J. Space Weather Space Clim. 2018, 8, A34

https://doi.org/10.1051/swsc/2018022

JSWSC

Available online at:
www.swsc-journal.org

Developing New Space Weather Tools: Transitioning fundamental science to operational prediction systems

RESEARCH ARTICLE OPEN ACCESS

Flare forecasting using the evolution of McIntosh sunspot classifications

Aoife E. McCloskey[1,*], Peter T. Gallagher[1] and D. Shaun Bloomfield[2]

[1] School of Physics, Trinity College Dublin, College Green, Dublin 2, Ireland
[2] Northumbria University, Newcastle upon Tyne, NE1 8ST, UK

Received 4 December 2017 / Accepted 2 May 2018

Abstract – Most solar flares originate in sunspot groups, where magnetic field changes lead to energy build-up and release. However, few flare-forecasting methods use information of sunspot-group evolution, instead focusing on static point-in-time observations. Here, a new forecast method is presented based upon the 24-h evolution in McIntosh classification of sunspot groups. Evolution-dependent ≥C1.0 and ≥M1.0 flaring rates are found from NOAA-numbered sunspot groups over December 1988–June 1996 (Solar Cycle 22; SC22) before converting to probabilities assuming Poisson statistics. These flaring probabilities are used to generate operational forecasts for sunspot groups over July 1996–December 2008 (SC23), with performance studied by verification metrics. Major findings are: (i) considering Brier skill score (BSS) for ≥C1.0 flares, the evolution-dependent McIntosh-Poisson method ($BSS_{evolution} = 0.09$) performs better than the static McIntosh-Poisson method ($BSS_{static} = -0.09$); (ii) low BSS values arise partly from both methods over-forecasting SC23 flares from the SC22 rates, symptomatic of ≥C1.0 rates in SC23 being on average ≈80% of those in SC22 (with ≥M1.0 being ≈50%); (iii) applying a bias-correction factor to reduce the SC22 rates used in forecasting SC23 flares yields modest improvement in skill relative to climatology for both methods ($BSS^{corr}_{static} = 0.09$ and $BSS^{corr}_{evolution} = 0.20$) and improved forecast reliability diagrams.

Keywords: operational forecasting / solar flares / sunspot groups

1 Introduction

Solar flares are one of the most energetic space weather phenomena that affects the near-Earth environment. They most commonly originate within sunspot groups, where evolution of complex magnetic field leads to magnetic reconnection and subsequent large magnitudes of energy release. In the reconnection process, stored magnetic energy is rapidly converted to both thermal and kinetic energy in addition to non-thermal acceleration of particles (Priest & Forbes, 2002). Solar flares, or coronal mass ejections (CMEs) if material is ejected, are understood to be caused by this magnetic reconnection process. Due to the high-energy radiation release and particle acceleration, these phenomena can have damaging effects on both Earth and space-based technologies (e.g., satellites and radio communication). Unlike CMEs that typically take 1–3 days to propagate to Earth after launch is detected, flare-related space weather impacts begin within minutes of flare onset (e.g., ionospheric disturbances; Mitra, 1974). Therefore, it is of high priority that methods are developed to forecast when flares may occur, and the magnitude of energy release, in order to mitigate their effects.

Over the past several decades, there have been many published works focused on the classification of sunspot groups in terms of their complexity and their relation to flare production. The most well-known are the Mount Wilson (Hale et al., 1919) and McIntosh (McIntosh, 1990) schemes, classifying sunspots according to their magnetic and white-light structure, respectively. The relationship between these sunspot group classifications and flaring has been investigated in several studies and it was shown that the more "complex" sunspot-group classifications are associated with higher frequency and magnitude of flaring (Waldmeier, 1947; Bornmann and Shaw, 1994).

In terms of solar flare prediction, one of the most established methods that has been developed to forecast solar flares is based upon sunspot-group classification, namely the McIntosh classification scheme. Gallagher et al. (2002) developed a Poisson-based method for calculating flare probabilities from the historical flaring rates of McIntosh classifications (publicly available at www.solarmonitor.org). Later this method was expanded upon and the performance of interpreting probabilities as dichotomous yes/no forecasts was verified by Bloomfield et al. (2012), where it was shown that Poisson probabilities performed comparably to some of the

*Corresponding author: mccloska@tcd.ie

more complex flare prediction methods in use at that time. There currently exists a vast quantity of prediction/forecasting methods including the most recent development of applying machine learning techniques to flare forecasting (see, e.g., Colak & Qahwaji, 2009; Ahmed et al., 2013; Bobra & Couvidat, 2015). For more information on the multitude of prediction/forecasting methods, see the recent comparison paper by Barnes et al. (2016) and references therein.

Several space weather Regional Warning Centres (RWCs) make use of the Poisson-based flare forecasting approach. The US National Oceanographic and Atmospheric Administration (NOAA) Space Weather Prediction Centre (SWPC) RWC uses the McIntosh scheme as an input for their "expert" decision-rule system that is used to assign flaring probabilities to active regions (McIntosh, 1990) that are augmented by experienced space weather forecasters prior to being issued. The UK Met Office Space Weather Operations Centre (MOSWOC) RWC also uses the historical flaring rates of McIntosh classes to calculate an initial forecast, again later adjusted by human forecasters. The performance of these operational forecasts have been evaluated and shown to perform well compared to more complex methods, with improvement in performance achieved by including the human editing of probabilities (Crown, 2012; Murray et al., 2017), also true for the Belgian Solar Influences Data Center (SIDC) RWC (Devos et al., 2014).

Until now, few forecasting methods account for evolution in sunspot-group properties, but there have been some research-focused studies considering evolution in sunspot-group classifications. Lee et al. (2012) investigated a subset of McIntosh classes alongside their 24-h change in sunspot area, finding that groups which increased in area had a higher flaring rate compared to groups with steady or decreasing area. Comparatively, McCloskey et al. (2016) calculated evolution-dependent flaring rates for the three components of the McIntosh classification scheme. It was shown that when sunspot groups evolve upward in their McIntosh class higher 24-h flaring rates are observed, with lower flaring rates being true for downward evolution. So far, however, no verified forecasting methods have included the temporal evolution of sunspot-group classifications.

In this paper, we investigate the evolution of McIntosh sunspot-group classifications over 24-h time scales as a method for forecasting solar flare magnitude and occurrence. The data we use is based upon McCloskey et al. (2016), where historical flaring rates were calculated for McIntosh evolutions from the training period 1988 to 1996 (Solar Cycle 22; SC22), with more recent data from 1996 to 2008 (SC23) included for testing and forecast verification. In Section 2 we provide more details on the data used in this study and the method used to produce flare probabilities. Section 3 discusses the results of the forecasting method along with verification metrics and an exploration of the maximum performance possible when applying linear Cycle-to-Cycle rate corrections. Finally, in Section 4 we present our conclusions and outlook for future work.

2 Data analysis

2.1 Data sources

The data used in this study are that analysed by McCloskey et al. (2016), consisting of historical sunspot-group classifications and flare information collected by NOAA/SWPC. SWPC provide a daily Solar Region Summary (SRS) issued at 00:30 UT, with sunspot-group properties including NOAA active region number, heliographic coordinates, McIntosh and Mount Wilson classifications, and longitudinal extent. Additionally, solar flares associated with these regions were obtained from the Geostationary Operational Environment Satellite (GOES) event lists collated by SWPC. It is noted that the association of a flare to a specific NOAA active region is carried out by SWPC for up to three days after the event occurs. We chose to include all GOES 1–8 Å soft X-ray flares of C-class and above (i.e., $\geq 10^{-6}\,\mathrm{Wm}^{-2}$), with the reason for excluding flares below these magnitudes being the high background solar X-ray flux level at solar maximum that obscures B-class and lower flares.

The data used here as a training set for our forecasting method was taken from the SC22 period of 1 December 1988 to 31 July 1996, inclusive (Balch, 2011, private communication). It is noted that although SC22 is estimated to have commenced in September 1986 (Hathaway et al., 1999), the region-associated flare data from before December 1988 was not available and therefore could not be included here. This provided a data set of 24-h flaring rates calculated for individual evolutions in McIntosh classification parameters, i.e., modified Zurich, penumbral or compactness classes. However, it is important to note that in this study we chose to make use of the *evolution in the full McIntosh classification* of each sunspot group rather than the evolution in the three separate components studied in McCloskey et al. (2016). Section 2.2 outlines this distinction in further detail.

