An Adaptive Markov Chain Approach for Probabilistic Forecasting of Categorical Events

EDWARD C. D. POPE

Met Office, Exeter, United Kingdom

DAVID B. STEPHENSON

Department of Mathematics, University of Exeter, Exeter, United Kingdom

DAVID R. JACKSON

Met Office, Exeter, United Kingdom

(Manuscript received 29 July 2019, in final form 13 May 2020)

ABSTRACT

Categorical probabilistic prediction is widely used for terrestrial and space weather forecasting as well as for other environmental forecasts. One example is a warning system for geomagnetic disturbances caused by space weather, which are often classified on a 10-level scale. The simplest approach assumes that the transition probabilities are stationary in time—the homogeneous Markov chain (HMC). We extend this approach by developing a flexible nonhomogeneous Markov chain (NHMC) model using Bayesian nonparametric estimation to describe the time-varying transition probabilities. The transition probabilities are updated using a modified Bayes's rule that gradually forgets transitions in the distant past, with a tunable memory parameter. The approaches were tested by making daily geomagnetic state forecasts at lead times of 1–4 days and were verified over the period 2000–19 using the rank probability score (RPS). Both HMC and NHMC models were found to be skillful at all lead times when compared with climatological forecasts. The NHMC forecasts with an optimal memory parameter of ~100 days were found to be substantially more skillful than the HMC forecasts, with an RPS skill for the NHMC of 10.5% and 5.6% for lead times of 1 and 4 days ahead, respectively. The NHMC is thus a viable alternative approach for forecasting geomagnetic disturbances and could provide a new benchmark for producing operational forecasts. The approach is generic and is applicable to other forecasts that include discrete weather regimes or hydrological conditions (e.g., wet and dry days).

1. Introduction

In developing physical models for forecasting complex systems, it is often useful to be able to benchmark skill against that which can be obtained using empirically based statistical schemes. Furthermore, since the future states are not perfectly known, it is important for decision-making to communicate uncertainty in these forecasts. One way to do this is to issue probabilities of future states, for example, by running an ensemble of forecasts using a numerical weather/climate model (e.g., Arribas et al. 2011; MacLachlan et al. 2015; Swinbank et al. 2016), or alternatively by fitting an appropriate statistical model to past data [e.g., for solar flare forecasts (Bloomfield et al. 2012); for forecasts of relativistic electrons at geostationary orbits (Baker et al. 1990; Boynton et al. 2016)].

Motivated by these needs for space weather, this study presents a novel statistical approach for issuing probabilistic forecasts of categorical events and will demonstrate it by application to the problem of daily forecasting geomagnetic intensity classified into five levels of intensity. The simplest statistical approach would be to issue probabilities for each category that do not change in time; for example, one could estimate such probabilities by simply using the climatological long-run frequencies of occurrence of past events. However, this neglects dependency between states in successive days—tomorrow's state is generally expected to have some dependence on the states that have just occurred. One widely used approach for capturing such

Corresponding author: Edward C. D. Pope, edward.pope@metoffice.gov.uk

DOI: 10.1175/MWR-D-19-0239.1

conditional dependence is to use a stochastic model known as a Markov chain. A *k*th-order Markov chain assumes that the probability of the next state is a deterministic function of what states have occurred in the past *k* days (e.g., Wilks 2006). So, for example, a first-order Markov chain for a process with five states issues a different set of five probabilities for tomorrow depending upon which one of the five states occurred today (the probabilities can be conveniently represented in a 5×5 transition probability matrix). Markov chains have been widely used in meteorology—for example, Markov chain modeling of wet- and dry-day rainfall sequences (e.g., the early studies of Gabriel and Neumann 1962; Caskey 1963; Gates and Tong 1976).

In most Markov chain applications, it is assumed for simplicity that the transition probabilities are homogeneous (i.e., do not change) in time—these are referred to as homogeneous Markov chain (HMC) models. However, such a stationarity assumption is not justifiable for the majority of meteorological phenomena that are known to behave quite differently at different times (e.g., in different seasons) and in different phases of modes/cycles, etc. A few studies have attempted to address this by fitting nonhomogeneous Markov chain (NHMC) models. For example, Woolhiser and Pegram (1979) accounted for seasonal variations in the daily transition probabilities of a two-state first-order precipitation forecast by expressing the time dependence throughout the year as a Fourier series. In contrast, Rajagopalan et al. (1996) used a nonparametric kernel smoothing approach to estimate unknown changes in daily transition probabilities. Paulo and Pereira (2007) used both an HMC and an NHMC (with different transition matrix for each month) to forecast changes in the severity of droughts in Southern Portugal. Our study goes further by proposing a simple yet more rigorous nonparametric NHMC that is based on an autoregressive modification to Bayesian updating.

The remainder of this article is structured as follows. Section 2 provides more detail about how geomagnetic storm forecasts are produced, and the limitations of the climatology of observations used to construct the statistical models. Section 3 proposes a sequential updating scheme for calculating a one-step NHMC. Section 4 assesses its forecast skill relative to both a one-step homogeneous Markov chain and climatology.

2. Application: Probabilistic forecasts of geomagnetic disturbance categories

The Met Office Space Weather Operations Centre (MOSWOC) provides a range of operational space weather forecasts including arrival times of coronal mass ejections (CMEs), and probabilistic forecasts for relativistic electron fluences (time-integrated flux), proton and X-ray fluxes, and geomagnetic disturbance indices. Geomagnetic atmospheric disturbances can have severe impacts on critical technical infrastructure [including positioning with the global positioning system positioning, navigation and timing, satellite electronics, radio communications, and the electricity grid], and are quantified using a range of indices/scales. Many operational space weather centers use the planetary *K*-scale (K_p), which is derived from the maximum fluctuations of magnetic field horizontal components observed by a magnetometer during a 3-h interval (e.g., Bartels et al. 1939). The planetary 3-h-range index K_p is the mean standardized *K*-index from 13 geomagnetic observatories between 44° and 60° northern or southern geomagnetic latitude. The scale is ''quasi logarithmic'' and ranges from 0 to 9, with 0 being no disturbance and values greater than 5 indicating the occurrence of a geomagnetic storm. For this reason, MOSWOC characterizes geomagnetic storms using the National Oceanic and Atmospheric Administration (NOAA) G-scale (i.e., $G = K_p - 4$) in which G 1, G 2, G 3, G 4, and G 5 are used to signify $K_p = 5$, $K_p = 6$, $K_p = 7$, $K_p = 8$, and $K_p = 9$, respectively (http://www.swpc.noaa.gov/noaa-scales-explanation). To reduce the number of states, MOSWOC combines G 1 and G 2 events into a G 1/2 category. In this work, we also include the conjugate probability to the G 1/2 category, which is the probability that the disturbance does not exceed the G 1 threshold (i.e., <G 1), effectively making a five-category forecast. Figure 1 shows the time series of daily maximum geomagnetic disturbance from January 1998 to March 2019, as measured on the G-scale, with the corresponding histogram showing the climatological frequency of each category on a logarithmic scale. It is evident that the vast majority of days are in the low activity state (<G 1), and that the more excited states are increasingly rare, with only a few G 5 events recorded since 1998.

