulation: (a) the initial Carrington map, smoothed to an appropriate resolution; and (b) at a later time near solar maximum, as generated by the SFT model. The black and white regions indicate the negative and positive magnetic field, respectively. The magnetograms are saturated at ±10 G.

3. Total magnetic energy. The total magnetic energy (erg) of the simulated coronal magnetic field at each time step is given by

$$W_{\text{tot}} = \frac{1}{8\pi}\int_V |\boldsymbol{B}|^2\, dV, \tag{7}$$

where V is the volume of the simulation domain.

4. Potential field magnetic energy. For the Cartesian simulations, we compute the potential field magnetic energy, which is the total magnetic energy of the potential magnetic field, $\boldsymbol{B}_{\text{pf}}$, extrapolated from the same boundary conditions as the nonpotential field at each time (see Table 1):

$$W_{\text{pf}} = \frac{1}{8\pi}\int_V |\boldsymbol{B}_{\text{pf}}|^2\, dV.$$

A potential field is the minimum energy state for given boundary conditions, so $W_{\text{pf}}(t) \leqslant W_{\text{tot}}(t)$ for all t.

5. Free magnetic energy. The free magnetic energy is the amount of magnetic energy in the coronal volume in excess of the potential field magnetic energy, given by

$$W_{\text{free}} = W_{\text{tot}} - W_{\text{pf}}. \tag{8}$$

This is typically considered to be energy that is available for release, e.g., due to a solar flare.

6. Mean electric current density. The mean electric current density in the simulation is given by

$$\langle J\rangle = \frac{\int_V |\nabla \times \boldsymbol{B}|\, dV}{\int_V dV}. \tag{9}$$

This is an indicator of nonpotentiality, as well as a useful diagnostic quantity for identifying eruptive behavior in global simulations (V. Aslanyan et al. 2024).

We plot time series of solar quantities for the active regions, QS, and Solar Cycle 24 in Figure 3 and quantities for the toy model in Figure 4. Note that not all quantities are calculated for every simulation. For example, the potential field magnetic energy (and hence the free magnetic energy) requires the extrapolation of a potential magnetic field at every time step of the simulation, which would be time-consuming for the 10 yr long Solar Cycle 24 simulation. We find similar results in our persistence analysis across each solar quantity within a simulation, and several solar quantities can be compared across all simulations (e.g., the total magnetic energy and total absolute flux).

3. Solar Quantities Exhibit Long Memory

In this work, we view the outputs of data-driven solar physics simulations as stochastic. Although the coronal models in Section 2.2 are deterministic, the magnetofrictional method assimilates observational magnetogram data through time. These observational data—see Figures 1 and 2—are a source of both aleatoric uncertainty, i.e., natural and random variation that is irreducible, and epistemic uncertainty, e.g., arising from measurement errors and missing data. Due to observational limitations, many epistemic uncertainties, such as those arising from inaccessible data, are also irreducible for solar simulations. Therefore, we regard each solar quantity in Section 2.4 as inherently stochastic and consider each as a stochastic process, $\{Y(t)\}_{t\geqslant 0}$, a collection of random variables indexed by time. We analyze time series, or sequences of recorded observations $\{Y(t_1) = y(t_1), \ldots, Y(t_n) = y(t_n)\}$, through (equidistant) discrete times $t = t_1, \ldots, t_n$, typically $t \in [0, 1, \ldots, n] \subset \mathbb{N}_0$ (see, e.g., the monographs by J. D. Hamilton 1994; P. J. Brockwell & R. A. Davis 1991; G. E. P. Box et al. 2015), as displayed in Figure 3.

This stochastic perspective provides additional tools for understanding and analyzing the strength and persistence of various physical processes that underpin the simulation, such as the memory of flux connectivity, which is a key feature of magnetofrictional methods. For instance, we anticipate that simulated observations made closer together in time may be more strongly associated than those made farther apart. Figure 5 displays scatter plots of the free magnetic energy at time t against lagged versions at time $t-h$, for $h = 1, 2, \ldots, 9$ time steps (defined in Table 1), for an active region and QS simulation. Although the cadences of the data assimilation and the timescales over which each simulation evolves differ, Figure 5 displays qualitative evidence that the sample correlations (a measure of linear association) remain high at moderate to high lag distances. Moreover, the time series in Figure 3 are characterized by long excursions or local trends, in contrast to the stationary time series for the toy model in Figure 4. All of these features suggest some level of predictability about the underlying process. In the next section, we recall the statistical concept of long memory required to make the observation of long-range correlations precise.

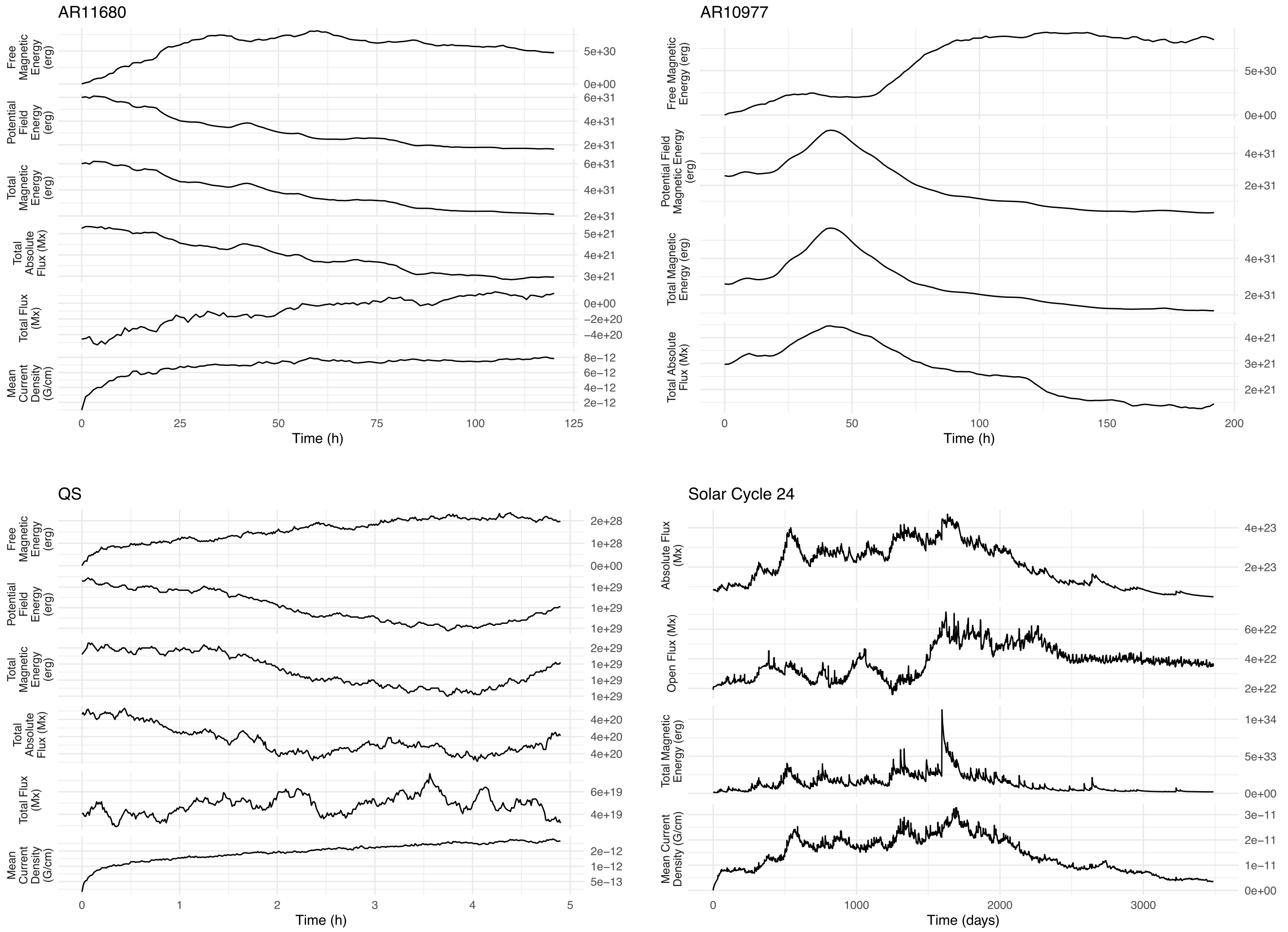

Figure 3. Time series of solar quantities for AR11680 (a), AR10977 (b), and QS (c), based on data-driven magnetofrictional simulations including the free magnetic energy, total magnetic energy, potential field magnetic energy, total flux, total absolute flux, and mean current density. Time series of solar quantities for global simulations over Solar Cycle 24 (d), from 2008 December to 2019 December, including the total absolute flux, open flux, total magnetic energy, and mean current density. The magnetofrictional method's initial condition results in nonphysical predictions, which can be observed in the free magnetic energy and mean current density values of zero at simulation time zero.

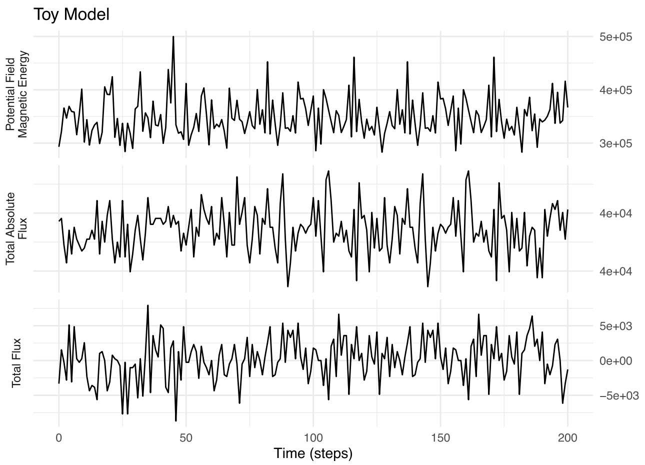

Figure 4. Time series of quantities for the toy model including total flux, total absolute flux, and potential field magnetic energy; note the absence of local trends or excursions compared to the time series in Figure 3.

3.1. Characterization of Long Memory

Long memory or persistence in a time series is marked by significant long-range dependence among observations (see, e.g., J. Beran 1994; J. Beran et al. 2013). For a stationary time series, persistence is characterized in the time domain by autocorrelations $\rho(h)$, a function of the time lag h, that decay algebraically to zero,

$$\rho(h) \simeq h^{2d-1}, \qquad \text{as } h \to \infty, \tag{10}$$

for $d \in (0, 0.5)$, i.e., slower than geometrically (J. R. M. Hosking 1981).[9] From Equation (10), we observe that long memory is an asymptotic property of the data-generating process. In the frequency domain, Equation (10) is equivalent to the spectral density function $f(\cdot)$ being unbounded for frequencies, λ, near zero—that is,

$$f(\lambda) \simeq |\lambda|^{-2d}, \qquad \text{as } \lambda \to 0^{+}. \tag{11}$$

In Equations (10) and (11), d is a parameter that captures "long memory," "persistent autocorrelation," or "long-term dependence."