The data used here as a test set was obtained from the publicly available NOAA/SWPC website (ftp://ftp.swpc.noaa.gov/pub/warehouse/) over the SC23 period of 31 July 1996 to 13 December 2008, inclusive, in order to ensure an independent data set for forecast verification. Using the same method as McCloskey et al. (2016), McIntosh classifications were extracted for each unique NOAA sunspot group along with the region-associated GOES X-ray flares. A total of 21 476 individual daily sunspot-group entries were extracted in the test period, corresponding to 3017 unique NOAA active regions. The total number of GOES soft X-ray flares associated with these regions was 8647, consisting of 7434 C-class, 1106 M-class, and 107 X-class flares.

2.2 Full McIntosh classification evolution

The McIntosh classification scheme is a long-established method for classifying the white-light structure of sunspot groups. It was first developed by Cortie (1901) and later expanded upon and modified to include additional parameters (Waldmeier, 1947; McIntosh, 1990). The scheme is comprised of 17 different parameters which combine to form 60 different allowed classifications. These parameters are divided into three component classes: modified Zurich, penumbral and compactness (*Zpc*). In summary, "*Z*" describes the longitudinal extent of the sunspot group, "*p*" describes the size and symmetry of the penumbra of the leading spot and "*c*" describes the distribution of sunspots in the interior of the group. For a more detailed description of these components and their allowed combinations, see McIntosh (1990), Bornmann & Shaw (1994), and McCloskey et al. (2016).

Previously, it has been shown that the McIntosh classifications of sunspot groups and their flare productivity are related. Importantly, there is evidence that the McIntosh classification can capture differences in flaring rates for sunspot groups, with more complex classifications producing higher flaring rates overall (Bornmann & Shaw, 1994). Building upon this, McCloskey et al. (2016) showed that the 24-h evolution of McIntosh sunspot-group classifications show comparable results in terms of the rate of flare production – sunspot groups that evolved upward in a classification component produced higher flaring rates, while downward evolution produced lower flaring rates. In this paper we make use of this statistical relationship to implement a method for flare forecasting using the 24-h evolution of McIntosh classifications.

As previously mentioned, instead of considering the evolution in only a single McIntosh component (i.e., $Z_1 \rightarrow Z_2$ or $p_1 \rightarrow p_2$ or $c_1 \rightarrow c_2$), the full McIntosh class evolution of a sunspot group is extracted over 24 h (i.e., $\{Zpc\}_1 \rightarrow \{Zpc\}_2$). The main reasoning for this was to better capture the information in the evolution of the complete white-light structure of each sunspot group that was naturally excluded by considering only evolution in a single McIntosh component. Here, the average flaring rate associated with one unique $\{Zpc\}_1 \rightarrow \{Zpc\}_2$ evolution is determined by extracting all instances of active regions that underwent that McIntosh class evolution. From this subset of active regions, the total number of flares that were produced within 24 h of that specific evolution are divided by the total number of regions in that subset.

To verify that the previously observed relationship between McIntosh-class evolution and flaring rate is also present when considering the full McIntosh classification, Figure 1 depicts flaring rates for a selection of full McIntosh-class evolutions. This selection was chosen to represent evolution by evolving sequentially in at least one parameter (e.g., a DSO evolving to a BXO, followed by a DSO evolving to a CSO). Note that this graphical representation is less continuous to that shown in McCloskey et al. (2016), since bars that lie two steps apart may depict evolution in two separate McIntosh components (rather than two steps in one component in the previous work). Figure 1a plots the occurrence-frequency distribution, with the most frequent occurrence once again being no evolution in McIntosh class over 24 h (black bar). When evolution does occur, a DSO-type is most likely to evolve upward in penumbral class (i.e., to DAO) or downward in modified Zurich class (i.e., to CSO). This reflects the previous findings of McCloskey et al. (2016) where sunspot groups are most likely to remain the same classification and are not likely to evolve significantly over a 24-h period (i.e., rarely more than two evolution steps in any one McIntosh component).

Figure 1b displays the ≥C1.0 flaring rates associated with these selected McIntosh evolutions. This plot indicates that there are increasingly higher flaring rates associated with greater evolution steps upward in at least one McIntosh parameter, with the opposite true for greater evolution steps downward (i.e., sequentially decreasing rates). Additionally, for flaring rates λ, associated Poisson errors are calculated as $\Delta\lambda = 1/\sqrt{M}$, where M is the total number of sunspot groups that underwent that evolution in McIntosh class. These are shown as error bars in both Figures 1b and c, where the maximum error in flaring rate is ±1. Similar behaviour is seen for ≥M1.0 flaring rates in Figure 1c, with higher flaring rates seen for evolution upward in McIntosh class, however due to low occurrence numbers these rates are deemed not statistically significant (i.e., $\lambda \pm \Delta\lambda$ encompasses zero). This relationship of McIntosh class evolution and flaring rates is comparable to the findings of McCloskey et al. (2016).

2.3 Issuing Poisson probabilities

For the purpose of testing the forecast method in an operational manner, forecasts for ≥C1.0 and ≥M1.0 flares are issued for each 24-h time window from 00:00 UT in the form of probabilities of flare occurrence. It has been previously shown that the waiting-time distributions of soft X-ray flares from individual active regions is well represented by a time-dependent Poisson process with typical piece-wise constant flaring-rate timescales of >2–3 days (Wheatland, 2001). As that work encompasses the full lifetime of individual active regions, and hence their evolution across McIntosh classes, we find the assumption of Poisson statistics suitable for our work. Here, we convert our evolution-dependent 24-h flaring rates to probabilities as follows,

$$P_\lambda(N_f) = \frac{\lambda^{N_f}}{N_f!} e^{-\lambda}, \tag{1}$$

where N_f is the number of flares expected to occur in a 24-h period following an evolution and λ is the average number of flares observed within the 24 h immediately following each unique evolution in McIntosh class. Note, these flaring probabilities are calculated separately for each unique full McIntosh evolution using the training set data of SC22. Hence, the probability of one or more flares occurring in a given time interval following an evolution is then calculated by,

$$P_\lambda(N_f \geq 1) = 1 - P_\lambda(N_f = 0) = 1 - e^{-\lambda}. \tag{2}$$

By using a 24-h flaring rate, the issued flaring probabilities are then valid for the following 24-h period (i.e., 00:00 UT–00:00 UT). Although the SWPC SRS files used to determine McIntosh-class evolution are issued at 00:30 UT, here the forecast interval begins at 00:00 UT as this is the end-time at which McIntosh classifications are constructed from the previous 24 h. It is worth noting that there are certain circumstances where our evolution-dependent forecasting method will not be able to issue a forecast. This includes the first day a sunspot group appears on disk and therefore no evolution can have been observed, while there are a small number of full McIntosh-class evolutions that were not observed in the training data set and therefore no evolution-dependent flaring rate can be assigned in the test data set. Rather than disregard these sunspot groups from the analysis, we have chosen instead to use the standard static point-in-time flaring rates and hence probabilities for these cases based on the currently observed full McIntosh class. This satisfies the purpose of creating an operational forecasting method and allows for a more fair comparison of our evolution-dependent method with the original static method.