Geomagnetic disturbances are the result of the physical interaction between the extreme driving of solar plasma (i.e., the solar wind), arising chiefly from CMEs, and Earth's magnetic field. It is possible to predict the passage of CMEs (and other changes in the solar wind) from Sun to Earth. However, because of our limited understanding of the interaction between the solar wind and Earth's geomagnetic field, there are currently no physics-based numerical models that can be used to accurately predict the occurrence of significant geomagnetic disturbances. It is therefore necessary for prediction of these disturbances to make use of empirical statistical approaches based on past data. Inspection

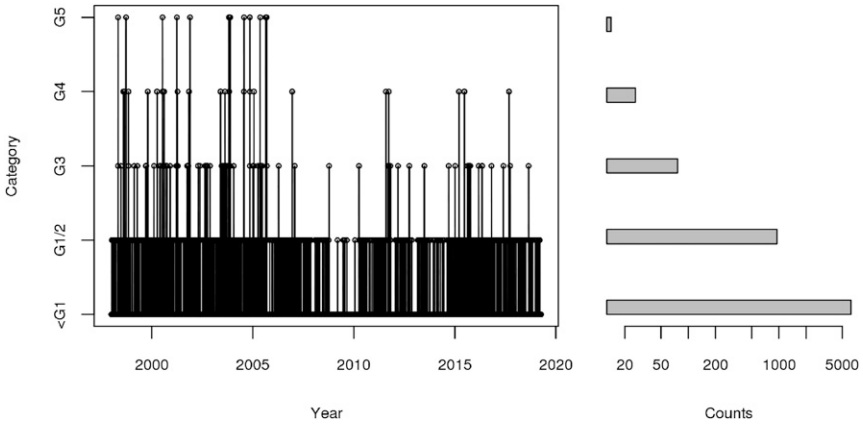

FIG. 1. (left) Time series of maximum geomagnetic disturbance each day from January 1998 to March 2019, as measured on the G-scale. (right) Histogram showing the corresponding total counts (on a logarithmic scale) in each category.

of the geomagnetic disturbance data suggests that the values on a given day are conditional on the previous day. For example, fluctuations in the geomagnetic field resulting from the arrival of an Earth-directed CME can take at least a day to relax. This behavior suggests it may be possible to produce skillful statistical forecasts based on current conditions and an understanding of the likelihood of the system transitioning between different states. For this reason, statistical forecasting of geomagnetic disturbances suggests a Markov modeling approach. Currently, methods for producing statistical geomagnetic storm forecasts include Auto-Regressive Moving Average (ARMA) models (e.g., Thomson et al. 2001) and neural networks (NN) (e.g., Thomson et al. 1995; Thomson 1996; Boberg et al. 2000).

As described above, MOSWOC issues forecasts for the probability of exceeding each of the following four G-scale thresholds: G 1/2, G 3, G 4, and G 5. These categorical forecasts are produced every day for the next 4 days and are based on solar wind prediction models, existing ARMA or NN forecasts (mentioned above), but also rely heavily on the forecasters' experience and knowledge. Sharpe and Murray (2017) explored MOSWOC forecast performance during 2015–16. The MOSWOC forecasts for 1, 3 and 4 days ahead usually performed better than a benchmark forecast based on a rolling climatology from the previous 180 days, and 2-day-ahead forecasts usually performed slightly worse than the benchmark. However, the difference between the MOSWOC and benchmark forecasts, or between the Day 1–4 MOSWOC forecasts, was usually not statistically significant. Future work will explore the verification of MOSWOC geomagnetic disturbance forecast relative to a range of benchmarks, building on the work of Sharpe and Murray (2017).

High intensity disturbances are generally more frequent during the solar maximum than near the solar minimum, although the times of peak occurrence are not perfectly aligned with the solar cycle (see Fig. 2). The main physical reason for this relationship is that CMEs are strongly modulated by the ~11-yr solar cycle. Phases of the solar cycle are identifiable by numbers of sunspots on the surface of the Sun, with the solar maximum corresponding to larger numbers and vice versa. Direct observations of sunspots since the 1700s and indirect estimates based on dendroclimatology indicate substantial solar variability from intercycle (i.e., 11-yr) time scales to millennia and beyond (e.g., Owens et al. 2017). Recent observations indicate that adjacent solar cycles can be sufficiently different from each other that averaging over only a few cycles can produce misleading results when applied to any individual cycle. Accordingly, modeling the solar cycle parametrically using a particular mathematical function is not guaranteed to produce optimal results. It is, therefore, necessary to take account of this nonstationarity in a more adaptive nonparametric way. For clarity, we note that Fig. 2 demonstrates the existence of natural variability in the occurrence of geomagnetic disturbances and its association with sunspot numbers—we do not use sunspots to predict geomagnetic disturbances.