The implied spectral densities of the solar quantities from Section 2.4 exhibit the hallmarks of long memory. In Figure 6, we plot periodograms, i.e., the log-estimated spectral density for the demeaned series versus frequency, using a fast Fourier transform; for the sample rate f_s, the spectral density scaling is

[9] Further details about autocorrelations can be found in Appendix A.

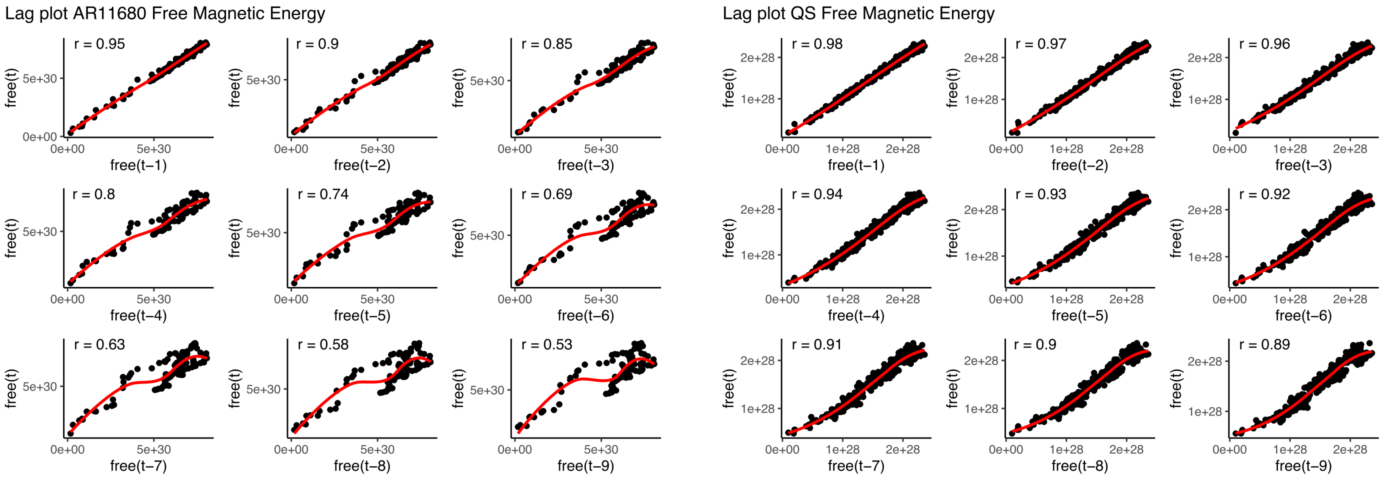

Figure 5. Lag plots of the free magnetic energy for AR11680 (a) and QS (b) indicating high levels of correlation r at moderate lags. Each lag plot includes a local polynomial regression (LOESS) fit, using a second-degree polynomial with smoothing parameter $\alpha = 0.75$.

$1/f_s$ and the frequency λ has been matched to the timescale of the series (cycles per hour or cycles per unit time). For stationary time series, the periodogram with scaling $1/f_s$ is a density over $(-f_s/2, f_s/2)$, whose integral is the variance of the series. However, long-run estimates of the periodogram for a long-memory process will have an infinite density at frequency zero, consistent with Equation (11). We observe such blowup in the implied periodograms of the free magnetic energy and total absolute flux for AR11680 (Figures 6(a) and (b)), AR10977 (Figures 6(a) and (d)), and QS (Figures 6(a) and (f)), and in the mean current density and total absolute flux in Solar Cycle 24 (Figures 6(g) and (h)). Hence, solar quantities derived both from observed magnetograms (total absolute flux) and magnetofrictional simulations (free magnetic energy) exhibit behavior consistent with long memory. This contrasts sharply with the bounded behavior observed in the periodogram for the toy model in Figure 7, which is constant.

Although it is possible to attempt to fit the parameter d in Equations (10) and (11) directly, a more robust approach is to consider a class of generative time-series models that capture the desired power-law behavior of the spectral density. Notably, a fractionally integrated process of order $d \in (0, 0.5)$ has the spectral density of Equation (11); see, e.g., J. Geweke & S. Porter-Hudak (1983).[10] Fractionally integrated processes are related to FGN (see Section 1) and are extended by autoregressive fractionally integrated moving average models, first introduced independently by C. W. J. Granger & R. Joyeux (1980) and J. R. M. Hosking (1981; see also the modern monographs by J. Beran 1994; P. J. Brockwell & R. A. Davis 1991; J. Beran et al. 2013). We let $I(d)$ denote the class of fractionally integrated process of order $d \in \mathbb{R}$.

3.2. Memory Parameter Inference

Typical inference approaches in the context of $I(d)$ processes rely on the spectral density of the series. Within such approaches, likelihood methods involve the numerical minimization of a likelihood function, which for time series takes the form of the Whittle likelihood (H. R. Kuensch 1987; P. M. Robinson 1995a). In the analysis below, we present inferences using the exact local Whittle (ELW) estimator (K. Shimotsu & P. C. B. Phillips 2005). Possible alternatives include the two-step ELW estimator (K. Shimotsu 2010), which uses tapers to extend the range of consistency, and the modified local Whittle estimator of Hou and Perron (J. Hou & P. Perron 2014). These likelihood estimators are all semiparametric, in that they do not require estimating intermediate parameters relating to short memory in the system (that is, they estimate the parameter d only). Assuming that our time series are observations from the class $I(d)$ processes, we will quantify the nature and strength of the persistence over the timescale of the simulation by inferring the parameter d.

For the sake of comparison with other studies, we also present the log-periodogram estimator of Geweke and Porter-Hudak (GPH; J. Geweke & S. Porter-Hudak 1983; P. M. Robinson 1995a), a rescaled-range (R/S-AL) estimator (A. A. Annis & E. H. Lloyd 1976; R. Weron 2002), and a DFA (C. K. Peng et al. 1994). GPH estimates d based on ordinary least-squares estimates of the slope parameter in a linear regression of the log-periodogram on a deterministic regressor; alternatively, one could consider the McCloskey and Perron estimator, which is reported to be robust to low-frequency contamination (A. McCloskey & P. Perron 2012). R/S-AL is an estimator for the Hurst exponent H in the FGN model (B. B. Mandelbrot & J. R. Wallis 1969b), with the Anis–Lloyd correction for small-sample bias. In our tables and figures, we report the corresponding value $\hat{d} = \hat{H} - 0.5$ (where we adopt the typical use of “hats” to distinguish estimated quantities), with the caveat that the latter relation is only validated for stationary regimes, i.e., $d \in (-0.5, 0.5)$. The DFA parameter $\alpha \approx H$ and we report $\hat{d} = \hat{\alpha} - 0.5$, noting that α is anticipated to overestimate H in nonstationary regimes (where $H \approx \alpha - 1$; J. W. Kantelhardt et al. 2002). We further caution that DFA introduces uncontrolled bias and may be inappropriate for nonstationary series (R. M. Bryce & K. B. Sprague 2012).

While ELW can be applied directly to estimate any $d \in \mathbb{R}$, GPH and R/S-AL are in principal for stationary series with $d \in (-0.5, 0.5)$. We recall that an $I(d)$ process is difference-stationary, in that it can be differenced d times to obtain a stationary white-noise process. As our ELW estimates typically suggest $d > 0.5$, we utilize our ELW estimate $\hat{d}$ to determine a suitable number of integer differences for each series. To have a fair comparison of the estimators (see Remark 1), we apply

[10] Further details about fractionally integrated processes can be found in Appendix B.

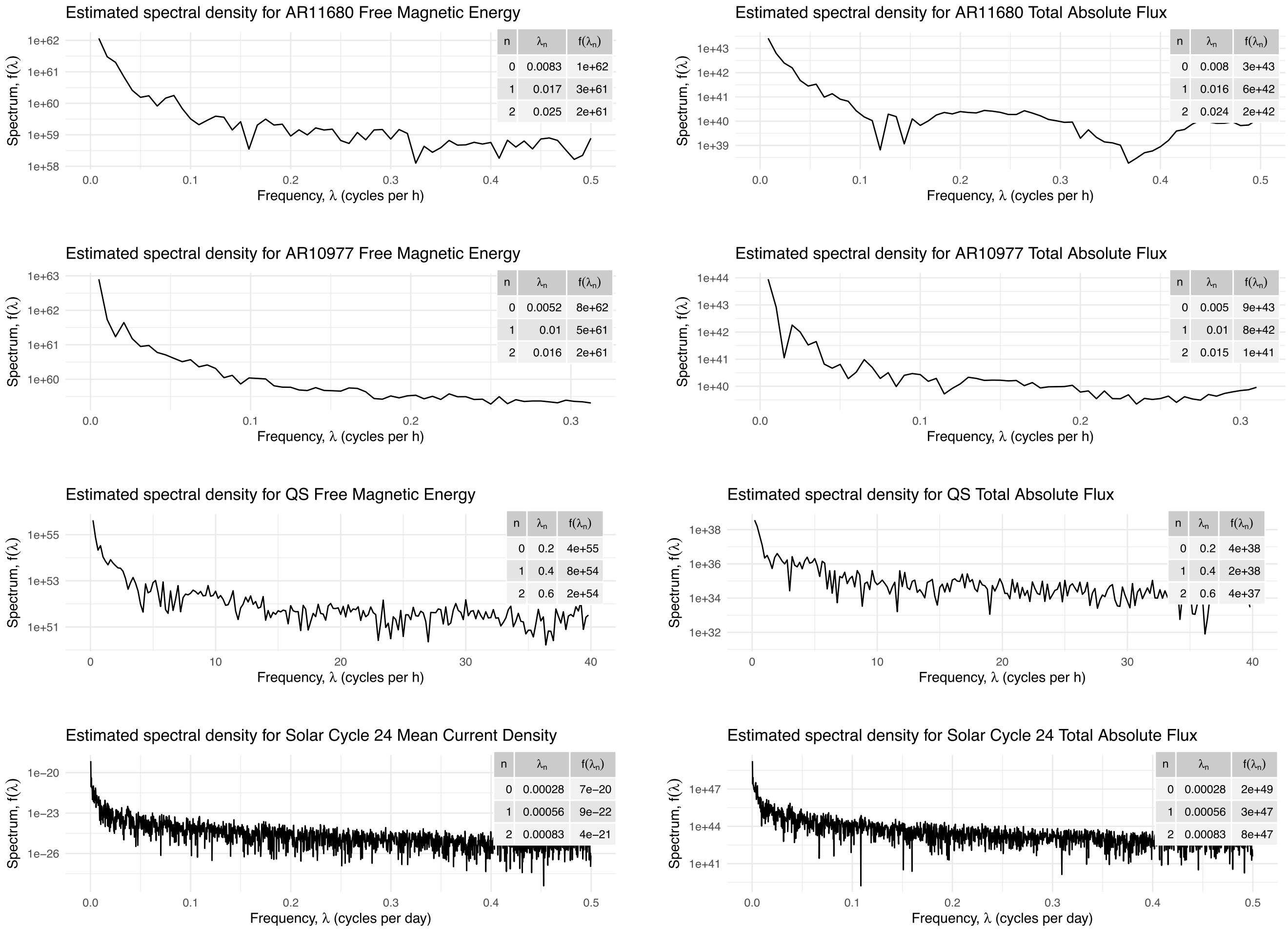

Figure 6. Periodograms for the free magnetic energy and total absolute flux of AR11680 (a) and (b), AR10977 (c) and (d), and QS (e) and (f), and for the mean current density and total absolute flux for Solar Cycle 24 (g) and (f). Each demonstrates blowup for frequencies near zero, consistent with long memory (see the similar spectral behavior in Figure 15(b)).