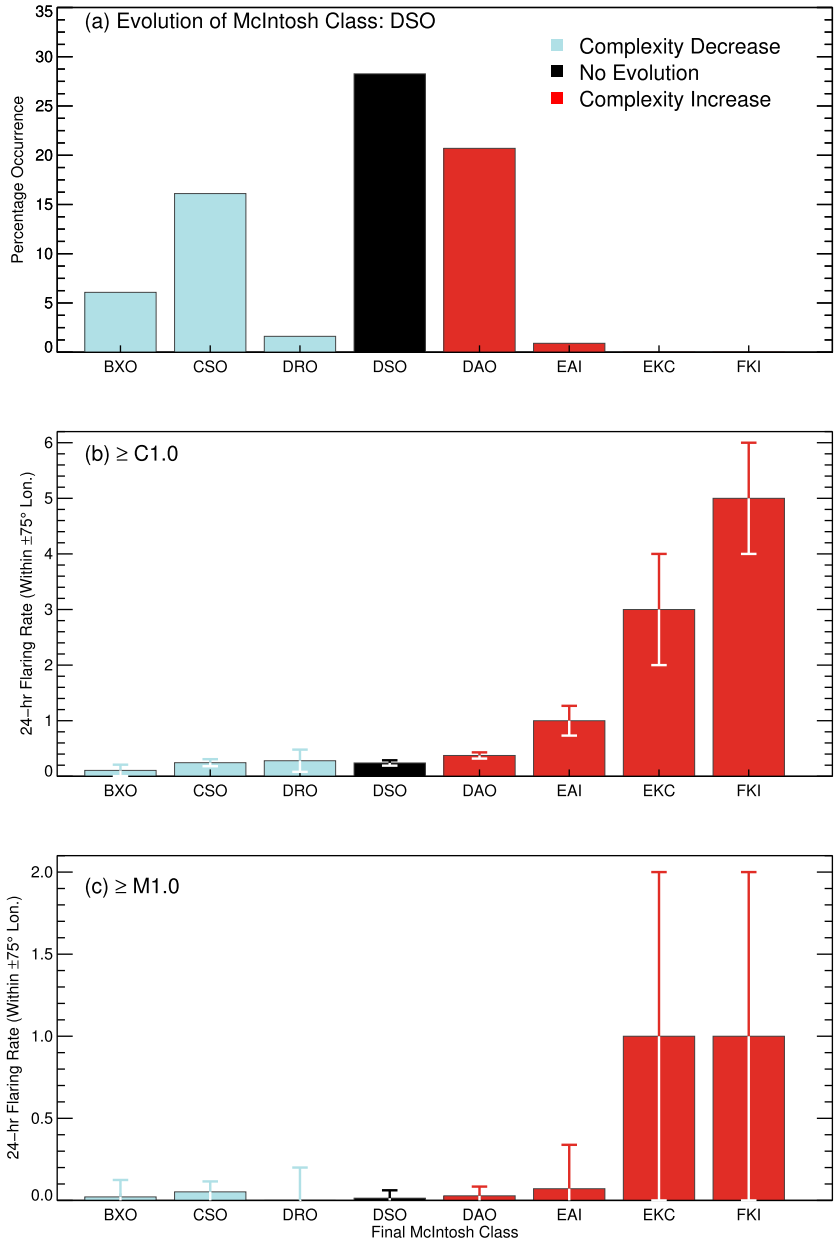

Fig. 1. Histograms showing the 24-h evolution of sunspot groups starting as a DSO-type McIntosh classification (a), with bars representing the percentage of evolutions observed starting as DSO and evolving to a sub-group of McIntosh classifications. The corresponding evolution-dependent ≥C1.0 and ≥M1.0 flaring rates are shown in panels (b) and (c), respectively. Histogram bars are coloured by evolution: no evolution (black); upward evolution (dark red); downward evolution (light blue).

3 Results

3.1 Forecast verification

Various verification metrics can be investigated to quantify the performance of a forecasting method. There are two main types of forecasting methods that are widely used, namely categorical and probabilistic. Dichotomous categorical forecasts have only two possible values when predicting if an event will occur (i.e., yes/no), whereas probabilistic forecasts yield a range of values (i.e., decimal percentage between 0 and 1). Here, we evaluate the performance of both the original static McIntosh method (Gallagher et al., 2002) and our new evolution-dependent McIntosh method focusing on verification techniques suited for probabilistic forecasts. This allows for direct comparison of the two methods using probabilistic verification metrics that were not explored in the previous benchmarking study of Bloomfield et al. (2012).

One of the main quantities that assesses the performance of a probabilistic forecast is the Brier score (BS). In its simplest form, BS is equivalent to the mean-squared error between the issued forecast probability, f (i.e., 0–1), and the observed binary outcome for that forecast, o (i.e., 0 or 1),

Table 1. Decomposed Brier score (BS) components and Brier skill score (BSS) for the McIntosh static and evolution-dependent forecast methods.

Flaring magnitude	Forecast method	BS components			BSS
		Reliability	Resolution	Uncertainty	
≥C1.0	Static	0.037	0.025	0.146	−0.09
≥C1.0	Evolution	0.033	0.046	0.146	0.09
≥M1.0	Static	0.017	0.003	0.038	−0.36
≥M1.0	Evolution	0.014	0.009	0.038	−0.15

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{3}$$

where N is the total number of forecasts issued and i identifies specific forecast-observation pairs. If the issued forecasts can be identified as groups of unique forecast probabilities, the BS can be decomposed into three components (Murphy, 1973),

$$\begin{aligned}\mathrm{BS} &= \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{o}_k - \overline{o})^2 \\ &\quad + \overline{o}(1-\overline{o}), \\ &= \text{reliability} - \text{resolution} + \text{uncertainty},\end{aligned} \tag{4}$$

where k identifies unique forecast-probability groups, n_k is the number of occurrences in each k group, $\overline{o}_k$ is the corresponding observed frequency of events in that k group (i.e., the climatology for that unique forecast group) and $\overline{o}$ is the overall climatology of events for all valid forecast days. Climatology of events refers to the long-term average value of binary flare occurrence (i.e., 0 or 1) over the period of testing (i.e., SC23). Reliability is a measure of how close the issued probability of a unique forecast group is to the frequency of observed outcomes for that unique forecast group (i.e., the average binary outcome of their observed events), where a reliability value of 0 corresponds to a perfectly reliable forecast. The resolution term measures the difference between the climatology of the unique forecast groups and the overall climatology, which can be interpreted as the potential ability of the unique forecast groups to perform better than unskilled climatology (i.e., the higher the value of resolution the better). Finally, the uncertainty term measures the variability in the observed event frequency, which is independent of unique forecast grouping and is largest when an event is difficult to predict (i.e., occurring 50% of the time) and smallest when an event occurs almost always or never. In the context of this work, the issued forecast probabilities can be considered as binned into k unique bins where each represents a unique McIntosh-class evolution (e.g., AXX to BXO).

To interpret the performance of a forecast set, it is standard practice to normalise a verification metric score, S, to that of a reference forecast, S_{ref}, by means of a skill score (SS),

$$\mathrm{SS} = \frac{S - S_{\mathrm{ref}}}{S_{\mathrm{perfect}} - S_{\mathrm{ref}}}, \tag{5}$$

where S_{perfect} is the score of a perfect forecast for the chosen verification metric. In the case of BS, a perfect forecast has a value of 0 and the reference forecast is typically taken to be that achieved by climatology, $\mathrm{BS}_{\mathrm{clim}}$ (equivalent to the uncertainty term in equation (4), as reliability and resolution cancel each other out). The Brier skill score (BSS) is then given as,

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{clim}}}{0 - \mathrm{BS}_{\mathrm{clim}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}}. \tag{6}$$

This can also be represented via the decomposed form of equation (4) by the three components as,

$$\begin{aligned}\mathrm{BSS} &= 1 - \frac{\text{reliability} - \text{resolution} + \text{uncertainty}}{\text{uncertainty}} \\ &= \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}}.\end{aligned} \tag{7}$$

Table 1 presents the three decomposed BS components and BSS for ≥C1.0 and ≥M1.0 flares for both the McIntosh static and evolution-dependent forecasting methods. Focusing on BSS values for ≥C1.0 flares, both methods achieve similar reliability values of 0.037 and 0.033, respectively. Considering now the resolution, as these values contribute to the overall BSS positively, if the value of resolution is greater than reliability the overall BSS will be positive. For the static method, despite being reasonably reliable it does not achieve a positive BSS (−0.09) as the value of resolution is too low (0.025) – the climatology for many of the unique forecast groups are indistinguishable from the overall climatology (i.e., little forecast discrimination ability). Although the evolution-dependent method has a similar reliability value, its resolution (0.046) is higher, relative to both the static method and its own reliability term, contributing to a positive BSS (0.09). Achieving a positive value for BSS indicates that the evolution-dependent method is performing better than the climatology reference forecast, while the static method does not.