3. Markov Chain modeling

This section describes the key concepts and notation in first-order Markov chain modeling and then introduces our new nonhomogeneous approach.

a. *First-order Markov Chain models*

Consider a system that at any time t can be in any of $J > 1$ possible states $\{1, 2, \ldots, J\}$ denoted by variable

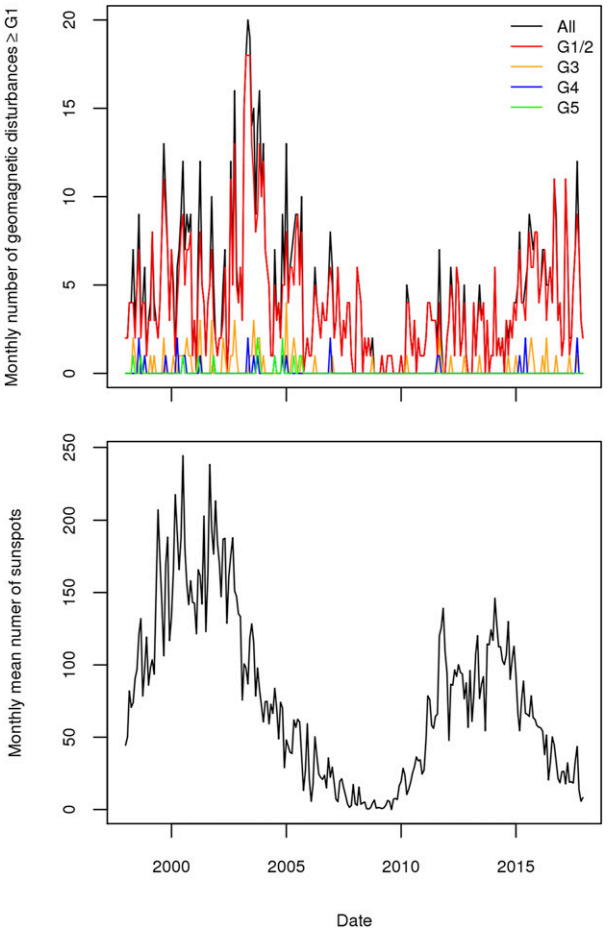

FIG. 2. Comparison of (top) the monthly number of geomagnetic disturbances (of all magnitudes), and (bottom) the mean number of sunspots each month from January 1998 to May 2018 (https://solarscience.msfc.nasa.gov/SunspotCycle.shtml).

$X(t)$. Following Sung (2014), let the $J \times J$ random matrix $\mathbf{N}(t)$ denote the number of transitions between states of this single chain at time $t-1$ and t, [i.e., $N_{ij}(t) = 1$ when $X(t-1) = i$ and $X(t) = j$, and 0 otherwise]. The expectation of random matrix $\mathbf{N}(t)$ gives the $J \times J$ transition probability matrix $\mathbf{P}(t)$ that contains conditional probabilities $P_{ij}(t) = \Pr[X(t) = j | X(t-1) = i]$. A first-order Markov chain model assumes that the conditional probability of $X(t) = j$ given *any* past sequence is only determined by the previous state; that is, $\Pr[X(t) = j | X(t-1), X(t-2), \ldots] = P_{X(t-1)j}(t)$. Homogeneous Markov chain models assume that $\mathbf{P}(t) = \mathbf{P}$ is constant in time, whereas nonhomogeneous chain models allow $\mathbf{P}(t)$ to vary in time. The mth-step-ahead probability forecast of the states of $X(t+m-1)$ made at time $t-1$ is given by the $X(t-1)$ th row of $\mathbf{P}^m(t)$ (the mth power of the transition matrix at time t).

b. *Frequentist estimation of transition probabilities*

The simplest approach for estimating $\mathbf{P}$ for homogeneous chains is to approximate the expectation of the transition counts by the long-term sample mean of all previously observed transitions:

$$\hat{\mathbf{P}}(t) = \frac{1}{t-1} \sum_{t'=1}^{t-1} \mathbf{N}(t'). \tag{1}$$

Figure 3 shows these relative frequencies of transitions between all possible pairs of the five states on the daily G-scale observations from 1998 to 2019. These estimates of daily transition probabilities demonstrate that all states are accessible, and that all transitions have been observed since 1998. However, the daily probability of the system transitioning out of the inactive (<G 1) state is small, being ~0.1, with decreasing probabilities of transitioning to more excited states (i.e., larger upward transitions have a correspondingly lower probability). Furthermore, the probability of the system staying in the same state tomorrow as today also decreases rapidly in the more active states (e.g., 0.071 for G 5 vs 0.38 for G 1/2). For the active states, the daily probability of transitioning to a less excited state exceeds the probability of transitioning to a more excited state.

This frequentist approach has the disadvantage that it can produce zero probabilities (i.e., impossible transitions) for states where no transitions have yet been observed (e.g., between rare states). It also ignores any prior beliefs one may have about the transition matrix before observing the data. For example, one may believe that transitions to more extreme states should be less likely than transitions to less extreme states. To incorporate such beliefs, one can use a Bayesian estimation approach.

c. *Bayesian updating for a homogeneous processes*

By use of Bayes's theorem, observed values $\tilde{\mathbf{n}}_i$ of the transition vector $\tilde{\mathbf{N}}_i$ can be used to sequentially update uncertain beliefs about probabilities $\tilde{\mathbf{P}}_i$. Suppose, without loss of generality, that the system is in state i at time $t-1$ [i.e. $X(t-1) = i$]. The number of transitions to other states in the next time step is then given by random J-vector $\tilde{\mathbf{N}}_i(t) = (N_{i1}, N_{i2}, \ldots, N_{iJ})$, which may be considered to be a single random draw from the multinomial distribution $\tilde{\mathbf{N}}_i(t) \sim \text{Multi}[\tilde{\mathbf{P}}_i(t)]$ having probability mass function $\Pr(\tilde{\mathbf{N}}_i = \mathbf{n}) = P_{i1}^{n_1} P_{i2}^{n_2} \ldots P_{iJ}^{n_J}$ (a generalization of the Bernoulli distribution to $J > 2$ categories). If one then assumes that the prior probabilities have a Dirichlet distribution (Kotz et al. 2000), the posterior distribution – the product of the prior and the likelihood–is also Dirichlet distributed:

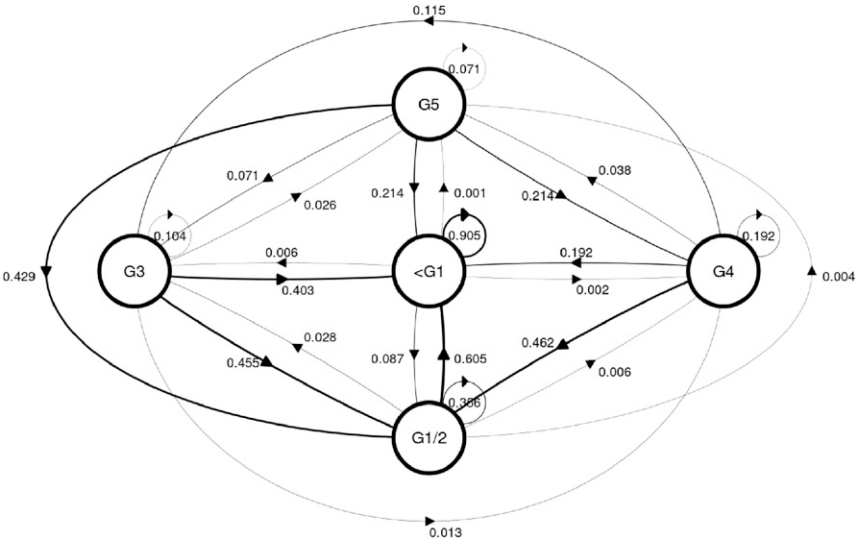

FIG. 3. Transition probability to and from each state from one day to the next, estimated from daily observations (January 1998–March 2019). To help to illustrate the most dominant transitions, the line thickness is proportional to the probability.