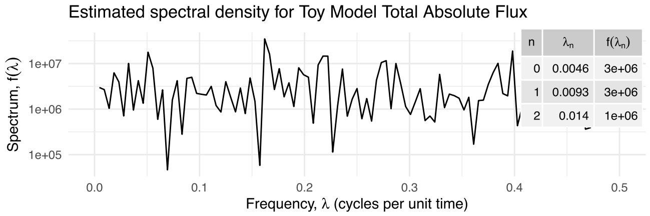

Figure 7. Periodogram for the toy model, where the lack of blowup near zero contrasts with the periodograms in Figure 6.

GPH and R/S-AL to data preconditioned with the indicated difference filter; GPH and R/S-AL applied to preconditioned data are recorded as GPH* and R/S-AL*.[11]

The estimation methods considered in our persistence analysis are summarized in Table 4. In the next section, we identify quantitative evidence of long memory in our solar quantities using these inference procedures.

3.3. Quantitative Evidence of Long Memory

We display estimates $\hat{d}$ for a range of solar quantities for regional and global simulations in Figure 8 (see also Table 2). Estimates for the memory parameter all exceed $\hat{d} > 0.5$, indicating the existence of long memory in these time series over the timescales of the simulation (respectively, hours to days, in Figures 8(a)–(c), and years, in Figure 8). The identification of long memory in simulation-based quantities and over timescales of hours to days is a new finding in solar physics. We observe in Figure 8 that the estimated persistence strengths are largely consistent across the solar quantities within each simulation. In particular, the observation-based solar quantities (e.g., absolute flux) are comparable to the simulation-based quantities (e.g., free magnetic energy), demonstrating that magnetofrictional simulations capture the memory structure present in magnetogram data. Across simulations, the persistence estimates for AR11680 and AR10977 exceed the estimates for QS, which in turn all exceed the estimates for Solar Cycle 24. In contrast, the estimates for the toy model in Figure 9 suggest $\hat{d} = 0$, consistent with a 1D white noise.

Previous studies of solar records have identified long memory in nonperiodic variations over a wide range of timescales, from 20 days to 3000 yr (see, e.g., B. B. Mandelbrot & J. R. Wallis 1969a; A. Ruzmaikin et al. 1994; R. W. Komm 1995; F. Lepreti et al. 2000; M. G. Ogurtsov 2004). These studies have tended to rely on Hurst rescaled-range analysis or DFA and on a variety of different proxies for solar activity. B. B. Mandelbrot & J. R. Wallis (1969a) look at monthly sunspot activity from

[11] An alternative to estimating the integrated order that uses unit root testing is described in Appendix C.

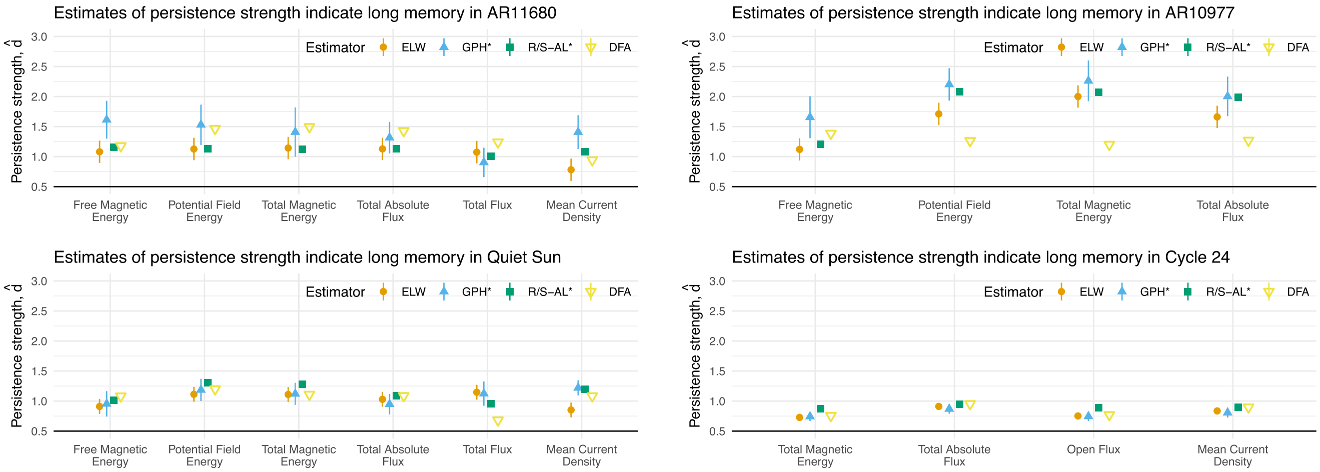

Figure 8. Estimates of persistence in (a) AR11680, (b) AR10977, (c) QS, and (d) Solar Cycle 24 indicate long memory (i.e., $\hat{d} > 0.5$). The levels of persistence across all solar quantities in active regions are higher than in QS, which in turn are higher than in Solar Cycle 24 (see Table 2 for full details; a summary of the estimators is given in Table 4).

1749 to 1948 and estimate a high Hurst index $H \approx 0.93$; A. Ruzmaikin et al. (1994) consider Carbon-14 data from dendrochronology records from the period 6000 BC to 1950 AD and estimate $H \approx 0.8$ over timescales from 100 to 3000 yr; and R. W. Komm (1995) considers daily differential rotation measurements from 1967 to 1992 and estimates $H \approx 0.83$ over timescales of 20 days to 11 yr. F. Lepreti et al. (2000) consider both the daily averaged intensity of optical flares from 1976 to 1996 and daily averaged sunspot numbers from 1951 to 1996 and find $H = 0.74 \pm 0.02$ over timescales of 24 to 450 days and $H = 0.76 \pm 0.01$ over timescales of 20 to 350 days, respectively.

Using the heuristic $H = d + 0.5$, the persistence strength we observe in the present study of regional and global simulations exceeds those identified in previous studies by a large margin.[12] There are several explanations for this phenomenon. First, we are considering solar quantities that are integrals of the magnetic field and anticipate that integrated quantities would yield higher persistence estimates (i.e., the free magnetic energy will have a higher level of persistence than the magnetic field, especially over the sunspot or flare numbers). Second, we utilize higher-cadence data than previous studies (owing in part to our access to modern magnetogram data), and downsampling is anticipated to have a degrading effect on the estimated persistence strength. Third, the Hurst R/S and DFA methods may consistently underestimate the true persistence strength for nonstationary time series. For this last point, we turn to a simple numerical experiment in Remark 1 to illustrate this bias.

Remark 1 (**Nonstationarity Bias**). The GPH, Hurst R/S, and DFA methods may consistently underestimate the true persistence strength for nonstationary time series with $|d| > 0.5$. In Figure 10, we present density estimates for a data-generating process with known $d = 1.6$ using DFA, ELW, GPH, R/S-AL, GPH*, and R/S-AL* (i.e., also with second-difference preconditioning). The density estimates are based on 1000 simulated $I(d = 1.6)$ series of length $N = 5000$. Even for very large time series, the GPH, DFA, and R/S-AL estimates are biased—the highest likelihood does not correspond to 1.6. One should further keep in mind that the reported DFA, $\hat{d} = \hat{\alpha} - 0.5$, is assumed to *overestimate* the persistence in nonstationary regimes.

4. Quantifying Simulation Burn-in Time

Quantifying a simulation's burn-in time, or the time required for a solar quantity to have sufficiently "forgotten" the simulation's initial state, is a critical computational question for automating research workflows in solar physics. For example, in data-driven magnetofrictional simulations, the choice of the initial 3D force-free magnetic field is somewhat arbitrary and a source of uncertainty, because routine 3D magnetic field observations of the Sun's corona are currently unavailable. We observe in Figures 3(a) and (c) that the initial potential field induces starting values of the free magnetic energy that are nonphysical, in the sense that they are zero-valued. In this context, the burn-in time is the time after which a solar quantity, such as the free magnetic energy, has evolved away from the initial potential field to a physically more realistic nonpotential state.

A similar challenge arises in molecular dynamics simulations, where the equilibrium statistics of a molecular system are calculated from simulations initiated from atypical (i.e., nonequilibrium) configurations (A. Grossfield et al. 2018). In such systems, correlations—e.g., the decay of particle velocity autocorrelation functions (ACFs; see, e.g., G. A. Pavliotis 2014; B. Leimkuhler & C. Matthews 2015)—are often used to assess the mixing or equilibration time. For example, advanced equilibration time detection algorithms exist for molecular dynamics that maximize the number of effectively uncorrelated samples in the simulation time span used to compute equilibrium statistics (J. D. Chodera 2016), and these are important for eliminating starting condition bias. A key difference in solar simulations is that the underlying time series exhibit long memory (Section 3) and will, therefore, have no characteristic decorrelation time.

Presently, we demonstrate an information-theoretic approach for calculating simulation burn-in time based on mutual information. Mutual information is a measure of dependence between random variables that quantifies the amount of shared information gained about one variable from observing the other (T. M. Cover & J. A. Thomas 2006). Mutual information can

[12] For the studies reported above, the implied values of d are, respectively, $d \approx 0.43$, $d \approx 0.3$, $d \approx 0.33$, $d \approx 0.24$, and $d \approx 0.26$.