In addition to skill scores, it is useful to visualise the performance of the forecast method. The two most popular visual diagnostics are reliability diagrams and relative operating characteristic (ROC) curves like those provided in Figure 2a and c, respectively. Reliability diagrams indicate differences between forecast probabilities and the observed frequencies of events (similar to the reliability term of the BSS in Eq. (4)), with forecast probabilities plotted along the horizontal axis, binned into sub-groups of forecasts, and the frequency of observed events for each sub-group plotted on the vertical axis. Here we chose to use 10% probability intervals, p, with the associated Bayesian uncertainty for each bin shown

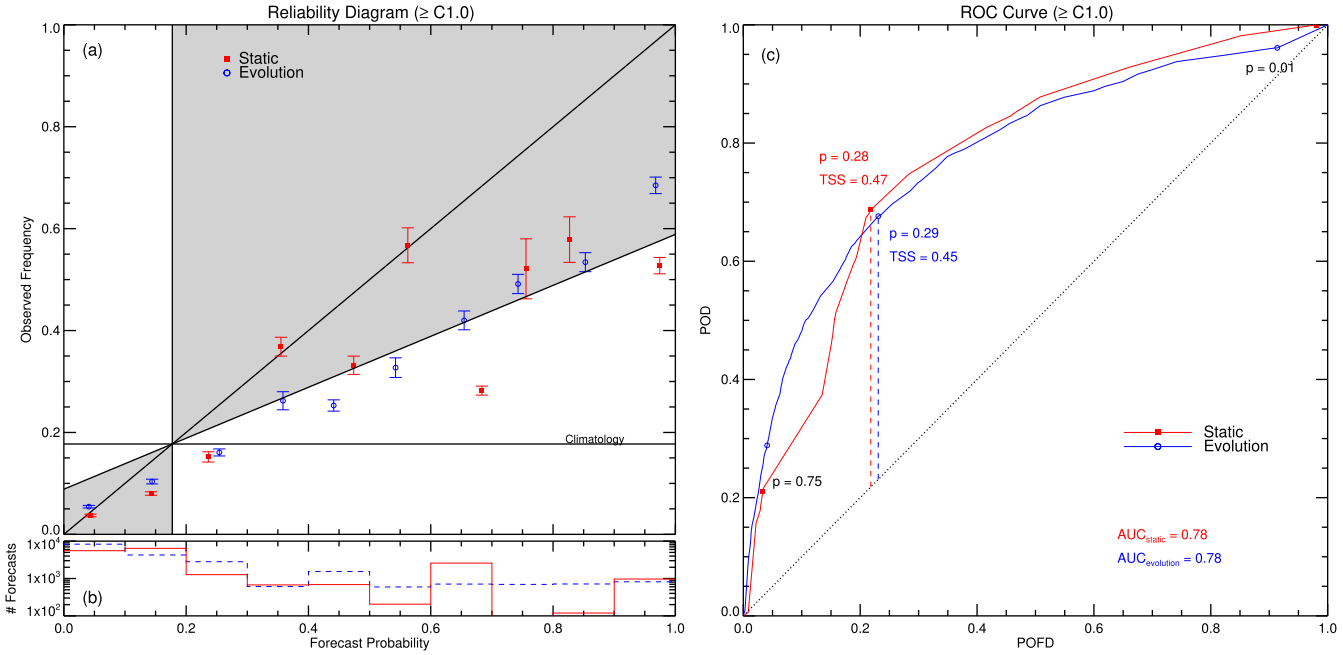

Fig. 2. Reliability diagrams (panel a), sharpness (i.e., probability occurrence) plots (panel b), and ROC curves (panel c) for ≥C1.0 flares. Data for the McIntosh static forecast method are indicted by red filled squares (panels a and c) and solid histogram (panel b), while the evolution-dependent method is depicted by blue open circles (panels a and c) and dashed histogram (panel b).

as error bars, $\sigma_p = \sqrt{p(1-p)/(T+3)}$, where T is the total number of forecast days in each probability bin (Wheatland, 2005), indicated in the sharpness plot of Figure 2b. The overall climatology of events is plot as a horizontal and a vertical line, with the shaded area indicating the region that data contribute positively to BSS.

Forecasts for the McIntosh static (red filled squares) and evolution-dependent (blue open circles) methods can be directly compared here, as both are applied to the same testing time period and so have the same climatology. For the static case, the majority of points lie within the shaded area, which can contribute positively to the BSS. However, while three points lie on the line of perfect reliability (i.e., $y = x$) most are found below this line, indicating the method is over-forecasting (i.e., the values of forecast probabilities are too high relative to the observed frequency of events for that forecast bin). It is interesting to note that the evolution-dependent case also appears to be over-forecasting, but in a more consistent manner (i.e., linearly biased from perfect reliability) than the static case. Notably, the static method achieves a worse (and negative) BSS compared to the evolution-dependent method, which is reflected in the reliability diagrams by more significant deviation of data points from the $y = x$ line and their relatively larger occurrence frequencies (e.g., for the static case, $p = 0.6$–0.7 is the greatest outlier while being the third-most populated bin).

For alternative verification purposes it is also possible to convert the probabilistic forecasts into dichotomous forecasts by probability thresholding. This is implemented by choosing a specific threshold and setting any forecast probability above that value to 1 (i.e., a "yes" forecast) and any forecast below it to 0 (i.e., a "no" forecast). The four possible arrangements of forecast-observation pairs can then be represented by a 2×2 contingency table consisting of: true positive forecasts (TP; hits), true negative forecasts (TN; correct rejections), false positive forecasts (FP; false alarms) and false negative forecasts (FN; missed flares). A ROC curve is then a visualisation of the probability of detection (also known as hit rate), POD = TP/(TP + FN), against the probability of false detection (also known as false alarm rate), POFD = FP/(FP + TN), as a function of probability threshold. A skillful forecast will have a higher success-ratio of events (POD) to failure-ratio of non-events (POFD), therefore the closer the curve is to the top left-hand corner the better. The ROC curve is also a visualisation of the True Skill Statistic (TSS = POD − POFD), where the vertical distance of the curve above the diagonal line is the TSS value at that probability threshold (i.e., curves below the diagonal have negative TSS).

Figure 2c displays the ROC curves for both the static (red filled squares) and evolution-dependent (blue open circles) methods, with probability thresholds of $p = 0.01$ and 0.75 as well as the threshold probability corresponding to the maximum TSS value indicated for each method. Initially the ROC curves of both methods behave similarly, with marginally larger TSS for the static case. However, after the threshold probabilities that yield maximum TSS, noticeable divergence occurs with the evolution-dependent curve remaining relatively smooth until converging once again at higher probability thresholds. This is a direct result of the evolution-dependent method containing more forecasts with mid-to-high probabilities relative to the static method (e.g., the sharpness plot of Fig. 2b). Furthermore, the area under the curve (AUC) is a measure of the accuracy of the forecast set, with areas of 1 corresponding to perfect forecasts and 0.5