$$f(\tilde{\mathbf{P}}_i) \propto P_{i1}^{a_{i1}-1} P_{i2}^{a_{i2}-1} \dots P_{iJ}^{a_{iJ}-1} \quad \text{prior}$$

$$\Pr(\tilde{\mathbf{N}}_i = \mathbf{n}_i | \tilde{\mathbf{P}}_i) = P_{i1}^{n_{i1}} P_{i2}^{n_{i2}} \dots P_{iJ}^{n_{iJ}} \quad \text{likelihood}$$

$$\rightarrow$$

$$f(\tilde{\mathbf{P}}_i | \tilde{\mathbf{N}}_i = \mathbf{n}) \propto P_{i1}^{a_{i1}-1+n_{i1}} P_{i2}^{a_{i2}-1+n_{i2}} \dots P_{iJ}^{a_{ij}-1+n_{iJ}} \quad \text{posterior}.$$

In other words, the Dirichlet distribution is the conjugate prior for the multinomial distribution (Sung 2014). This Bayesian updating can be written very simply as $\mathbf{a}(t+1) = \mathbf{a}(t) + \mathbf{n}(t)$, where the $J \times J$ matrix $\mathbf{a}$ consists of Dirichlet coefficient elements $\{a_{ij}\}$. It represents a random walk in Dirichlet coefficients with discrete nonnegative steps determined by the transitions between successive days.

The distribution of the transition probability P_{ij} from state i to state j is the marginal distribution of the Dirichlet distribution, which is known to be the beta distribution [i.e., $P_{ij} \sim \text{beta}(a_{ij}, s_i - a_{ij})$, where $s_i = \sum_j a_{ij}$ is the sum over column elements for row i]. The expectation of this distribution is $E(P_{ij}) = a_{ij}/s_i$, which converges to the long-term-relative frequency after many transitions from state i. The variance given by

$$\text{Var}(P_{ij}) = a_{ij}(s_i - a_{ij})/s_i^2(s_i + 1)$$

converges to zero (i.e., perfectly precise estimates) after many transitions since the denominator increases faster than the numerator. Confidence intervals on transition probabilities can be obtained by considering quantiles of the beta distributions described above (e.g., as shown later in Fig. 5).

d. Bayesian updating for nonhomogeneous processes

Bayesian updating is unsuitable for nonhomogeneous processes since, as more observations are made, the sums for each row of the $\mathbf{a}$ matrix increase indefinitely and so become dominated by past transitions rather than being able to adapt to recent changes in transition probabilities. One pragmatic approach for avoiding this problem is to include a "discounting" factor λ, which allows older data to be exponentially down-weighted (Bertuccelli and How 2008). For example, one could modify the updating equation to be the first-order autoregressive process $\mathbf{a}(t+1) = \lambda \mathbf{a}(t) + \mathbf{n}(t)$ where $\lambda = e^{-1/\tau}$ and $\tau > 0$ is an e-folding memory time scale (in days). When there are no transitions from a state, the row of $\mathbf{a}$ for that state relaxes to the null vector. More generally, one could consider relaxing back to any preferred reference beliefs $\mathbf{a_0}$ using

$$\mathbf{a}(t+1) - \mathbf{a}_0 = \lambda[\mathbf{a}(t) - \mathbf{a}_0] + \mathbf{n}(t). \quad (2)$$

TABLE 1. Transition probabilities estimated from the relative frequencies of transitions between observed daily maximum geomagnetic disturbance data over the whole period 1998–2019.

		Today				
		<G 1	G 1/2	G 3	G 4	G 5
Tomorrow	<G 1	0.905	0.605	0.403	0.192	0.214
	G 1/2	0.087	0.356	0.455	0.462	0.429
	G 3	0.006	0.028	0.104	0.115	0.071
	G 4	0.002	0.006	0.013	0.192	0.214
	G 5	0.001	0.004	0.026	0.038	0.071

TABLE 2. Climatological probabilities for each G state estimated from the relative frequencies of transitions over the whole period 1998–2019.

<G 1	G 1/2	G 3	G 4	G 5
0.858	0.127	0.0099	0.0033	0.0018

In what follows, we have chosen the reference to be a multiple of the climatological probabilities $\hat{\mathbf{p}}$ for the states estimated by fitting a homogeneous Markov chain to the entire dataset (1998–2019), that is, $\mathbf{a}_0 = \kappa(\hat{\mathbf{p}}, \hat{\mathbf{p}}, \hat{\mathbf{p}}, \hat{\mathbf{p}}, \hat{\mathbf{p}})^{\mathrm{T}}$. In the limit that the dimensionless $\kappa \to 0$, the reference state also tends to zero, while for large κ (>1000) the daily updates to the transition probability are fractionally very small. Experimentation shows that $\kappa \sim 10$ allows the nonhomogeneous Markov chain to behave similarly to the homogenous Markov chain if no transitions have been observed for a long time, but with the freedom to adapt to time-varying transition probabilities by exponentially weighting past transitions.

4. Results

The method outlined above has been applied to forecasting daily G-scale data observed over the period 1998–2019 [the data are freely available online from the NOAA Space Weather Prediction Center (ftp://ftp.swpc.noaa.gov/pub/indices/old_indices/)]. This period experienced a wide range of geomagnetic states since it contained a substantial fraction of the previous solar cycle 23 (1996–2008) and the current solar cycle 24 (from 2008 to the present).