Table 2
Details of ELW, GPH*, R/S-AL, and DFA Estimators for Persistence in Observation-based and Simulation-based Solar Quantities for Regional and Global Simulations All Satisfying $\hat{d} > 0.5$, Indicating the Presence of Long Memory (See Figures 8 and 9)

		ELW		GPH*		R/S-AL*	DFA
Simulation	Quantity	$\hat{d}$	$\hat{d} \pm 2 \cdot \mathrm{se}(\hat{d})$	$\hat{d}$	$\hat{d} \pm 2 \cdot \mathrm{se}(\hat{d})$	$\hat{d}$	$\hat{d}$
AR11680	Free magnetic energy	1.08	(0.89, 1.27)	1.61	(1.3, 1.93)	1.16	1.18
AR11680	Potential field energy	1.13	(0.94, 1.31)	1.53	(1.19, 1.87)	1.13	1.47
AR11680	Total magnetic energy	1.14	(0.96, 1.33)	1.41	(1, 1.82)	1.12	1.50
AR11680	Total absolute flux	1.13	(0.94, 1.31)	1.32	(1.05, 1.58)	1.13	1.43
AR11680	Total flux	1.07	(0.89, 1.26)	0.90	(0.66, 1.15)	1.01	1.24
AR11680	Mean current density	0.78	(0.6, 0.97)	1.41	(1.13, 1.69)	1.08	0.94
AR10977	Free magnetic energy	1.12	(0.93, 1.31)	1.66	(1.31, 2)	1.21	1.38
AR10977	Potential field energy	1.71	(1.53, 1.9)	2.20	(1.93, 2.47)	2.08	1.26
AR10977	Total magnetic energy	2.00	(1.81, 2.19)	2.26	(1.92, 2.6)	2.07	1.20
AR10977	Total absolute flux	1.66	(1.48, 1.85)	2.00	(1.67, 2.33)	1.99	1.27
QS	Free magnetic energy	0.91	(0.79, 1.03)	0.95	(0.74, 1.16)	1.01	1.08
QS	Potential field energy	1.11	(0.99, 1.24)	1.19	(1, 1.37)	1.31	1.20
QS	Total magnetic energy	1.11	(0.99, 1.23)	1.12	(0.94, 1.3)	1.28	1.11
QS	Total absolute flux	1.03	(0.91, 1.15)	0.95	(0.78, 1.12)	1.08	1.09
QS	Total flux	1.15	(1.02, 1.27)	1.12	(0.92, 1.33)	0.95	0.68
QS	Mean current density	0.85	(0.73, 0.98)	1.22	(1.09, 1.35)	1.19	1.08
Toy Model	Potential field energy	−0.19	(−0.35, −0.03)	-0.27	(−0.53, −0.01)	−0.40	−0.24
Toy Model	Total absolute flux	−0.08	(−0.23, 0.08)	-0.10	(−0.35, 0.14)	0.11	−0.02
Toy Model	Total flux	0.14	(−0.02, 0.3)	0.27	(0.03, 0.52)	−0.11	0.18
Solar Cycle 24	Total absolute flux	0.91	(0.85, 0.97)	0.87	(0.79, 0.95)	0.94	0.95
Solar Cycle 24	Open flux	0.75	(0.69, 0.81)	0.75	(0.67, 0.82)	0.89	0.77
Solar Cycle 24	Total magnetic energy	0.73	(0.67, 0.79)	0.74	(0.67, 0.82)	0.87	0.76
Solar Cycle 24	Mean current density	0.84	(0.78, 0.89)	0.81	(0.72, 0.89)	0.90	0.90

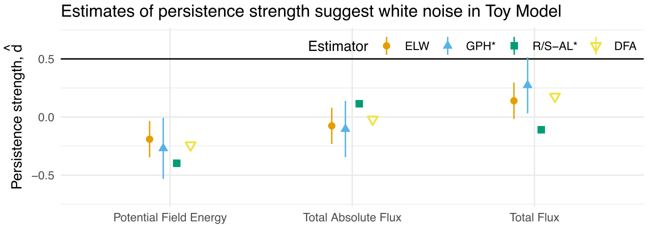

Figure 9. Estimates of persistence for the toy model. In contrast to Figure 8, these are not significantly different from zero, suggesting a white-noise model.

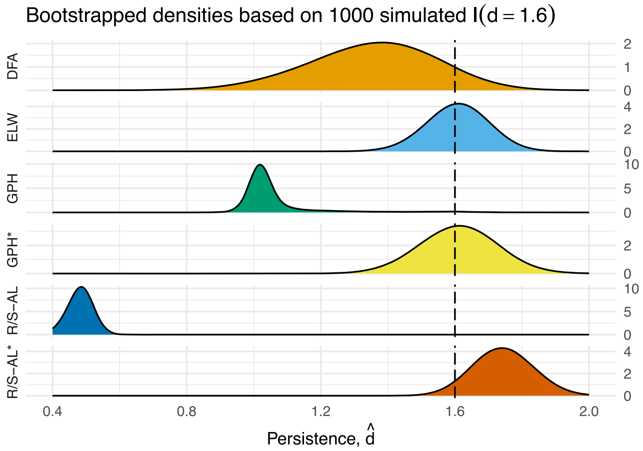

Figure 10. Bootstrapped KDEs for the persistence parameter $\hat{d}$, based on 1000 simulated $I(d = 1.6)$ series of length 5000, illustrate the bias in DFA, GPH, and R/S-AL for nonstationary time series, in contrast to the ELW method, which selects the correct value, $d = 1.6$, with high likelihood. If the appropriate difference filter is known a priori (in this case, two differences), the bias can be mitigated for GPH and (to a lesser extent) R/S-AL.

be applied to time series with long memory, where correlation-based tools would be inappropriate.

4.1. TDMI

In our specific context, we are interested in the mutual information between probability distributions generated by the time series of a solar quantity. That is, for Y, we consider $Y_{[0:t]}$, the distribution of which is given by the occupation density over $(Y(0), \ldots, Y(t))$. The TDMI of Y (also known as the average mutual information) considers the mutual information between the lagged distributions $Y_{[\tau:t]}$ and $Y_{[0:t-\tau]}$. For each τ, the TDMI quantifies the information we already possess about $Y_{[\tau:t]}$ if we already know $Y_{[0:t-\tau]}$. TDMI has been successfully used for investigations of nonlinear and complex phenomena, including the statistical analysis of nonlinear dynamics (J. A. Vastano & H. L. Swinney 1988; H. Kantz & T. Schreiber 2003) and in causal inference for nonlinear systems (S. Li et al. 2018).

Formally, the TDMI for y is a function of the lags (time steps) τ, given by

$$I(\tau) := \iint f_{Y_{[0:t-\tau]}, Y_{[\tau:t]}}(x, x') \times \log \frac{f_{Y_{[0:t-\tau]}, Y_{[\tau:t]}}(x, x')}{f_{Y_{[0:t-\tau]}}(x) f_{Y_{[\tau:t]}}(x')} \, dx dx' , \tag{12}$$

where $f_{Y_{[0:t-\tau]}, Y_{[\tau:t]}}$, $f_{Y_{[0:t-\tau]}}$, and $f_{Y_{[\tau:t]}}$ represent the joint and marginal densities for the path distributions over $[0: t - \tau]$ and $[\tau: t]$, respectively. The TDMI has the important property that $I(\tau) \geqslant 0$ is strictly nonnegative, with equality if and only if $Y_{[0:t-\tau]}$ and $Y_{[\tau:t]}$ are independent—that is, if Y over the interval

$[0\colon t-\tau]$ provides no information about Y over the interval $[\tau\colon t]$.

The densities required in Equation (12) can be estimated using kernel density estimates (KDEs), a nonparametric approach to estimating a probability density function that has improved statistical properties over histogram binning (see, e.g., the monograph by T. Duong 2018 and references therein). For an independent and identically distributed (iid) sample ($\boldsymbol{X}_i$, …, $\mathbf{X}_n$) from a common k-dimensional density f, the general form of a KDE is

$$\hat{f}_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{b}) = n^{-1}\sum_{i=1}^{n} K_{\boldsymbol{b}}(\boldsymbol{x}-\boldsymbol{X}_i)\,, \quad \boldsymbol{x}\in\mathbb{R}^k\,, \tag{13}$$

where K_b is a kernel smoothing function (an integrable function with unit integral), with smoothing bandwidth $\boldsymbol{b}$ (a symmetric, positive, definite $k\times k$ matrix of smoothing parameters). The bandwidth $\boldsymbol{b}$ is the key parameter that impacts the quality of the estimates. For our study, we utilize the (data-driven) plug-in selector of T. Duong (2010) for unconstrained bandwidths, a class of bandwidths recommended for most data analysis in M. P. Wand & M. C. Jones (1993), and a standard Gaussian kernel.[13]

A plug-in estimator for Equation (12) using KDEs is then given by

$$\begin{aligned}\hat{I}(\tau) \coloneqq &\sum_{y_{[0:t-\tau]}}\sum_{y_{[\tau:t]}}\hat{f}(y_{[0:t-\tau]}, y_{[\tau:t]})\\ &\times \log\frac{\hat{f}(y_{[0:t-\tau]}, y_{[\tau:t]})}{\hat{f}(y_{[0:t-\tau]})\hat{f}(y_{[\tau:t]})}\,,\end{aligned} \tag{14}$$

for a time series $\{y(t_i)\}_{i=0,\ldots,n}$, where we have omitted subscripts when obvious from the context. The estimator of Equation (14) may perform poorly for short time series; a similar problem is noted for the calculation of the transfer entropy, a quantity related to Equation (12), where corrections based on a permutation-based resampling related to bootstrapping are introduced to mitigate small-sample bias (R. Marschinski & H. Kantz 2002; A. Papana et al. 2011).

To calculate an effective TDMI, one approach (borrowing from the calculation of transfer entropy) is to randomly shuffle the components of the series $\tilde{y}_{[0:t-\tau]}$, where the permutation is chosen uniformly at random from the set of all possible permutations. The random shuffle theoretically destroys all dependence between $\tilde{y}_{[0:t-\tau]}$ and $y_{[\tau:t]}$, thus any observed nonzero mutual information is an artifact of the finite sample size, which is expected to decrease to zero as the number of permutations utilized increases. We consider the shuffle correction:

$$\begin{aligned}\hat{I}_{\text{shuffle}}(\tau) = &M^{-1}\sum_{m=1}^{M}\\ &\times\left(\sum_{\tilde{y}^{(m)}_{[0:t-\tau]}}\sum_{y_{[\tau:t]}}\hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]}, y_{[\tau:t]})\log\frac{\hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]}, y_{[\tau:t]})}{\hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]})\hat{f}(y_{[\tau:t]})}\right),\end{aligned} \tag{15}$$

which averages M plug-in estimators of Equation (14) with a random permutation $\tilde{y}^{(m)}_{[0:t-\tau]}$ in place of $y_{[0:t-\tau]}$. The correction is detailed in Algorithm 1. Other corrections are also possible; a block bootstrap correction is utilized for transfer entropy in T. Dimpfl & F. J. Peter (2012) and investigating generalizations to long-memory time series might be interesting.

An analysis of the local minima of the effective TDMI,

$$\hat{I}_{\text{eff}}(\tau) = \hat{I}(\tau) - \hat{I}_{\text{shuffle}}(\tau)\,, \tag{16}$$

can be used to identify the simulation burn-in time. Specifically, the first local minimum,

$$\tau_0 = \min\{\tau\colon \hat{I}'_{\text{eff}}(\tau) = 0 \text{ and } \hat{I}''_{\text{eff}}(\tau) > 0\}\,, \tag{17}$$

corresponds to the first time lag for which $y_{[0:t-\tau]}$ is minimally informative of $y_{[\tau:t]}$ (conversely, $y_{[\tau:t]}$ represents maximal information beyond the knowledge we have from $y_{[0:t-\tau]}$). We take the corresponding simulation time,

$$t^* = t_{\tau_0} > t_0, \tag{18}$$

as our simulation burn-in. This burn-in criterion is a reasonable choice to represent the first time at which $y(t^*)$ has "forgotten" $y(t_0)$ and mirrors the criterion used in equilibrium systems—namely, the first lag at which the ACF decays to zero. As in the nonlinear dynamics literature (H. Kantz & T. Schreiber 2003), we take for granted that such a minimum will exist. The full procedure for calculating the simulation burn-in time is summarized in Algorithm 2.