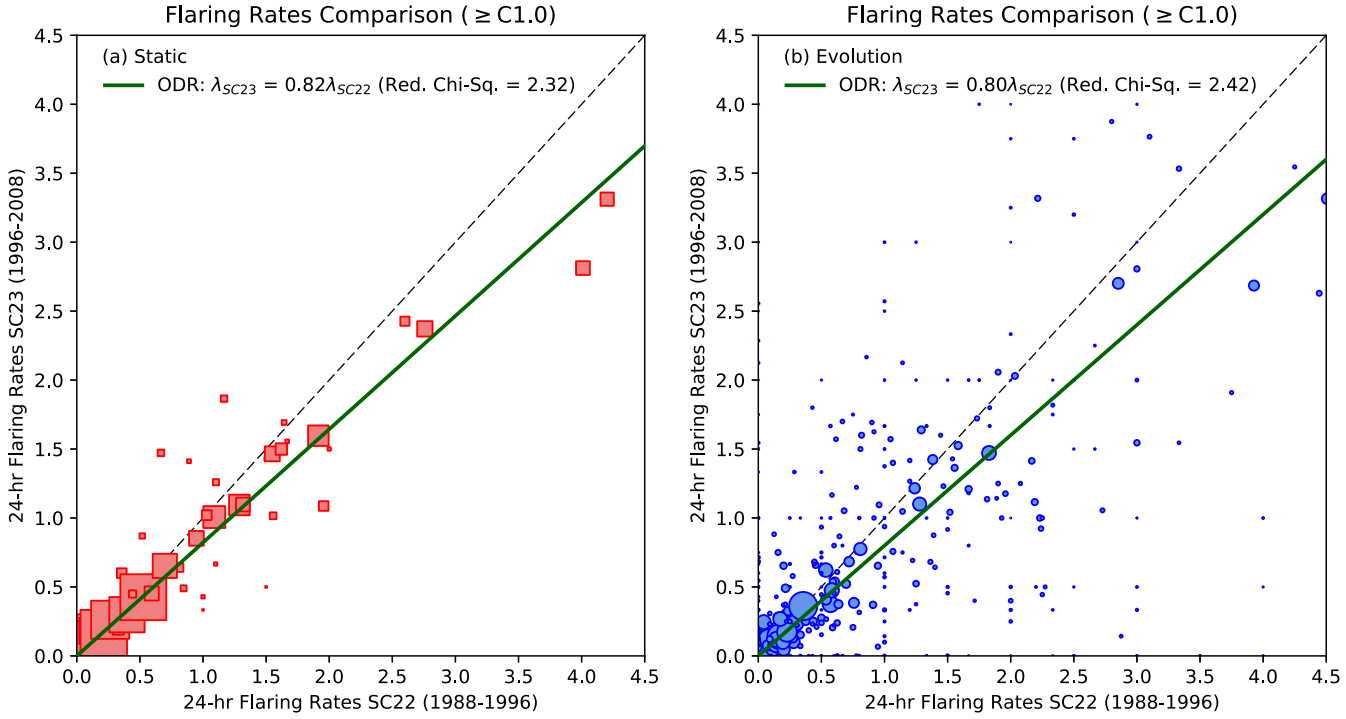

Fig. 3. Comparison of ≥C1.0 24-h flaring rates between SC22 (1988–1996) and SC23 (1996–2008) for the McIntosh static (panel a) and evolution-dependent (panel b) Poisson forecast methods. Dashed diagonal lines indicate the unity relation, while ODR best-fit linear relations are overlaid as thick lines. Best-fit slopes and reduced chi-squared values are also included.

corresponding to no-skill forecasts (indicated by the diagonal dashed line in Fig. 2c). Both methods have AUC values of 0.78, indicating they have comparable dichotomous forecast accuracy when considering performance across the entire probability space. Equivalent figures for ≥M1.0 flares can be found in Appendix A, showing qualitatively similar behaviour between the methods in terms of over-forecasting relative to the observed event frequency and similar values of AUC and maximum TSS.

Considering the overall performance of the static and evolution-dependent methods, both appear to perform similarly when only considering their categorical forecast representation. However, with probabilistic verification metrics it becomes evident that the methods do not achieve the same level of performance. For BSS, the evolution-dependent method was shown to perform better in skill by a value of ≈0.2 when considering either ≥C1.0 or ≥M1.0 flares. In the decomposition of BS, while both methods achieve similar reliability values they differ in resolution, leading to better performance by the evolution-dependent method. In terms of optimising a forecasting method, it is possible to apply forecast-bias corrections to achieve more reliable forecasts. However, for those methods with unique forecast-probability groupings the resolution is fundamentally invariant to such corrections (i.e., with the sets of forecast-observation pairs remaining the same in each unique group, $\overline{o}_k$ and hence resolution in Eq. (4) does not change). Considering that both methods are known to be over-forecasting (see Fig. 2a), in Section 3.2 we consider a basic bias correction to explore what the best performance of the methods could be in an ideal scenario.

3.2 Forecast-bias correction

Based on the results of verification performance for the static and evolution-dependent forecasting methods, we chose to investigate techniques to compensate for the over-forecasting of events in both cases. As both are Poisson-based methods derived from historical average flaring rates, the distributions of flaring rates were examined in the training (SC22) and test (SC23) data sets to investigate if a Cycle-to-Cycle variation existed. Figure 3 presents this comparison for ≥C1.0 flaring rates between SC22 (horizontal axes) and SC23 (vertical axes), for static (panel a) and evolution-dependent cases (panel b). The size of each data point corresponds to the total number of sunspot group occurrences, $M_{\mathrm{tot}} = M_{\mathrm{SC22}} + M_{\mathrm{SC23}}$, that are associated with each McIntosh class (panel a) or each evolution in full McIntosh class, such that larger data points were more frequently observed in both Solar Cycles.

Considering the McIntosh static case in Figure 3a, 49 McIntosh classifications were observed in both the training and test data sets, while 518 full McIntosh-class evolutions were observed in both data sets (Fig. 3b). These rate-rate plots were fit using the Orthogonal Distance Regression (ODR) method, as it takes account of uncertainties in both variables (i.e., $\Delta\lambda_{\mathrm{SC22}} = 1/\sqrt{M_{\mathrm{SC22}}}$ and $\Delta\lambda_{\mathrm{SC23}} = 1/\sqrt{M_{\mathrm{SC23}}}$). Fit intercepts were set to 0 to obtain slopes that can be later compared to rate-correction factors (RCFs) used to examine the possible influence of bias correction on forecast performance (see Sect. 3.3). Dashed diagonal lines in each panel indicate the unity slope (i.e., $\lambda_{\mathrm{SC23}} = \lambda_{\mathrm{SC22}}$), while ODR best-fit lines are displayed as thick lines. For the static method, the

ODR best-fit is found (with a reduced chi-squared of 2.32) to be $\lambda_{SC23} = (0.82 \pm 0.02)\lambda_{SC22}$. As the fit slope is below unity, this indicates that the flaring rates for sunspot groups in the training period (SC22; 1988–1996) are on average higher than the those with the same McIntosh classifications in the test period (SC23; 1996–2008). For the evolution-dependent case, the same behaviour is found (i.e., $\lambda_{SC23} = (0.80 \pm 0.02)\lambda_{SC22}$ with a reduced chi-squared of 2.42). Given that the flaring rates deduced for both methods produce the same relationship within error, this indicates that the rate of flares produced by sunspot groups is Cycle-dependent. These differences in underlying flaring rates between training and testing periods directly contributes to over-forecasting by both methods when using the Poisson approach.

Equivalent figures for ≥M1.0 flares can be found in Appendix A. Qualitatively similar results are presented, but with even greater differences in flaring rates observed between SC22 and SC23 (i.e., $\lambda_{SC23} = (0.52 \pm 0.02)\lambda_{SC22}$ and $\lambda_{SC23} = (0.49 \pm 0.02)\lambda_{SC22}$ for the McIntosh static and evolution-dependent cases, respectively).

3.3 Forecast performance exploration

As mentioned previously, it is possible to alter the performance of a forecasting method using bias-correction techniques. The results of the Cycle-to-Cycle flaring-rate comparison indicate that there is on average a difference in flaring rates for the same sunspot group type between the training and test data sets. Instead of relying solely on the best-fit slopes obtained from the rate-rate comparison, a range of RCFs were examined to find the optimum RCF conditioned on the BSS performance of the "corrected" forecasting methods. This technique works by adjusting the flaring rates obtained from the SC22 training period by multiplication with a RCF to produce new "corrected" flaring rates, with the standard Poisson approach once again applied to produce new "corrected" forecast probabilities.

The results of this analysis are presented in Figure 4, showing the variation with RCF value of BSS and its components following the decomposition given in equation (7). Figure 4a displays the variation of the resolution/uncertainty and reliability/uncertainty terms observed for the McIntosh static case, while the same for the evolution-dependent case is provided in Figure 4b. The BSS-decomposed uncertainty term is constant (with a value of 0.146) and equal in both cases, as it only depends on the climatological frequency of events that is common to both methods. It is important to note that when using the decomposition of BSS correctly (i.e., when the forecast method comprises of distinctly unique forecast-probability groups), the resolution of the method is invariant under the bias correction performed by applying the RCF; evidenced by the normalised resolution term remaining constant as a function of RCF in both cases (i.e., horizontal lines). As the uncertainty-normalised reliability term is always positive and contributes negatively to BSS (see Eqs. (4) and (7)), achieving the smallest possible value is highly desirable.