Table 1 shows the HMC transition probabilities estimated over the entire period 1998–2019. Note that the probability of a transition from a low to a high disturbance state is small (e.g., the daily probability of a transition from <G 1 to G 3 is ~0.006). Furthermore,

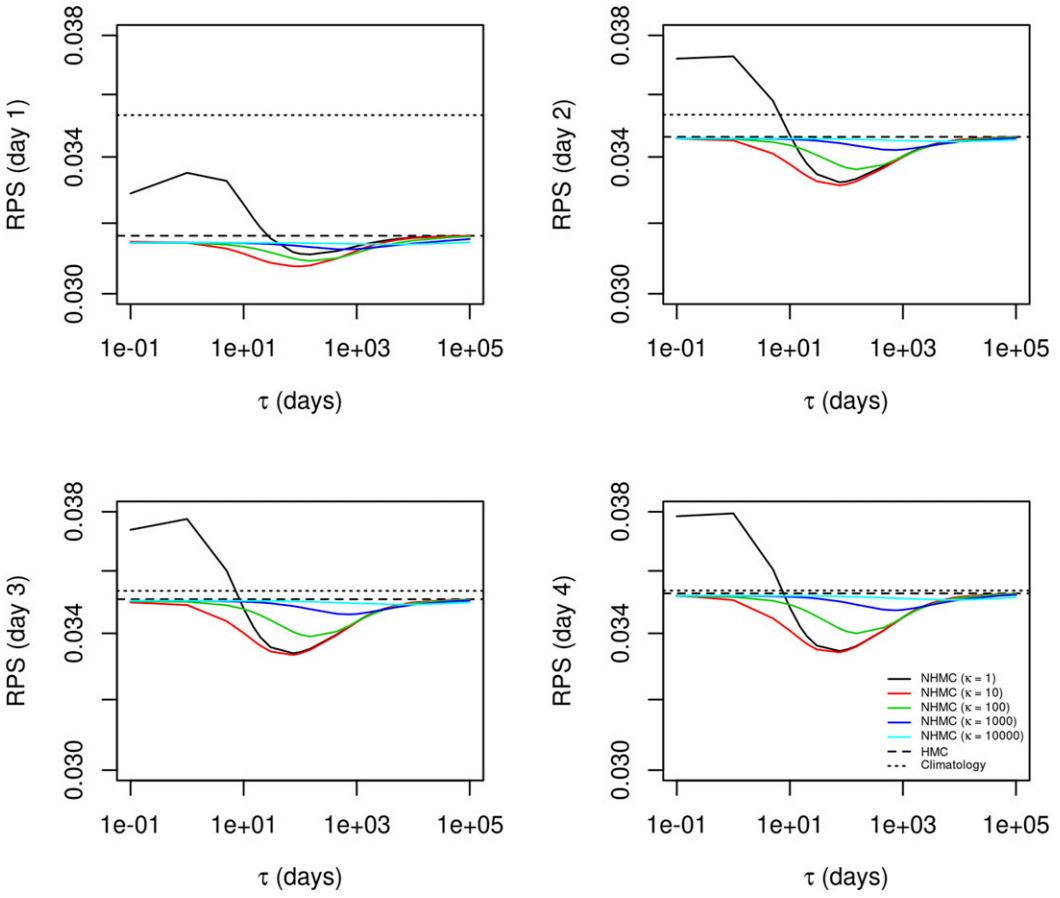

FIG. 4. Variation of RPS with τ and κ for lead times of 1–4 days, comparing NHMC with both HMC and climatology during the validation period (2000–19). The lowest RPS for the NHMC occurs for $\kappa \sim 10$ at all lead times, and for $\tau \sim 100$ days at a lead time of 1 day, with $\tau \sim 70$ days for a 2-day lead time, and $\tau \sim 60$ days for lead times of 3 and 4 days. For both larger and smaller values of τ, the performance of the NHMC tends asymptotically toward the HMC.

TABLE 3. Rank probability score and rank probability skill score for lead times of 1–4 days across the verification period (2000–19). The predictions from the HMC and NHMC (τ = 100 days; κ = 10) models are compared with climatology.

	Climatology	HMC	HMC RPSS	NHMC RPS	NHMC RPSS
Day 1	0.0353	0.0316	10.5%	0.0307	13.0%
Day 2	0.0353	0.0346	1.98%	0.0331	6.2%
Day 3	0.0353	0.0351	0.57%	0.0334	5.4%
Day 4	0.0354	0.0353	0.28%	0.0334	5.6%

the highest transition probabilities are those from a high state to the lowest state; for example, the daily probability of a transition from G 5 to G 1/2 is ~0.43. Table 2 gives the climatological probabilities of the states estimated by the relative frequencies of each category over the whole period. The probabilities indicate that nearly 99% of the days were in the lowest two categories.

The parameters in the HMC and NHMC models were sequentially updated each day. The HMC transition probabilities were updated daily by taking the time mean of all previous observed transitions since 1 January 1998. The NHMC model was updated daily using Eq. (2) starting from an initial state of **a** that had all elements set to 1 on 1 January 1998. The reference state was taken to be a multiple of the climatological probabilities of the states given in Table 2. Daily probability forecasts of states 1, 2, 3, and 4 days ahead issued by the two models were evaluated using the ranked probability score (RPS; Jolliffe and Stephenson 2011) over a validation period from 1 January 2000 to 31 March 2019, which allowed for a short burn-in training period from 1 January 1998 to 31 December 1999. The best values of parameters τ and κ for the NHMC model were obtained by finding those that gave smallest RPS calculated over the validation period. The use of data from the validation period to train the NHMC model potentially leads to an overestimate in the evaluated performance of this model. Although out-of-sample validation or cross validation might appear to be fairer, it is nonetheless problematic to implement because the NHMC model relies upon capturing the serial dependency of past states in order to estimate evolving transition probabilities. It would, therefore, be of interest in future studies to try to develop more objective out-of-sample procedures for estimating these two parameters.

Figure 4 shows the RPS for forecasts at each lead time as a function of τ and κ. For 1-day ahead forecasts, there is a clear minimum in RPS at $\tau \sim 100$ days and $\kappa \sim 10$, and for larger τ values the NHMC performance tends toward more homogeneous behavior as to be expected. More generally, the optimal value of τ appears to be positively correlated with κ, and dependent on lead time. For lead times of 2–4 days, small κ results in similar optimal values of τ and a comparable RPS. Experimentation shows that τ = 100 days and κ = 10 appear to be roughly optimal in allowing the NHMC to behave similarly to the HMC if no transitions have been observed for a long time, but with the freedom to adapt to time-varying transition probabilities. The following results were obtained with this choice of parameter values.