Although it is possible to compute the TDMI over the whole time range of the simulation, in most instances, it will not be necessary. To reduce computational overhead, the calculation of the TDMI can be carried out over a window,

$$[0\colon t_{W(n)}] \subset [0\colon t_n]\,,$$

a subset of the time range, provided it includes the solar quantity's transition to a nonpotential state. One approach to choosing the window would be to first calculate a potentially biased estimate of $\hat{I}(\tau)$ and then use the observed first minimum to calculate the effective TDMI over a window about twice the range.

We comment that there are alternative approaches to calcating Equation (12) that replace KDEs with other representations of the path distribution of the time series. These include symbolization, or converting the path-space distribution into a categorical variable (M. Staniek & K. Lehnertz 2008), and closely related binning procedures (e.g., H. Kantz & T. Schreiber 2003), which create a histogram rather than a smooth representation of the path-space distribution. However, using symbolization or histograms in place of KDEs would necessitate the adoption of various approaches to attenuate the bias introduced by the choice of anchor points, such as additional sensitivity analyses, averaging several shifted estimators, and/or the use of hexagonal bins for joint densities.

[13] See Appendix D for the additional software packages utilized for density estimation.

Algorithm 1. Shuffle (Small-sample Bias Correction for TDMI)

```
Input: Time series $\{y(t_i)\}$, for $i = 0, \ldots, n$
Input: Time lag $\tau$
Input: Number of replicates $M$
Output: Bias correction $\hat{I}_{\rm shuffle}$ at lag $\tau$
For each $m = 1, \ldots, M$ do
    Random shuffle $\tilde{y}^{(m)}_{[0:t-\tau]} \leftarrow \mathrm{Permute}\{y(0),\ldots,y(t-\tau)\}$
    Estimate densities $\hat{f}_{\tilde{y}^{(m)}_{[0:t-\tau]}, y_{[\tau:t]}}$, $\hat{f}_{\tilde{y}^{(m)}_{[0:t-\tau]}}$, and $\hat{f}_{y_{[\tau:t]}}$        ▷Using KDEs (13)
    Plug-in estimator $\hat{I}^{(m)} \leftarrow \sum_{\tilde{y}^{(m)}_{[0:t-\tau]}} \sum_{y_{[\tau:t]}} \hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]}, y_{[\tau:t]}) \log \frac{\hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]}, y_{[\tau:t]})}{\hat{f}(\tilde{y}^{(m)}_{[0:t-\tau]})\hat{f}(y_{[\tau:t]})}$        ▷Following (15)
End for
Mean $\hat{I}_{\rm shuffle} \leftarrow M^{-1}\sum_{m=1}^{M} \hat{I}^{(m)}$
return $\hat{I}_{\rm shuffle}$
```

Algorithm 2. Simulation Burn-in Time

```
Input: Time series $\{y(t_i)\}$, for $i = 0, \ldots, n$
Output: Simulation time $t^{*} \in \{t_0, \ldots, t = t_n\}$
For each lag $\tau = 1, \ldots, W(n)2$:        ▷E.g., $W(n) = n$ for short regional simulations
    Estimate densities $\hat{f}_{y_{[0:t-\tau]}, y_{[\tau:t]}}$, $\hat{f}_{y_{[0:t-\tau]}}$ and $\hat{f}_{y_{[\tau:t]}}$        ▷Using KDEs (13)
    Calculate $\hat{I}(\tau)$        ▷Using plug-in estimator (14)
    $\hat{I}_{\rm correction}(\tau) \leftarrow 0$
    If correction then
        Calculate $\hat{I}_{\rm correction}(\tau)$        ▷E.g., using Algorithm 1
    End if
    Effective TDMI $\hat{I}_{\rm eff}(\tau) \leftarrow \hat{I}(\tau) - \hat{I}_{\rm correction}(\tau)$        ▷Following (16)
End for
Local minima $\mathcal{T} \leftarrow \{\tau\colon \hat{I}'_{\rm eff}(\tau) = 0$ and $\hat{I}''_{\rm eff}(\tau) > 0\}$
First local minimum $\tau_0 \leftarrow \min \mathcal{T}$        ▷Following (17)
Simulation time $t* \leftarrow t_{\tau_0}$        ▷See (18)
Return Burn-in time $t^{*}$
```

4.2. Burn-in Time for Magnetofrictional Simulations

In determining the simulation burn-in, we consider a single key solar quantity from each simulation related to the nonpotentiality of the simulation. For the regional simulations AR10977, AR11680, and QS, that quantity is the free magnetic energy; for Solar Cycle 24, it is the mean current density. For the toy model, it is not meaningful to run a magnetofrictional simulation, since each "magnetogram" is unrelated, therefore we choose the potential field magnetic energy. On the left, in Figures 11–13, we plot the TDMI profiles calculated using Algorithm 2 with the location of the first minimum, τ_0 (Equation (17)), for both uncorrected (i.e., none) and shuffle-corrected profiles, with the latter using Algorithm 1 with $M = 30$. On the right, in Figures 11–13, we plot the time series and the burn-in time; the portion of the series used to reckon the burn-in time is highlighted in black. As prescribed in Algorithm 2, we select a window (a function of the length of the series n) over which to calculate the burn-in time that we believe to contain the transition to the nonpotential state. For the relatively short active regional simulations AR10977 and AR11680, we select the window $W(n) = n = 121$ (i.e., the whole series); for the QS, we select $W(n) = \lfloor 0.306\, n \rfloor = 120$; for Solar Cycle 24, we select $W(n) = \lfloor .\ 104n \rfloor = 365$ (i.e., the first year); and for the toy model, we select $W(n) = \lfloor 0.6n \rfloor = 120$. These window selections guarantee that the TDMI profiles displayed on the left in Figures 11–13 are calculated with kernel densities containing at least 60 samples each. The first minima and burn-in times are summarized in Table 3.

Across Figures 11(a), (c), and (e), Figure 12(a), and Figure 13(a), we note that the general shapes of the uncorrected and shuffle-corrected TDMI profiles are largely coherent, with the shuffle-corrected TDMI curves satisfying the nonnegativity property (even for a small number of shuffles $M = 30$). The burn-in time for AR10977 in Figure 11(b) is approximately 2 days, and there is a rather large disparity between the first minima observed using the uncorrected and shuffle-corrected TDMIs, owing to the overall flatness of the profiles. On the one hand, the suggested burn-in time may be viewed as overly conservative, due to the simulation start time coinciding with the potential field increasing and then decreasing rapidly over approximately 50 hr (starting around simulation hour 25). On the other hand, any interesting phenomena occurring near the simulation start should be viewed with caution, due to their temporal proximity to the initial potential state. AR10977 is undergoing emergence during roughly the first 40 hr of the simulation, which can be seen in the increasing total absolute flux in Figure 3(b). The free magnetic energy increases during this time, as energy is injected into the simulation by the emerging

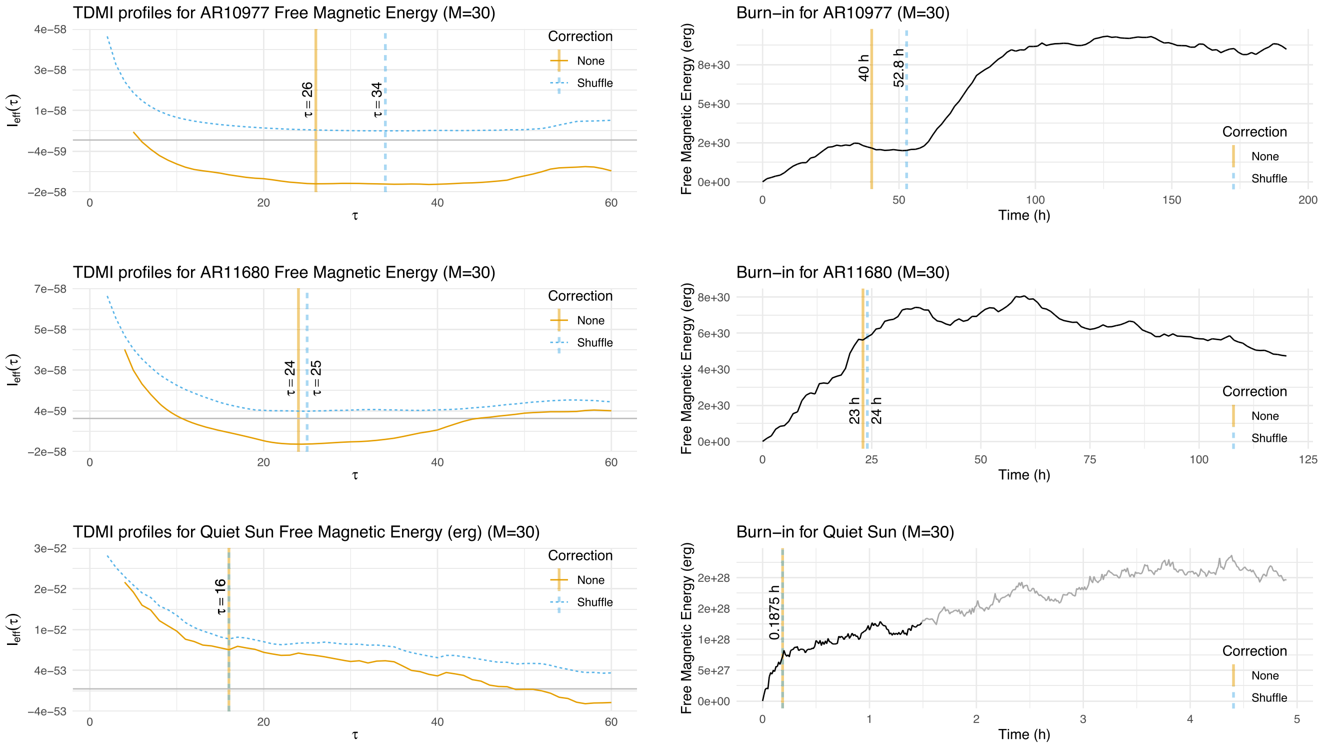

Figure 11. For the regional simulations, the automated burn-in time calculated using Algorithm 2 provides a conservative estimate of when the free magnetic energy has transitioned away from the initial potential state. Left: the TDMI profiles calculated for the free magnetic energy with the first minimum lag (Equation (17)). Right: the free magnetic energy with the associated burn-in time; the window used for calculating the burn-in time is the portion of the path highlighted in black. See also Table 3.