For the McIntosh static method in Figure 4a, the uncertainty-normalised reliability is optimized (i.e., minimized) at a value of 0.08 for a RCF of 0.32. Similarly for our evolution-dependent method, the minimum normalised reliability value of 0.11 is achieved for a RCF of 0.48 (Fig. 4b). For both cases this leads to the opposite behaviour for BSS as a function of RCF (Fig. 4c), with maximum BSS values of 0.09 and 0.20 achieved for the static and evolution-dependent methods, respectively. The optimal BS decomposed values and BSS are presented for ≥C1.0 and ≥M1.0 flares in Table 2. As mentioned before, the main difference between the two forecast methods is that our new evolution-based method achieves a resolution nearly twice that of the original static method, with uncertainty-normalised resolution values of 0.18 (static) and 0.31 (evolution-dependent). Optimising method reliabilities using a simple (admittedly *post facto*) RCF technique as presented here leads to an approximately two-fold increase in BSS from the values in Table 1.

"Corrected" reliability diagrams and ROC curves are presented in Figure 5 using the optimized RCF values conditioned on maximising BSS to visualise the effect it has on forecast performance. The reliability diagrams of Figure 5a confirm the McIntosh static (red filled squares) and evolution-dependent (blue open circles) forecast probabilities are both shifted to smaller values due to the RCFs applied being less than unity. Although this improves BSS for both methods, it does not appear to achieve a more reliable visual representation for the static method as several points appear to lie far from the line of perfect reliability (Fig. 2a, red filled squares for comparison). In contrast, the evolution-dependent method appears to achieve a much more reliable visual representation than its equivalent uncorrected version (Fig. 2a, blue open circles) with more points lying close to, or on, the line of perfect reliability. The "corrected" version of the ROC curves are presented in Figure 5b, with no significant changes to the overall shape, area under the curve or maximum departure from the diagonal no-skill line. This is to be expected, as the application of the RCF only acts to shift the probability thresholds that the dichotomous categorical forecast statistics are calculated from (i.e., the forecast observation outcomes are unaltered). This could have implications for use in an operational situation: if bias-corrections are applied to create more reliable probabilistic forecasts, then the choice of probability threshold for evaluating the performance of subsequently-derived categorical metrics (or issuing of yes/no flare forecasts) needs to be reconsidered.

Equivalent plots for the RCF analysis and "corrected" reliability diagrams and ROC curves for ≥M1.0 flares can be found in Appendix A, showing qualitatively similar results to the ≥C1.0 case (i.e., improvement in reliability and BSS).

4 Discussion and conclusion

In this paper, we have examined the evolution of McIntosh sunspot group classifications and its application as a method for forecasting solar flares. Flaring rates calculated from sunspot-group evolution in McIntosh classifications during SC22 were used to produce probabilities for ≥C1.0 and ≥M1.0 flares within 24-h forecast windows under the assumption of Poisson statistics. The reason for excluding flares below these magnitudes is the high background solar X-ray flux level at solar maximum that obscures B-class and lower flares. Additionally, due to the small number of X-class flares we chose to exclude the analysis of X-class and above as the large statistical errors lead to difficult interpretation of results.

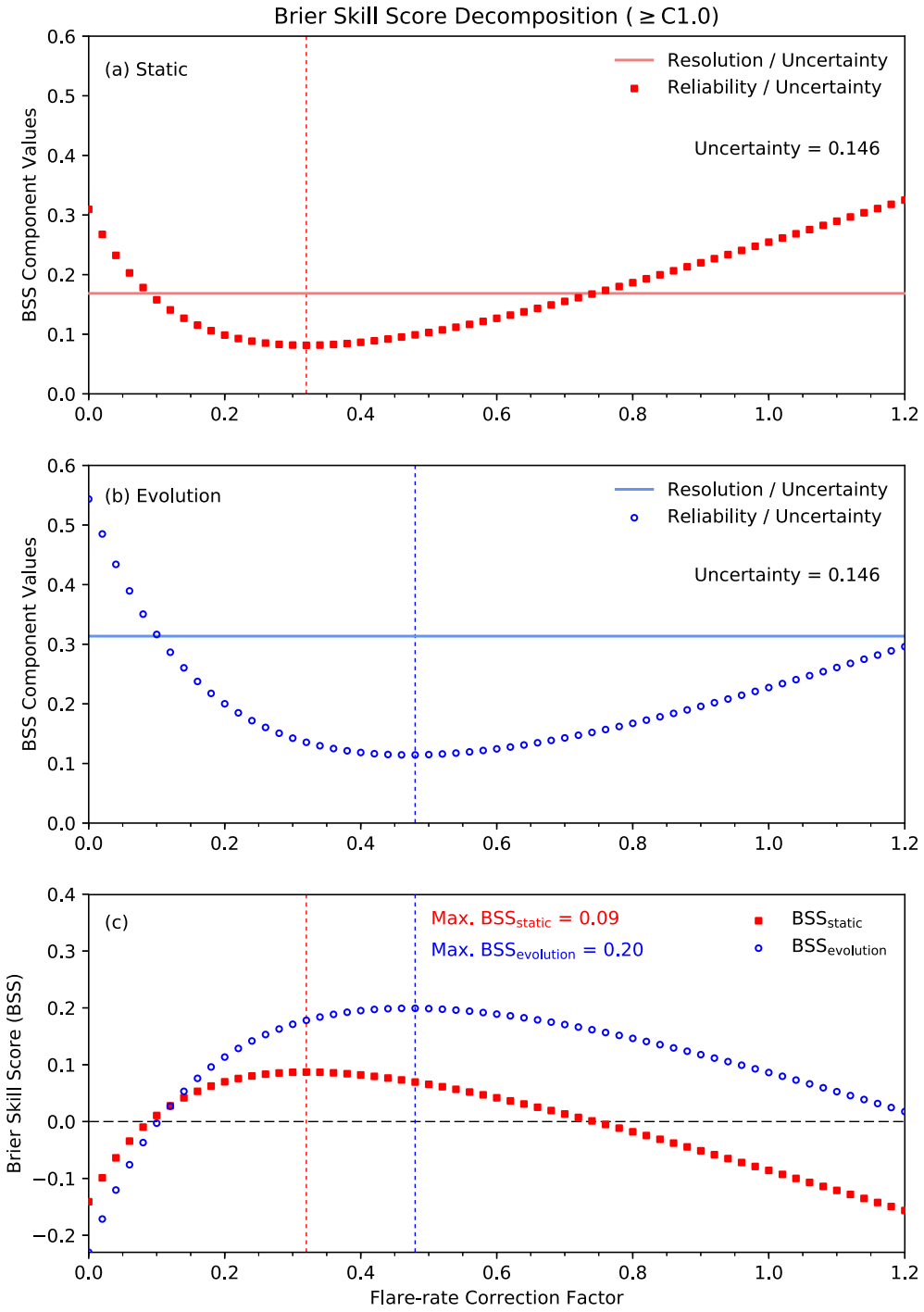

Fig. 4. Brier skill score (BSS) decomposition for the McIntosh static (panel a) and evolution-dependent (panel b) forecast methods for ≥C1.0 flares. BS components of reliability (data points), resolution (solid horizontal lines), and uncertainty (printed values) are displayed in panels a and b as a function of rate-correction factor (RCF) applied to the SC22 flaring rates. The resulting BSS is presented in panel c, also as a function of RCF applied to the SS22 flaring rates, for the static (red filled squares) and evolution-dependent (blue open circle) methods, with maximum values of BSS indicated by vertical dashed lines for both cases.

Similar to the results of McCloskey et al. (2016), we find that upward evolution in at least one McIntosh classification component leads to higher flaring rates, with lower flaring rates occurring for downward evolution (Fig. 1). Additionally, when sunspot groups evolve across multiple McIntosh classification components at the same time this behaviour is amplified – i.e., increasingly higher (lower) flaring rates observed for greater upward (downward) evolution.