The performance of the forecasts is summarized by the RPS values and RPS skill scores (RPSS) given in Table 3, which gives the RPS values over the validation period of the HMC and NHMC models, and compares them with RPS values of no-skill climatological forecasts (made by issuing constant probabilities for each state) to obtain skill scores. The scores are naturally small because ~86% of the days are spent in the lowest state (<G 1). Nevertheless, Table 3 and Fig. 4 demonstrate that the HMC is more skillful than climatology for lead times of 1 day, with a corresponding RPS skill of $100 \times (1 - 0.0316/0.0353) = 10.5\%$; the HMC performance then rapidly converges to climatology at longer lead times. The NHMC outperforms the HMC at all lead times and remains skillful at longer lead times, with an RPS skill at day 4 of $100 \times (1 - 0.0334/0.0354) = 5.6\%$.

We use the statistical test developed by Diebold and Mariano (1995) to compare the predictive accuracies of the three forecast models (i.e., NHMC, HMC and climatology), following the procedure outlined by Gilleland and Roux (2015). The Diebold and Mariano test is appropriate because it makes no assumptions about the distribution of forecast errors, and incorporates temporal autocorrelations, and can be applied even when the forecast models are correlated with each other. In this case, we compare the daily RPS values across the validation period (2000–19) for

TABLE 4. The p values for the Diebold and Mariano (1995) test comparing daily RPS values for each pair of forecast models, at lead times of 1–4 days across the verification period (2000–19). The predictions from the HMC and NHMC use τ = 100 days and κ = 10.

	NHMC/HMC	NHMC/Climatology	HMC/Climatology
Day 1	4.12×10^{-6}	1.08×10^{-11}	1.25×10^{-11}
Day 2	1.19×10^{-6}	5.63×10^{-7}	2.00×10^{-4}
Day 3	1.43×10^{-6}	8.03×10^{-7}	2.76×10^{-4}
Day 4	1.43×10^{-6}	1.62×10^{-6}	1.38×10^{-3}

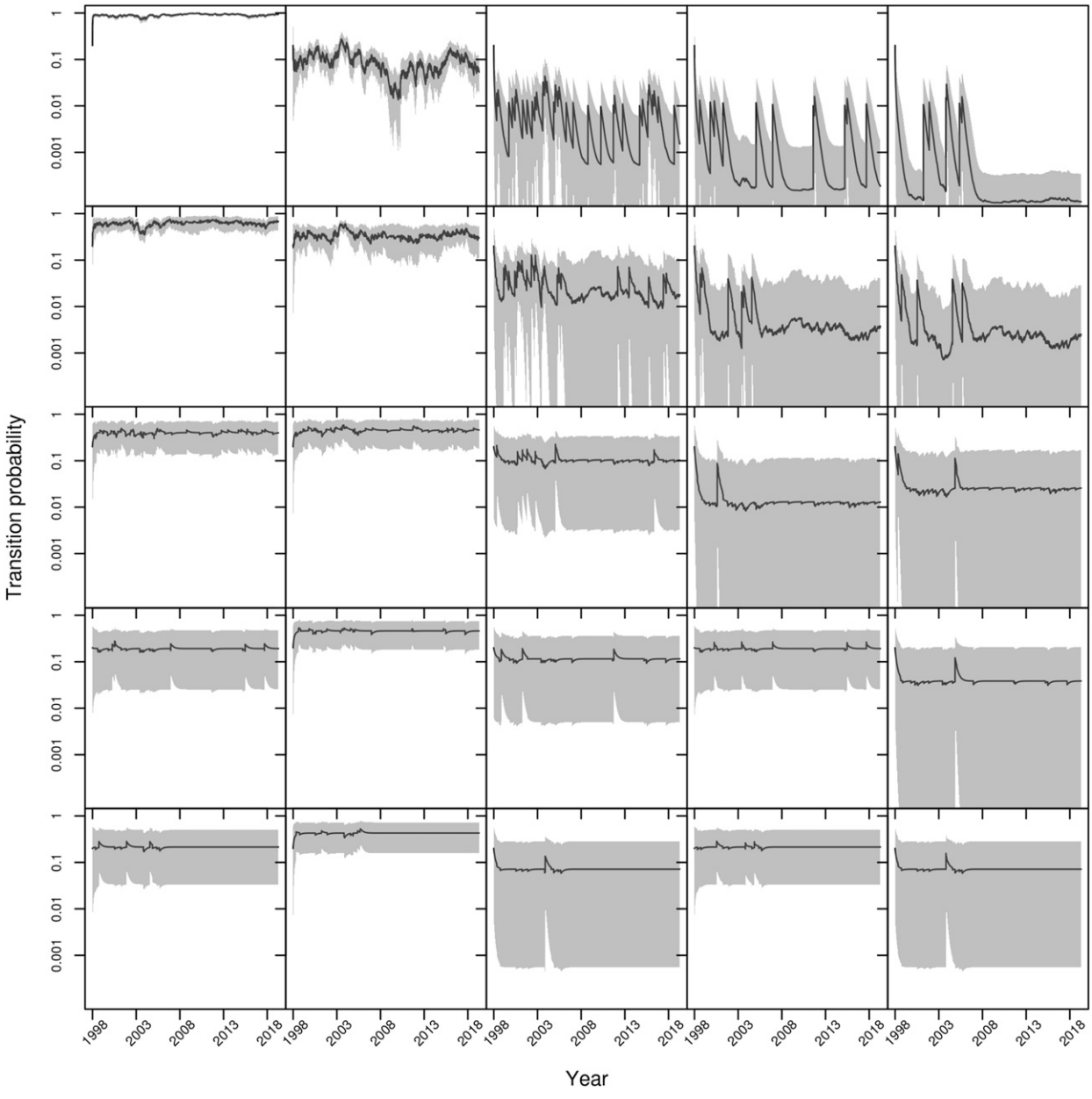

FIG. 5. Time evolution of all NHMC transition probabilities for 1998–2019, for $\tau = 100$ days and $\kappa = 10$. The solid black lines represent the NHMC estimates of the transition probabilities, and the gray shading indicates the central 95% confidence interval constructed from the 2.5th and 97.5th percentiles of the beta distribution for each transition probability. The panel matrix shows the probabilities of transitions from rows to columns. The rows, from top to bottom, and columns, from left to right, both correspond to <G 1, G 1/2, G 3, G 4, and G 5. For example, the panel in the first row and second column shows the probability of a transition from state <G 1 to G 1/2 and the panel in the third row and fourth column shows the probability of a transition from state G 3 to G 4.

pairs of forecast models, under the null hypothesis that the model forecasts are equally accurate on average. The p values given in Table 4 show that the null hypothesis can be rejected at the 1% level for each comparison and lead time. This supports the conclusion that the NHMC is, on average, more skillful than the HMC, which is also more skillful than the climatology.