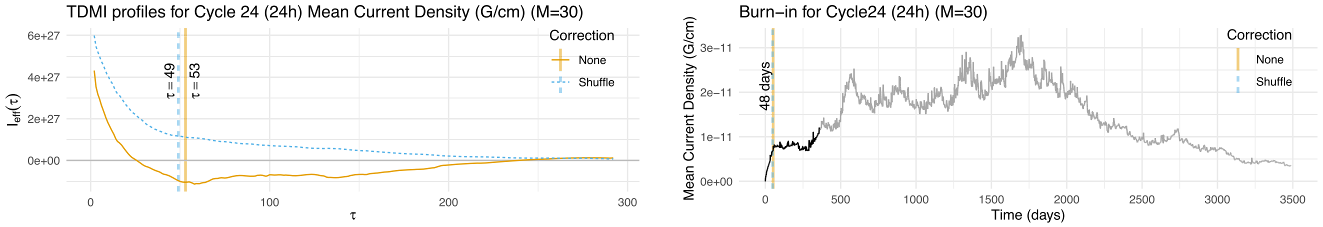

Figure 12. For Solar Cycle 24, the automated burn-in time calculated using Algorithm 2 provides a conservative estimate of when the mean current density has transitioned away from the initial potential state. (a) The TDMI profiles calculated for the mean current density with the first minimum lag (Equation (17)). (b) The mean current density with the associated burn-in time; the window (first year) used for calculating the burn-in time is the portion of the path highlighted in black. See also Table 3.

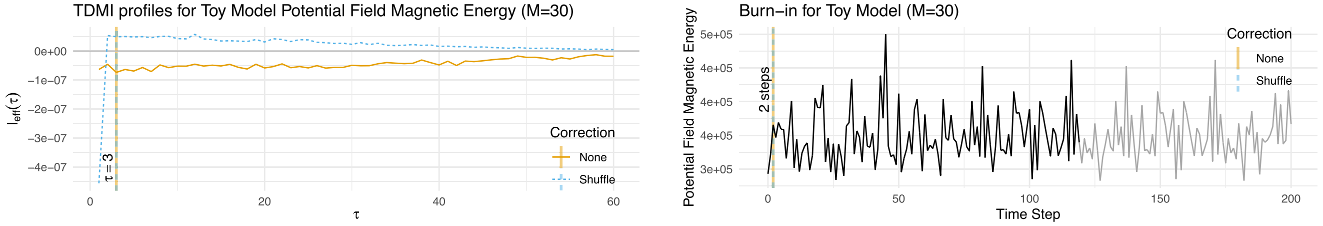

Figure 13. As expected, given the absence of a nonpotential state, Algorithm 2 produces a relatively flat TDMI profile for the toy model, with an almost instantaneous first minimum in (a) and burn-in time corresponding to the second step in (b); the window (first year) used for calculating the burn-in time is the portion of the path highlighted in black. See also Table 3.

flux, but it levels off around the peak in total absolute flux. The free magnetic energy begins to increase again at around 60 hr, this time due to a shearing motion between the positive and negative polarities of the active region during flux cancellation. In the case of AR10977, it is impossible to decouple the initial evolution of the active region from the initial potential state without additional experiments or simulations. For AR11680 in Figure 11(d), the burn-in time is approximately 1 day; for QS in Figure 11(b), it is approximately 10 minutes; and for Solar Cycle 24 in Figure 12(b), it is approximately 50 days. The burn-in time for the toy model in Figure 13(b) is instantaneous, as

Table 3
First Minimum, τ_0 (Equation (17)), and Corresponding Burn-in times, t^* (Equation (18)), for Regional, Global, and Toy Model Simulations, Using Both Uncorrected (i.e., None) and Shuffle-corrected TDMI Profiles (see Algorithms 1 and 2)

	None		Shuffle ($M = 30$)	
Simulation and Solar Quantity	τ_0	t^*	τ_0	t^*
AR10977, free magnetic energy	26	40.0 hr	34	52.8 hr
AR11680, free magnetic energy	24	23.0 hr	25	24.0 hr
QS, free magnetic energy	16	0.18750 hr	16	0.18750 hr
Toy model, potential field magnetic energy	3	2 steps	3	2 steps
Solar Cycle 24, mean current density	53	52 days	49	48 days

expected, given the absence of a nonpotential state from the white-noise construction of the "magnetogram" data for the model. Overall, the TDMI provides a systematic approach to quantifying simulation burn-in that is largely consistent with the assumed burn-in times for these regions used by other authors that are based on prior experience and knowledge of the simulation method (see, e.g., K. A. Meyer 2013; P. Pagano et al. 2019; P. Bhowmik et al. 2022).

5. Conclusion

We provide Algorithm 2 for determining burn-in times for solar coronal magnetic field simulations that are consistent with the assumed burn-in times employed in previous studies (K. A. Meyer 2013; P. Pagano et al. 2019; P. Bhowmik et al. 2022). Our algorithmic approach relies on the TDMI, the mutual information between the path-space distributions induced by lagged versions of a time series. This information-theoretic quantity is calculated using a simple plug-in estimator built from KDEs. Together with Algorithm 2, we also provide a small-sample bias correction, Algorithm 1, based on permutation resampling.

The information-theoretic measure of discrepancy in our burn-in algorithm is necessitated by the presence of long memory in solar quantities. We quantify persistence in the nonperiodic variation in time series of solar quantities, using fractionally integrated models and inference procedures popular in econometrics. We identify long memory ($d > 0.5$) in regional simulations related to AR11680 and AR10977 and in QS and global simulations related to Solar Cycle 24. The persistence analysis uses magnetofrictional simulations, a novel proxy of solar activity in the context of persistence analysis. In particular, we identify that levels of persistence in observation and simulation-based solar quantities are comparable, suggesting that magnetofrictional studies capture the memory structure in the observational data used to drive the simulations.

In the future, we plan to extend the study to consider the burn-in time for a more extensive sample of simulations. This will allow us to investigate, for example, the impact of active region evolution (e.g., flux emergence, cancellation, and rotation and shearing motions), in an effort to decouple these effects from the burn-in time calculation. In principle, this could form the basis for a rule of thumb for burn-in time in magnetofrictional simulations of the solar corona. Further, the burn-in algorithm presented here may be applied to other nonequilibrium systems, including, for example, MHD simulations, and perhaps also to simulations of open systems in molecular dynamics.

In this study, we observed that the magnetofrictional simulations (perhaps unsurprisingly) captured the memory structure present in the magnetogram data. It will be of interest to investigate the memory structure in independent observational proxies of solar activity, such as the total solar irradiance and F10.7 cm radio flux, which can be compared directly to the solar quantities derived from simulations. We noted in Remark 1 that for highly persistent times series, some estimators traditionally used in the physical sciences (such as Hurst rescaled-range analysis, log-periodogram estimators, and DFA) may provide biased estimates for persistence strength. An interesting and useful exercise would be to redo such analyses using the more robust approaches outlined here.

Acknowledgments

SOHO is a project of international cooperation between ESA and NASA. K.A.M. and A.R.Y. would like to thank the STFC for support via consortium grant Nos. ST/W001098/1 and ST/W00108X/1, respectively.

Appendix A
Autocorrelations

Autocorrelation measures the linear dependence of a time series with its own lagged values. The autocorrelation of a time series is the Pearson correlation between the series at different times as a function of the time lag:

$$\rho(h) = \mathrm{corr}(y_t, y_{t+h}) .$$

For a stationary (i.e., a weakly stationary, covariance stationary, or second-order stationary) time series y_t, the sample ACF is given by

$$\mathsf{ACF}(h) = \frac{n^{-1}\sum_{t=1}^{n-h}(y_{t+h} - \bar{y})(y_t - \bar{y})}{n^{-1}\sum_{t=1}^{n}(y_t - \bar{y})^2} , \tag{A1}$$

in terms of the lag h, where $\bar{y} = n^{-1}\sum_{t=1}^{n} y_t$ denotes the sample mean. Similarly, the partial ACF (PACF) measures the correlation of a time series, with its own lagged observations, with the linear effect of the intervening lags removed:

$$R(h, h) = \begin{cases} \mathrm{corr}(y_{t+1}, y_t) , & h = 1 , \\ \mathrm{corr}(y_{t+h} - \hat{y}_{t+h}, y_t - \hat{y}_t) , & h \geqslant 2 , \end{cases}$$

where $\hat{y}_{t+h}$ (respectively, $\hat{y}_t$) denotes the regression of y_{t+h} (resp., y_t) on $\{y_{t+h-1}, y_{t+h-2}, \ldots, y_{t+1}\}$:

$$\mathsf{PACF}(h) = \frac{\rho(h) - \sum_{k=1}^{n-1} R(n-1, k)\rho(n-k)}{1 - \sum_{k=1}^{n-1} R(n-1, k)\rho(k)} . \tag{A2}$$

The ACF and PACF can be readily estimated from data for stationary time series provided that $h \ll n$, and their calculation is automated in most statistical software packages. For stationary series, the ACF and PACF decay are related to the short-memory components in an autoregressive moving average (ARMA) model, following the Box–Jenkins approach