These flaring rates were converted to Poisson probabilities and applied to an independent test data set from SC23 to assess forecast performance, both for the original static point-in-time McIntosh forecasting method and our new evolution-dependent

Table 2. Optimized RCF-adjusted decomposed Brier score (BS) components and Brier skill score (BSS) for the McIntosh static and evolution-dependent forecast methods.

Flaring magnitude	Forecast method	Applied SC22 RCF	BS components			BSS
			Reliability	Resolution	Uncertainty	
≥C1.0	Static	0.32	0.012	0.025	0.146	0.09
≥C1.0	Evolution	0.48	0.017	0.046	0.146	0.20
≥M1.0	Static	0.20	0.001	0.003	0.038	0.06
≥M1.0	Evolution	0.30	0.005	0.009	0.038	0.09

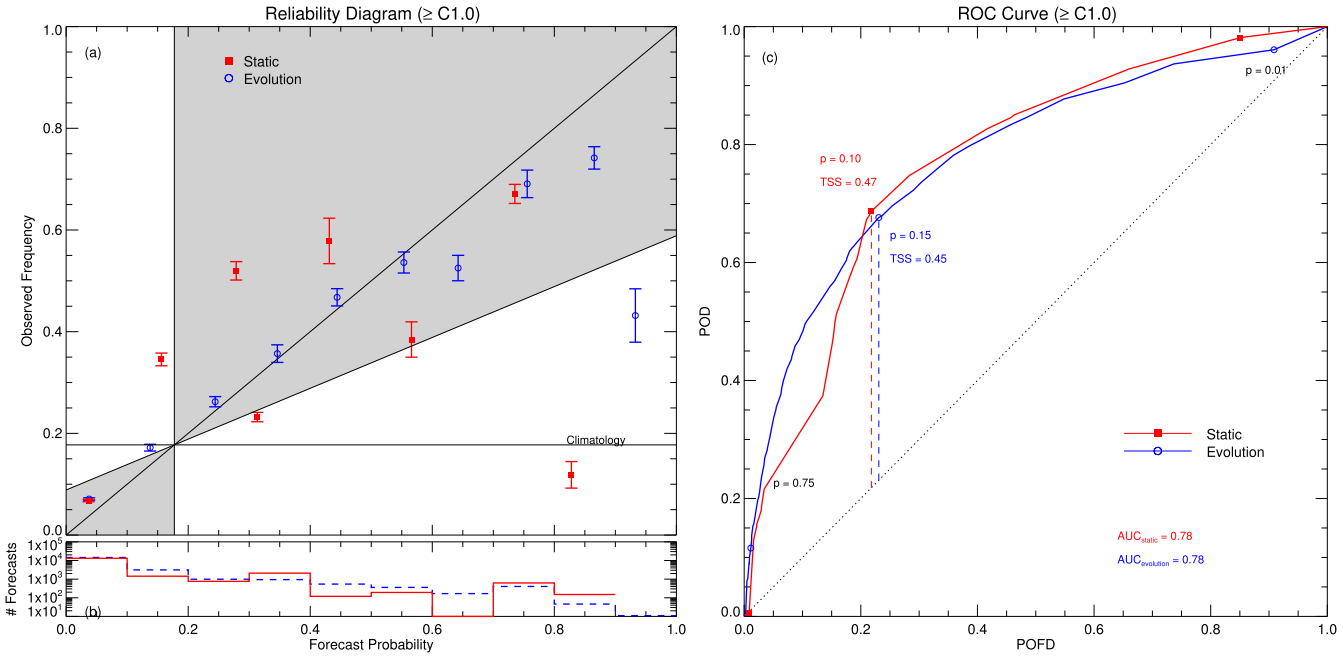

Fig. 5. As Figure 2, but using the BSS-optimised RCFs of 0.32 and 0.48 applied to the SC22 ≥C1.0 flaring rates for the McIntosh static and evolution-dependent forecast methods, respectively.

method. BSS was calculated for both, with the evolution-dependent method achieving a positive value for ≥C1.0 flares ($BSS_{evolution} = 0.09$), indicating that its performance surpasses that of climatology. In contrast, the static method performed worse than climatology ($BSS_{static} = -0.09$). Importantly, the determining factor for the difference in performance is that the evolution-dependent method achieves greater resolution than its static counterpart, indicating that the observed event occurrence averaged across the individual full-McIntosh class evolutions (i.e., unique forecast probability groups in the decomposed form of BS) is more separated from climatology than the same quantity averaged across individual static McIntosh classes. For ≥M1.0 flares the evolution-dependent method again performs better than the static method, but as both BSS values are negative ($BSS_{evolution} = -0.15$ and $BSS_{static} = -0.36$) this indicates that they do not perform as well as climatology. Reliability diagrams and ROC curves were also investigated, with a bias of over-forecasting clear in both methods (Fig. 2a and b).

This tendency to over-forecast was investigated by comparing the flaring rates for the training data from SC22 with those of the test data from SC23 using an ODR technique to fit the rate-rate relations. Considering previous studies, it has been shown that the level of activity in SC23 is lower compared to earlier Cycles. For example, Joshi and Pant (2005) report that the number of Hα flare events was lower in SC23 compared to SC21 and SC22, while Joshi et al. (2015) found that there was a significant decrease in the total soft X-ray flare index (a measure of flare activity) in SC23 compared to SC21 and SC22. These results agree well with our finding SC23 rates being ≈80% and ≈50% of those in SC22 for ≥C1.0 and ≥M1.0 flares, respectively (Figs. 3 and A.2).

To explore the maximum-achievable performance by the McIntosh-Poisson forecasting methods, a range of RCFs were explored through application to the original SC22 flaring rates to bias-correct the forecast probabilities. The optimal value of RCF for ≥C1.0 flares (i.e., that achieving maximum BSS) was found to be 0.32 for the static method, while the evolution-dependent method has a weaker correction factor of 0.48 (Fig. 4). Interestingly, these RCFs differ from the Cycle-to-Cycle ODR linear rate-rate slopes of ≈0.80, although the ODR-determined value is admittedly obtained with no

information feeding back from the application of the adjusted flaring rates in forecasting. The resulting maximum values for corrected BSS were found to be 0.09 and 0.20 for the static and evolution-dependent methods, respectively. These correspond to a two-fold increase in BSS that confirms the lowering of forecast probabilities issued for SC23 yields better performance for both methods, evidenced by improved reliability diagrams (Fig. 5a). To put these values in context, Barnes et al. (2016) compared several flare-forecasting methods using standard verification metrics to assess performance. To ensure direct comparison of the methods, a common data set was used where all methods issued forecasts for each data entry, analogous to daily operational flare forecasts and therefore the most suitable for comparing to the operational methods presented here. The maximum BSS achieved for ≥C1.0 flares in 24-h forecast windows by any of the methods in Barnes et al. (2016) was 0.32 (see their Tab. 4). After optimal bias-correction was determined and applied, our McIntosh evolution-dependent method achieved a BSS approaching but still less than this (i.e., $BSS^{corr}_{evolution} = 0.20$).

It is noted that the bias-correction method applied here determines the systematic differences in flaring rates between training and test periods from *post facto* analysis. To be truly operational, the application of pre-forecast bias correction requires prior knowledge of these differences in rates. Therefore, predictions for the next Solar Cycle could provide the bias-correcting RCF for the next forecast test period.

The authors thank Dr Chris Balch (NOAA/SWPC) for providing the 1988–1996 data used in this research and Dr Graham Barnes (NWRA/CoRA) for useful discussions on BS decomposition. A.E.McC. is supported by an Irish Research Council Government of Ireland Postgraduate Scholarship and D.S.B. is supported by the European Union Horizon 2020 research and innovation programme under grant agreement No. 640216 (FLARECAST project; flarecast.eu). The editor thanks two anonymous referees for their assistance in evaluating this paper.

References

Ahmed OW, Qahwaji R, Colak T, Higgins PA, Gallagher PT, Bloomfield DS. 2013. Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Sol Phys* **283**: 157–175. DOI:10.1007/s11207-011-9896-1.

Barnes G, et al. 2016. A Comparison of flare forecasting methods. I. Results from the "all-clear" workshop. *Astrophys J* **829**: 89. DOI: 10.3847/0004-637X/829/2/89.