Figure 5 shows the time evolution of the transition probabilities estimated using the NHMC model over the whole period. The gray shading shows the 95% confidence interval calculated from the 2.5th and 97.5th percentiles of the beta distribution for each transition probability (i.e., the appropriate marginals of the Dirichlet distribution). The panel matrix shows the

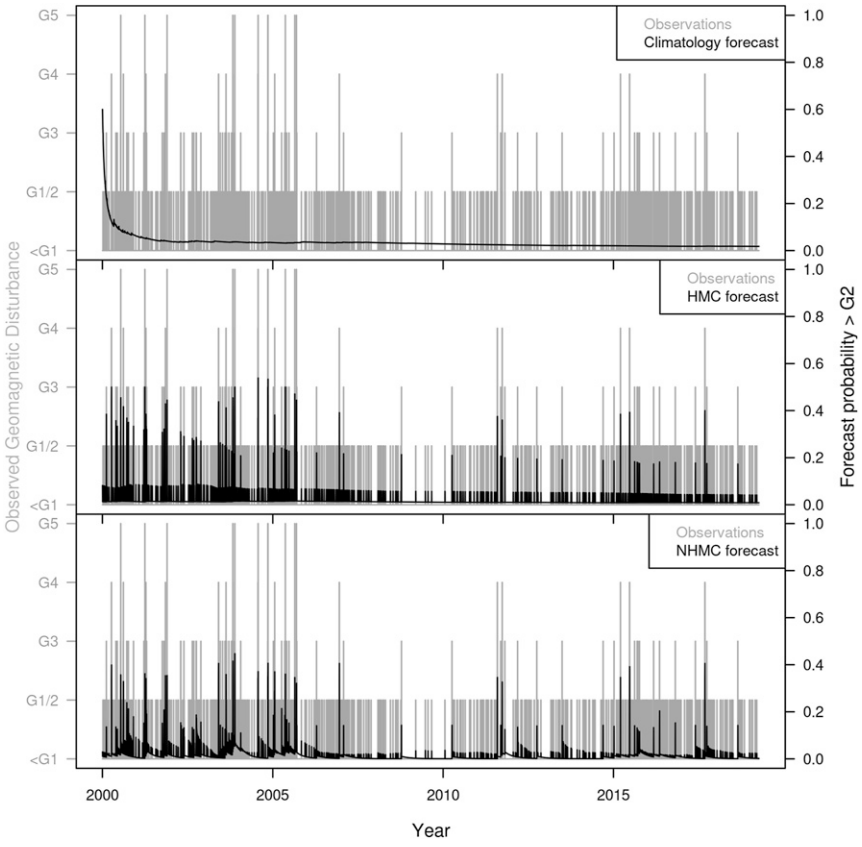

FIG. 6. Comparison of daily forecasts, shown in black, based on (top) climatology, (middle) HMC, and (bottom) NHMC for $\tau = 100$ days and $\kappa = 10$, with observed geomagnetic states shown in gray. The G 3–G 5 states have been aggregated to demonstrate the comparison more clearly.

probabilities of transitions from rows to columns, with the rows, from top to bottom, and the columns, from left to right, both labeled as <G 1, G 1/2, G 3, G 4, and G 5. For example, the panel in the first row and second column shows the probability of a transition from state <G 1 to G 1/2; the third row and fourth column shows the probability of a transition from state G 3 to G 4.

The discontinuities in the time series are the result of updates to the probability following an event that resulted in a geomagnetic disturbance. The rarity of discontinuities for the higher states illustrates the limited data on which to base the corresponding transition probabilities, which is also evident in the wider 95% confidence intervals. The exponential decay for these rare transitions following the discontinuity illustrates the relaxation back to the reference state estimated using the HMC. Furthermore, the temporal variations found in the different transition probabilities do not follow a single form nor do they closely match a known forcing such as the solar sunspot cycle.

Figure 6 shows examples of the daily forecasts based on climatology, HMC, and NHMC, with observed geomagnetic states given by NOAA-SWPC data. For clarity, we have aggregated forecast probabilities for the G 3–G 5 states to show the daily probability of high-impact geomagnetic disturbances. In the top panel, the climatological forecast probability decays rapidly from the uniform prior initial conditions to a more realistic long-term climatological mean probability. The initial conditions assume that each of the five categories are equally likely, giving an unrealistically high daily probability of 60% for G 3–G 5 events. However, within 1–2 years, enough events have occurred for the climatology forecast to be comparable to the sum of the G 3–G 5 probabilities in Table 2 (i.e., 1.5%).

The exponential decay described above is also evident in the middle and bottom panels; however, both the HMC and NHMC capture key natural variations of the geomagnetic disturbance which cannot be reproduced by the climatology forecast, such as the relatively active phase during 2000–07. This is particularly clear in the

short-term persistence which occurs during the days after a disturbance.

There are also important differences between the behavior of the HMC and the best-performing NHMC. In particular, the HMC forecast probabilities evolve rapidly from their initial uniform configuration toward the homogeneous values given in Table 1. The mathematical construction of the HMC means that each subsequent update to the Dirichlet parameters results in fractionally smaller changes to the transition probabilities. As such, by the end of the validation period, updates to the Dirichlet parameters become negligible and the transition matrix becomes practically constant. In contrast, since the NHMC is formulated to down-weight older information, it continues to closely follow natural variations in the geomagnetic disturbances throughout the validation period. The NHMC is also seen to exponentially relax toward the predefined reference state during less active periods. Therefore, unlike the HMC, the NHMC forecast accuracy will not degrade over time.

5. Conclusions

This study has proposed and demonstrated a simple yet flexible nonhomogeneous Markov Chain approach for creating probabilistic forecasts of categorical events, which provides an alternative data-driven benchmark for physics-based models. The approach captures persistence in the series of events by assuming that the probabilities of the next state depend on the current state. Furthermore, transition probabilities are allowed to vary slowly over time in order to account for possible modulation by unknown drivers.