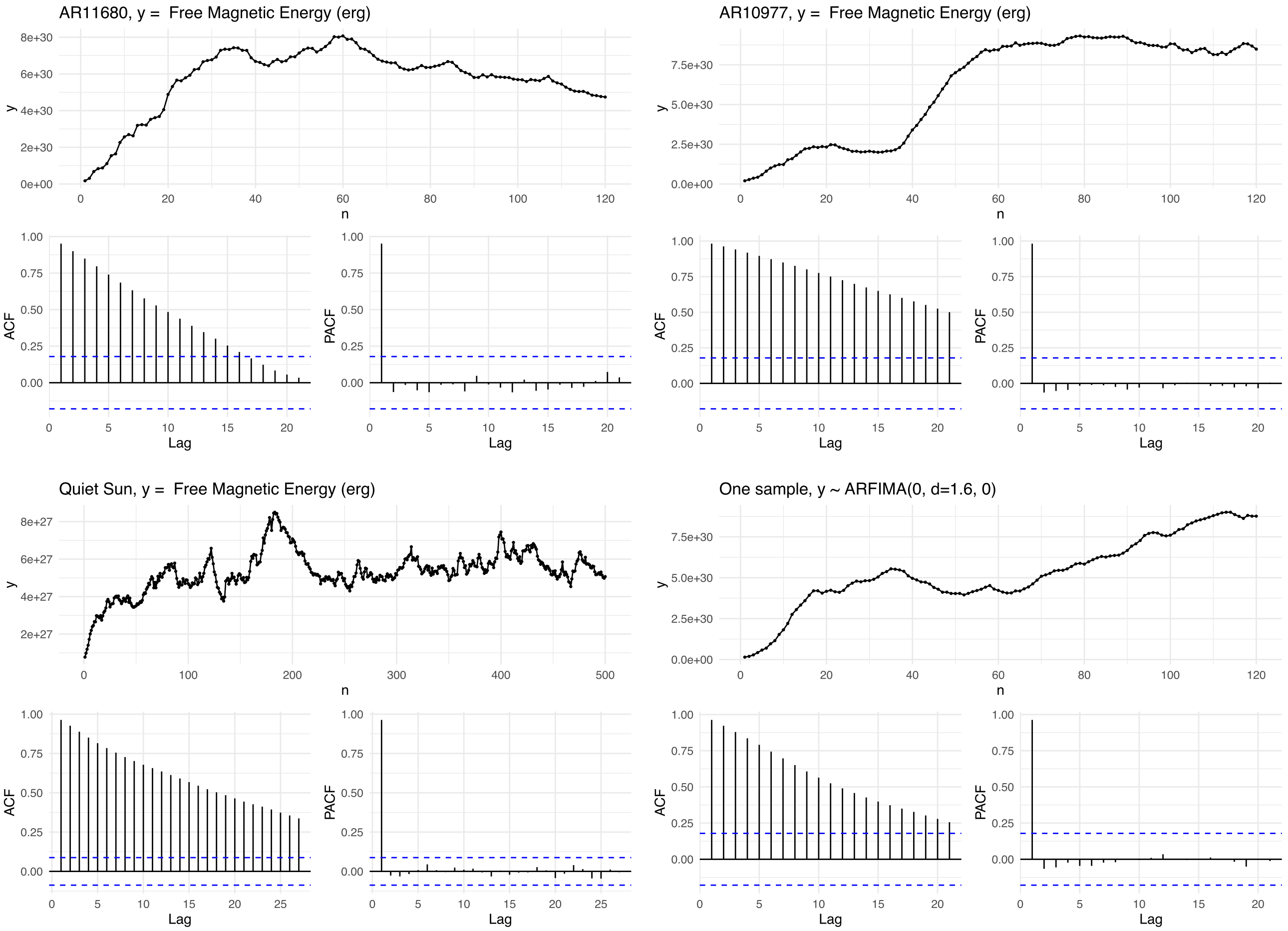

Figure 14. The slow decay of the implied ACF for the free magnetic energy in (a) AR11680, (b) AR10977, and (c) the QS provide qualitative evidence of long memory. A single time series from an $I(d = 1.6)$ model is displayed in (d) for comparison. Values outwith the white-noise-based 95% confidence bands are significantly different from zero (the bands are the normal quantile divided by the square root of the number of samples). Note that the autocorrelation is defined only for stationary processes.

to time-series analysis (G. E. P. Box et al. 2015). In Figure 14, the slow decay of the implied (or empirical) ACFs for the free magnetic energy in AR11680, AR10977, and the QS suggests that the autocorrelations persist over relatively long timescales in each simulation. In Figure 14(d), a single time series simulated from an $I(d = 1.6)$ process is plotted together with the implied ACF, to illustrate the slowly decaying autocorrelations (see the ACFs in Figures 14(a)–(c)). Note that the autocorrelation is not defined for nonstationary processes, such as $I(d)$ processes with $|d| > 0.5$ (the variance is infinite).

Appendix B
Fractionally Integrated Processes

Fractionally integrated processes are a popular class of generative time-series models with spectral densities exhibiting power-law behavior. Notably, a fractionally integrated process of order $d \in (0, 0.5)$ has the spectral density of Equation (11); see, e.g., J. Geweke & S. Porter-Hudak (1983). While long memory could theoretically be approximated by an ARMA model or ARMA(p, q) process (a perhaps more familiar generative model in time-series analysis), the high orders of p and q required would pose a challenge for parameter estimation (P. J. Brockwell & R. A. Davis 1991). Assuming that our time series are from the class of fractionally integrated candidate models, we will seek to infer the parameter d for each, thereby quantifying the nature and strength of the persistence over the timescale of the simulation.

Fractionally integrated processes are formed by fractionally integrating a stationary white noise. A fractionally integrated process $\{Y(t)\}_{t\in\mathbb{N}_0}$ of order d, denoted $Y(t) \sim I(d)$, is given by

$$Y(t) = \nabla^{-d}\ Z(t) = (1 - L)^{-d} Z(t)\,, \tag{B3}$$

where Z is white noise, i.e., $\{Z(t)\}_{t\in\mathbb{N}_0}$ is a sequence of iid zero-mean random variables with unit variance, L is the lag operator $LZ(t) = Z(t - 1)$ for $t > 1$, and the fractional integration operator is given by

$$(1 - L)^{-d} = \sum_{k=0}^{\infty} \frac{\Gamma(k + d)L^k}{\Gamma(d)\Gamma(k + 1)}\,,$$

for any $d \in (-\infty, \frac{1}{2})$, $d \neq 0, -1, -2, \ldots,$ where Γ is the Gamma function. Thus, any $Y(t) \sim I(d)$ can be expressed as a sum of white noises:

$$Y(t) = \sum_{k=0}^{\infty} \psi_k Z(t - k)\,,$$

where the coefficients $\psi_k = \Gamma(k + d)/(\Gamma(d)\Gamma(k + 1))$ are obtained from the expansion of $(1 - L)^{-d}$. In particular, we say that a process $Y(t)$ is difference-stationary if it can be fractionally differenced d times to obtain a stationary white-

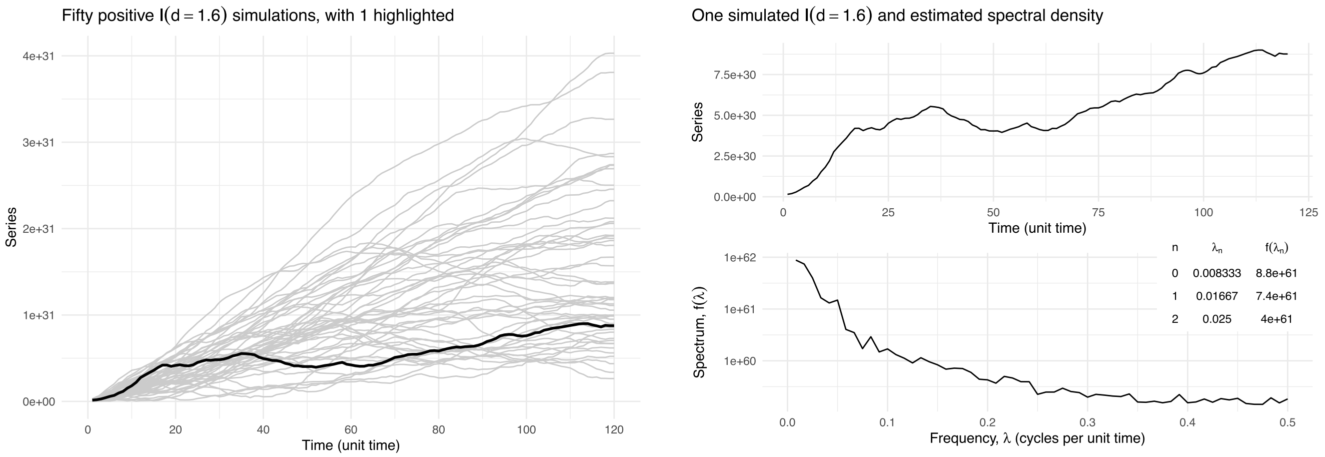

Figure 15. (a) Paths for 50 positive simulated $I(d = 1.6)$ series of length 120 demonstrate the behavior of the process. (b) For a single $I(d = 1.6)$ simulated series, observe the relative blowup of the spectral density near frequency zero, as in Equation (11); see also the similar spectral behavior near zero in Figures 6(b), (d), and (f).

noise process:

$$Z(t) = \nabla^d \quad Y(t) = (1 - L)^d Y(t) = \sum_{k=0}^{\infty} \pi_k Y(t - k) ,$$

where $\pi_k = \Gamma(k - d)/(\Gamma(- d)\Gamma(k + 1))$.

The model of Equation (B3) can be extended to all d by summing and differencing. If $X \sim I(\varphi)$, where $\varphi \in [-0.5, 0.5)$, then the process $Y(t)$ obtained by cumulative summing,

$$Y(t) = \sum_{j=0}^{t} X(j) , \quad t \in \mathbb{N}_0 ,$$

satisfies $Y(t) \sim I(d)$ for $d = \varphi + 1$, i.e., $Y(t)$ is a fractionally integrated process with order $d \in [0.5, 1.5)$. Likewise, the model can be extended to negative integer d by differencing. Consequently, for $d > 0$, we view $Y(t)$ as a fractionally integrated process, and if $d < 0$, we view it as a fractionally differenced process in this context.

The class of fractionally integrated processes has important properties depending on the parameter d. From standard facts about Gaussian distributions, for $Y(t) = \nabla^{-d}Z(t)$ with iid$Z(t) \sim \mathsf{N}(0, 1)$, we observe that if $d \in (-1, 0.5)$, then the fractionally integrated process is also Gaussian: $Y(t) \sim \mathsf{N}(0, \Gamma(1 - 2d)/\Gamma^2(1 - d))$. In J. R. M. Hosking (1981), it is shown that for $d \in (-0.5, 0.5)$, the process $Y(t) \sim I(d)$ is stationary and invertible (i.e., it can be expressed as a convergent sum of past values of the process). For $d \in (0, 0.5)$, the spectral density of $Y(t) \sim I(d)$ follows Equation (11). Processes with $d > 1$ are of integrated order, which is related to the existence of unit roots; shocks to unit root processes have more permanent effects (rather than rapid reversion to an equilibrium state), which is consistent with our understanding of the behavior of solar processes.

In Figure 15, we generate sample paths of an $I(d = 1.6)$ process with variance scaling $\sigma^2 = 10^{58}$. In Figure 15(a), we display 50 positive simulated series of length 120 to give a sense of the range of behaviors of such processes. In Figure 15(b), we display the highlighted simulated series and its periodogram to observe the blowup for frequencies near zero in the implied spectral density. In the next section, we outline inferential procedures for estimating d.

Appendix C
Unit Root Tests

The integrated order of a process can be roughly estimated by using two sequences of unit root tests with complementary hypotheses. The first test is the Augmented Dickey–Fuller, which considers the null hypothesis that the series has a unit root against a stationary alternative (D. A. Dickey & W. A. Fuller 1979; S. E. Said & D. A. Dickey 1984). The second test by Kwiatkowski, Phillips, Schmidt, and Shin (D. Kwiatkowski et al. 1992), considers the null hypothesis that the series is stationary around a deterministic trend. One then formulates a sequence of hypotheses to test the increasing order of *whole-number* differences for each series until the null hypothesis on nonstationarity is rejected and the null hypothesis on stationarity is retained, respectively. If an exact agreement is not reached, one should conservatively choose the smaller number of differences. We do not anticipate the integrated order being a whole number, but this procedure provides a systematic means of determining the number of differences to take to compare methods that are not robust against nonstationarities (without relying upon the ELW estimator or similar). This approach for estimating the integer-integrated order is automated by the `ndiffs` command in the `R` package `forecast` (R. J. Hyndman & Y. Khandakar 2008).

Appendix D
Software Utilized

The `R` packages used for memory parameter inference are summarized in Table 4.

The `R` package `ks` contains tools for univariate and multivariate KDE, including plug-in estimators for unconstrained bandwidth selectors (T. Duong 2007).

Table 4
Estimators for d

Estimator	Full Name	Indication	References	R Package	Parameters	Notes
ELW	Exact local Whittle	All d, if optimization covers interval width less than $9/2$.	K. Shimotsu & P. C. B. Phillips (2005)	`LongMemoryTS`	m	Requires mean estimation, but can use initial value.
GPH	Geweke and Porter-Hudak estimator	Based on the regression equation using the periodogram function as an estimate of the spectral density.	J. Geweke & S. Porter-Hudak (1983)	`LongMemoryTS` or `fracdiff`	m or δ	$d > 0.5$ requires differencing.
R/S-AL	Anis–Lloyd-corrected R/S statistics	R over S Hurst exponent H, with small-sample bias correction; report $d = H - 0.5$.	A. A. Annis & E. H. Lloyd (1976); R. Weron (2002)	`pracma`	Box size	$d > 0.5$ requires differencing.
DFA	Detrended fluctuation analysis	DFA α, which is related to $\alpha \approx H$ (overestimates H in nonstationary regimes); J. W. Kantelhardt et al. (2002); report $d = \alpha - 0.5$.	C. K. Peng et al. (1994)	`nonlinearTseries`	Window size range and number	DFA introduces uncontrolled bias and may be inappropriate for nonstationary series; R. M. Bryce & K. B. Sprague (2012).

Note. m is a bandwidth parameter specifying the number of Fourier frequencies used for the estimation, usually $\lfloor 1 + T^{\delta} \rfloor$, where $0 < \delta < 1$ and T is the length of the series. For DFA, 20 windows were selected from the range 10, 300.

ORCID iDs

Eric J. Hall https://orcid.org/0000-0003-0211-3335
Karen A. Meyer https://orcid.org/0000-0001-6046-2811
Anthony R. Yeates https://orcid.org/0000-0002-2728-4053

References

Adams, M., Hathaway, D. H., Stark, B. A., & Musielak, Z. E. 1997, SoPh, 174, 341
Annis, A. A., & Lloyd, E. H. 1976, Biometrika, 63, 111
Asensio Ramos, A., Cheung, M. C. M., Chifu, I., & Gafeira, R. 2023, LRSP, 20, 4
Asensio Ramos, A., Requerey, I. S., & Vitas, N. 2017, A&A, 604, A11
Aslanyan, V., Meyer, K. A., Scott, R. B., & Yeates, A. R. 2024, ApJL, 961, L3
Barczynski, K., Meyer, K. A., Harra, L. K., et al. 2022, SoPh, 297, 141
Beran, J. 1994, Statistics for Long-Memory Processes (1st ed.; Boca Raton, FL: CRC Press)
Beran, J., Feng, Y., Ghosh, S., & Kulik, R. 2013, Long-Memory Processes: Probabilistic Properties and Statistical Methods (Berlin: Springer)
Bhowmik, P., Yeates, A. R., & Rice, O. E. K. 2022, SoPh, 297, 41
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. 2015, Time Series Analysis: Forecasting and Control (5th ed.; New York: Wiley)
Brockwell, P. J., & Davis, R. A. 1991, Time Series: Theory and Methods (2nd ed.; Berlin: Springer)
Bryce, R. M., & Sprague, K. B. 2012, NatSR, 2, 315
Chacón, J. E., & Duong, T. 2010, TEST, 19, 375
Chacón, J. E., & Duong, T. 2018, Multivariate Kernel Smoothing and its Applications, Vol. 19 (Boca Raton, FL: Chapman and Hall/CRC), 375
Chodera, J. D. 2016, JCTC, 12, 1799
Cover, T. M., & Thomas, J. A. 2006, Elements of Information Theory (2nd ed.; Hoboken, NJ: Wiley-Interscience)
Dickey, D. A., & Fuller, W. A. 1979, JASA, 74, 427
Dimpfl, T., & Peter, F. J. 2012, Economic Risk Discussion Papers Sfb 649 2012- 051, Humboldt University
Ding, L., Luo, Y., Lin, Y., & Huang, Y. 2021, PhyA, 566, 125603
Domingo, V., Fleck, B., & Poland, A. I. 1995, SoPh, 162, 1
Duong, T. 2007, JSS, 21, 1
Geweke, J., & Porter-Hudak, S. 1983, J. Time Series Analysis, 4, 221
Gibb, G. P. S., Mackay, D. H., Green, L. M., & Meyer, K. A. 2014, ApJ, 782, 71
Gonzi, S., Weinzierl, M., Bocquet, F. X., et al. 2021, SpWea, 19, e02499
Gošić, M., Bellot Rubio, L. R., Cheung, M. C. M., et al. 2022, ApJ, 925, 188
Granger, C. W. J., & Joyeux, R. 1980, J. Time Series Analysis, 1, 15
Grossfield, A., Patrone, P. N., Roe, D. R., et al. 2018, Living J. Comput. Mol. Sci., 1, 5067
Hamilton, J. D. 1994, Time Series Analysis (Princeton, NJ: Princeton Univ. Press)
Hoeksema, J. T., Abbett, W. P., Bercik, D. J., et al. 2020, ApJS, 250, 28
Hosking, J. R. M. 1981, Biometrika, 68, 165
Hou, J., & Perron, P. 2014, J. Econometrics, 182, 309
Hurst, H. E. 1951, Trans. of the American Society of Civil Engineers, 116, 770
Hyndman, R. J., & Khandakar, Y. 2008, JSS, 27, 1
Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., et al. 2002, PhyA, 316, 87
Kantz, H., & Schreiber, T. 2003, Nonlinear Time Series Analysis (2nd ed.; Cambridge: Cambridge Univ. Press)
Komm, R. W. 1995, SoPh, 156, 17
Kuensch, H. R. 1987, Probability Theory and Applications, Vol. 1 (Boston, MA: De Gruyter), 67
Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. 1992, J. Econometrics, 54, 159
Leimkuhler, B., & Matthews, C. 2015, Molecular Dynamics: With Deterministic and Stochastic Numerical Methods, Vol. 39 (Berlin: Springer)
Lepreti, F., Carbone, V., & Vecchio, A. 2021, Atmos, 12, 733
Lepreti, F., Fanello, P. C., Zaccaro, F., & Carbone, V. 2000, SoPh, 197, 149
Li, S., Xiao, Y., Zhou, D., & Cai, D. 2018, PhRvE, 97, 052216
Lo, A. W. 1991, Econometrica, 59, 1279
Mackay, D. H., Green, L. M., & van Ballegooijen, A. 2011, ApJ, 729, 97
Mackay, D. H., & Upton, L. A. 2022, ApJ, 939, 9
Mackay, D. H., & Yeates, A. R. 2021, SoPh, 296, 178
Maddanu, F., & Proietti, T. 2022, SoPh, 297, 13
Mandelbrot, B. B., & Wallis, J. R. 1969a, WRR, 5, 321
Mandelbrot, B. B., & Wallis, J. R. 1969b, WRR, 5, 967
Marschinski, R., & Kantz, H. 2002, EPJB, 30, 275
McCloskey, A., & Perron, P. 2012, SSRN, doi: 10.2139/ssrn.2171907
Metcalf, T. R., Leka, K. D., Barnes, G., et al. 2006, SoPh, 237, 267
Meyer, K. A., Sabol, J., Mackay, D. H., & van Ballegooijen, A. A. 2013, ApJL, 770, L18
Ogurtsov, M. G. 2004, SoPh, 220, 93
Pagano, P., Mackay, D. H., & Yardley, S. L. 2019, ApJ, 883, 112
Pagano, P., Mackay, D. H., & Yeates, A. R. 2018, JSWSC, 8, A26
Panchev, S., & Tsekov, M. 2007, JASTP, 69, 2391
Papana, A., Kugiumtzis, D., & Larsson, P. G. 2011, PhRvE, 83, 036207
Pavliotis, G. A. 2014, Stochastic Processes and Applications, Texts in Applied Mathematics (New York: Springer)
Peng, C. K., Buldyrev, S. V., Havlin, S., et al. 1994, PhRvE, 49, 1685
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Price, D. J., Pomoell, J., Lumme, E., & Kilpua, E. K. J. 2019, A&A, 628, A114
Rahman, S., Shin, S., Jeong, H.-J., et al. 2023, ApJ, 948, 21
Rice, O. E. K., & Yeates, A. R. 2022, FrASS, 9, 849135
Rice, O. E. K., & Yeates, A. R. 2023, ApJ, 955, 114
Robinson, P. M. 1995a, AnSta, 23, 1048
Robinson, P. M. 1995b, AnSta, 23, 1630
Rodkin, D., Goryaev, F., Pagano, P., et al. 2017, SoPh, 292, 90
Ruzmaikin, A., Feynman, J., & Robinson, P. 1994, SoPh, 149, 395
Ruzmaikin, A. A., Feynman, J., Neugebauer, M., & Smith, E. J. 2000, AAS/Solar Physics Division Meeting, 31, 04.04
Rypdal, K., & Rypdal, M. 2012, in Multi-scale Dynamical Processes in Space and Astrophysical Plasmas, ed. L. M. P. & Z. Vörös (Berlin: Springer), 227
Rypdal, M., & Rypdal, K. 2012, JGRA, 117, A04103
Said, S. E., & Dickey, D. A. 1984, Biometrika, 71, 599
Salcedo-Sanz, S., Casillas-Pérez, D., Del Ser, J., et al. 2022, PhR, 957, 1
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129
Schou, J., Bush, R. I., Schou, J., et al. 2012, SoPh, 275, 229
Schrijver, C. J., De Rosa, M. L., Metcalf, T. R., et al. 2006, SoPh, 235, 161
Sheeley, N. R., Jr 2005, LRSP, 2, 5
Shimotsu, K. 2010, Econometric Theory, 26, 501
Shimotsu, K., & Phillips, P. C. B. 2005, AnSta, 33, 1890
Song, P., & Russell, C. T. 1999, SSRv, 87, 387
Staniek, M., & Lehnertz, K. 2008, PhRvL, 100, 158101
Sun, X. 2018, arXiv:1801.04265
Thalmann, J. K., Gupta, M., & Veronig, A. M. 2022, A&A, 662, A3
van Ballegooijen, A. A., & Cranmer, S. R. 2008, ApJ, 682, 644
van Ballegooijen, A. A., Priest, E. R., & Mackay, D. H. 2000, ApJ, 539, 983
Vastano, J. A., & Swinney, H. L. 1988, PhRvL, 60, 1773
Wand, M. P., & Jones, M. C. 1993, JASA, 88, 520
Weron, R. 2002, PhyA, 312, 285
Wiegelmann, T., & Sakurai, T. 2021, LRSP, 18, 1
Yardley, S. L., Mackay, D. H., & Green, L. M. 2018, ApJ, 852, 82
Yardley, S. L., Mackay, D. H., & Green, L. M. 2021, SoPh, 296, 10
Yeates, A. R. 2014, SoPh, 289, 631
Yeates, A. R. 2024, SoPh, 299, 83
Yeates, A. R., & Bhowmik, P. 2022, ApJ, 935, 13
Yeates, A. R., Cheung, M. C. M., Jiang, J., Petrovay, K., & Wang, Y.-M. 2023, SSRv, 219, 31
Yeates, A. R., Mackay, D. H., & van Ballegooijen, A. A. 2008, SoPh, 247, 103
Zuccarello, F. P., Aulanier, G., & Gilchrist, S. A. 2015, ApJ, 814, 126
Zuccarello, F. P., Pariat, E., Valori, G., & Linan, L. 2018, ApJ, 863, 41