The approach has been demonstrated here by application to geomagnetic storm forecasts. Rank probability scores show that the nonhomogeneous forecasts outperform those made with a homogeneous Markov chain for all lead times from 1 to 4 days. There is substantially improved skill when using a memory time scale of around 100 days compared to either shorter or longer memory time scales. There is also clear evidence of temporal evolution in the transition probabilities that is not directly related to the solar cycle. This probabilistic categorical approach is novel compared to more stationary approaches, such as ARMA models (e.g., Thomson et al. 2001) and neural networks (e.g., Thomson et al. 1995; Thomson 1996; Boberg et al. 2000). For operational forecasting, it would be of interest to combine this statistical approach with more physical information about the speeds and arrival times CMEs, which are known to trigger the onset of transitions to higher G-states.

The approach is also relevant to other applications in meteorology and other fields, for example, in forecasting discrete weather regime states, or wet and dry days of relevance for hydrology. It would also be of interest to extend the model to allow the memory time scale to be longer for states that occur less frequently where there is less recent transition information. At a more fundamental level, it would be of interest to better understand how the autoregressive updating of the Dirichlet parameters can be interpreted consistently using Bayes's theorem.

Acknowledgments. We are grateful to the Met Office Academic Partnership, which allowed Edward Pope to spend time on this work while on secondment at the University of Exeter.

REFERENCES

Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.*, **139**, 1891–1910, https://doi.org/10.1175/2010MWR3615.1.

Baker, D. N., R. L. McPherron, T. E. Cayton, and R. W. Klebesadel, 1990: Linear prediction filter analysis of relativistic electron properties at 6.6 RE. *J. Geophys. Res.*, **95**, 15 133–15 140, https://doi.org/10.1029/JA095iA09p15133.

Bartels, J., N. H. Heck, and H. F. Johnston, 1939: The three-hour-range index measuring geomagnetic activity. *J. Geophys. Res.*, **44**, 411–454, https://doi.org/10.1029/TE044i004p00411.

Bertuccelli, L. F., and J. P. How, 2008: Estimation of non-stationary Markov chain transition models. *Proc. 47th IEEE Conf. on Decision and Control*, Cancun, Mexico, IEEE, https://doi.org/10.1109/CDC.2008.4738904.

Bloomfield, D. S., P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher, 2012: Toward reliable benchmarking of solar flare forecasting methods. *Astrophys. J.*, **747**, L41, https://doi.org/10.1088/2041-8205/747/2/L41.

Boberg, F., P. Wintoft, and H. Lundstedt, 2000: Real time Kp predictions from solar wind data using neural networks. *Phys. Chem. Earth, Part C Sol.-Terr. Planet. Sci.*, **25**, 275–280, https://doi.org/10.1016/S1464-1917(00)00016-7.

Boynton, R. J., M. A. Balikhin, D. G. Sibeck, S. N. Walker, S. A. Billings, and N. A. Ganushkina, 2016: Electron flux models for different energies at geostationary orbit. *Space Wea.*, **14**, 846–860, https://doi.org/10.1002/2016SW001506.

Caskey, J. E., 1963: A Markov chain model for the probability of precipitation occurrence in intervals of various length. *Mon. Wea. Rev.*, **101**, 198–301, https://doi.org/10.1175/1520-0493(1963)091<0298:AMCMFT>2.3.CO;2.

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Gabriel, K. R., and J. Neumann, 1962: A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteor. Soc.*, **88**, 90–95, https://doi.org/10.1002/qj.49708837511.

Gates, P., and H. Tong, 1976: On Markov chain modeling to some weather data. *J. Appl. Meteor.*, **15**, 1145–1151, https://doi.org/10.1175/1520-0450(1976)015<1145:OMCMTS>2.0.CO;2.

Gilleland, E., and G. Roux, 2015: A new approach to testing forecast predictive accuracy. *Meteor. Appl.*, **22**, 534–543, https://doi.org/10.1002/met.1485.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 274 pp.

Kotz, S., N. Balakrishnan, and N. L. Johnson, 2000: *Continuous Multivariate Distributions: Models and Applications*. Vol. 1, 2nd ed. John Wiley and Sons, https://doi.org/10.1002/0471722065.

MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, https://doi.org/10.1002/qj.2396.

Owens, M. J., M. Lockwood, and P. Riley, 2017: Global solar wind variations over the last four centuries. *Sci. Rep.*, **7**, 41 548, https://doi.org/10.1038/srep41548.

Paulo, A. A., and L. S. Pereira, 2007: Prediction of SPI drought class transitions using Markov chains. *Water Resour. Manage.*, **21**, 1813–1827, https://doi.org/10.1007/s11269-006-9129-9.

Rajagopalan, B., U. Lall, and D. G. Tarboton, 1996: Nonhomogeneous Markov model for daily precipitation. *J. Hydrol. Eng.*, **1**, 33–40, https://doi.org/10.1061/(ASCE)1084-0699(1996)1:1(33).

Sharpe, M. A., and S. A. Murray, 2017: Verification of space weather forecasts issued by the Met Office Space Weather Operations Centre. *Space Wea.*, **15**, 1383–1395, https://doi.org/10.1002/2017SW001683.

Sung, M., 2014: Bayesian conjugate analysis for transition probabilities of non-homogeneous Markov Chain: A survey. *Commun. Stat. Appl. Methods*, **21**, 135–145, https://doi.org/10.5351/CSAM.2014.21.2.135.

Swinbank, R., P. Friederichs, and S. Bentzien, 2016: Forecasting high-impact weather using ensemble prediction systems. *Dynamics and Predictability of Large-Scale High-Impact Weather and Climate Events*, J. Li et al., Eds., Cambridge University Press, 5–112.

Thomson, A. W. P., 1996: Non-linear predictions of Ap by activity class and numerical value. *Pure Appl. Geophys.*, **146**, 163–193, https://doi.org/10.1007/BF00876675.

——, E. Clarke, and T. D. G. Clark, 1995: Study on forecasting of solar and geomagnetic activity, Phase 2. British Geological Survey Tech. Rep. WM/95/16C, 189 pp.

——, J. A. King, T. D. G. Clark, and E. Clarke, 2001: Improved prediction of solar and geomagnetic influence on the near Earth environment. British Geological Survey Tech. Rep. CR/01/162, 72 pp.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed.International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Woolhiser, D. A., and G. G. S. Pegram, 1979: Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *J. Appl. Meteor.*, **18**, 34–42, https://doi.org/10.1175/1520-0450(1979)018<0034:MLEOFC>2.0.CO;2.