Bloomfield DS, Higgins PA, McAteer RTJ, Gallagher PT. 2012. Toward reliable benchmarking of solar flare forecasting methods. *Astrophys J Lett* **747**: L41. DOI:10.1088/2041-8205/747/2/L41.

Bobra MG, Couvidat S. 2015. solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *Astrophys J* **798**: 135. DOI:10.1088/0004-637X/798/2/135.

Bornmann PL, Shaw D. 1994. Flare rates and the McIntosh active-region classifications. *Sol Phys* **150**: 127–146. DOI:10.1007/BF00712882.

Colak T, Qahwaji R. 2009. Automated Solar Activity Prediction: a hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather* **7**: S06001. DOI:10.1029/2008SW000401.

Cortie AL. 1901. On the types of sun-spot disturbances. *Astrophys J* **13**: 260. DOI:10.1086/140816.

Crown MD. 2012. Validation of the NOAA Space Weather Prediction Center's solar flare forecasting look-up table and forecaster-issued probabilities. *Space Weather* **10**: S06006. DOI:10.1029/2011SW000760.

Devos A, Verbeeck C, Robbrecht E. 2014. Verification of space weather forecasting at the Regional Warning Center in Belgium. *J Space Weather Space Clim* **4**: A29. DOI:10.1051/swsc/2014025.

Gallagher PT, Moon Y-J, Wang H. 2002. Active-region monitoring and flare forecasting I. Data processing and first results. *Sol Phys* **209**: 171–183. DOI:10.1023/A:1020950221179.

Hale GE, Ellerman F, Nicholson SB, Joy AH. 1919. The magnetic polarity of sun-spots. *Astrophys J* **49**: 153. DOI:10.1086/142452.

Hathaway DH, Wilson RM, Reichmann EJ. 1999. A synthesis of solar cycle prediction techniques. *J Geophys Res* **104**: 22375–22388. DOI:10.1029/1999JA900313.

Joshi B, Bhattacharyya R, Pandey KK, Kushwaha U, Moon Y-J. 2015. Evolutionary aspects and north-south asymmetry of soft X-ray flare index during solar cycles 21, 22, and 23. *Astron Astrophys* **582**: A4. DOI:10.1051/0004-6361/201526369.

Joshi B, Pant P. 2005. Distribution of H flares during solar cycle 23. *Astron Astrophys* **431**: 359–363. DOI:10.1051/0004-6361:20041986.

Lee K, Moon Y-J, Lee J-Y, Lee K-S, Na H. 2012. Solar flare occurrence rate and probability in terms of the sunspot classification supplemented with sunspot area and its changes. *Sol Phys* **281**: 639–650. DOI:10.1007/s11207-012-0091-9.

McCloskey AE, Gallagher PT, Bloomfield DS. 2016. Flaring rates and the evolution of sunspot group mcintosh classifications. *Sol Phys* **291**: 1711–1738. DOI:10.1007/s11207-016-0933-y.

McIntosh PS. 1990. The classification of sunspot groups. *Sol Phys* **125**: 251–267. DOI:10.1007/BF00158405.

Mitra AP, ed. Ionospheric effects of solar flares, vol. 46 of astrophysics and space science library, 1974. DOI:10.1007/978-94-010-2231-6.

Murphy AH. 1973. A new vector partition of the probability score. *J Appl Meteorol* **12**: 595–600. DOI:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murray SA, Bingham S, Sharpe M, Jackson DR. 2017. Flare forecasting at the Met Office Space Weather Operations Centre. *Space Weather* **15**: 577–588. DOI:10.1002/2016SW001579.

Priest ER, Forbes TG. 2002. The magnetic nature of solar flares. *Astron Astrophys Rev* **10**: 313–377. DOI: 10.1007/s001590100013.

Waldmeier M. 1947. Heliographische Karten der Photosphäre für das Jahr 1946. *Publ Zürich Obs* **9**: 1.

Wheatland M.S. 2001. Rates of flaring in individual active regions. *Sol Phys* **203**: 87–106. DOI:10.1023/A:1012749706764.

Wheatland MS. 2005. A statistical solar flare forecast method. *Space Weather* **3**: S07003. DOI:10.1029/2004SW000131.

Cite this article as: McCloskey AE, Gallagher PT, Bloomfield DS. 2018. Flare forecasting using the evolution of McIntosh sunspot classifications. *J. Space Weather Space Clim.* **8**: A34

Appendix A : forecast verification of flares at/above M1.0

Here we present equivalent figures to those in Section 3, but for ≥M1.0 flares. Reliability diagrams and ROC curves (equivalent to Fig. 2) are plotted in Figure A.1. Following from this, the flare rate comparison between SC22 and SC23 (equivalent to Fig. 3) is shown in Figure A.2. Finally, the BSS decomposition as a function of RCF and the "corrected" reliability diagrams and ROC curves are provided in Figures A.3 and A.4 , respectively (equivalent to Figs. 4 and 5).

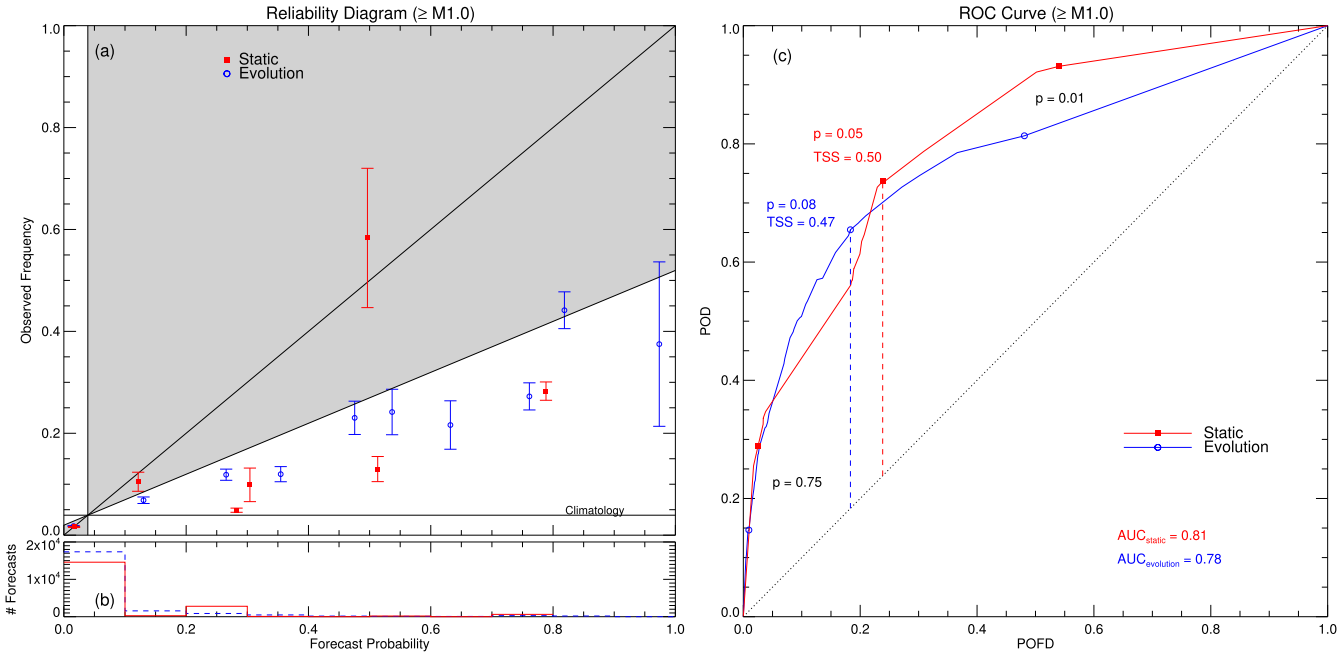

Fig. A.1. Reliability diagrams (panel a), sharpness (i.e., probability occurrence) plots (panel b), and ROC curves (panel c) for ≥M1.0 flares. Data for the McIntosh static forecast method are indicted by red filled squares (panels a and c) and solid histogram (panel b), while the evolution-dependent method is depicted by blue open circles (panels a and c) and dashed histogram (panel b).

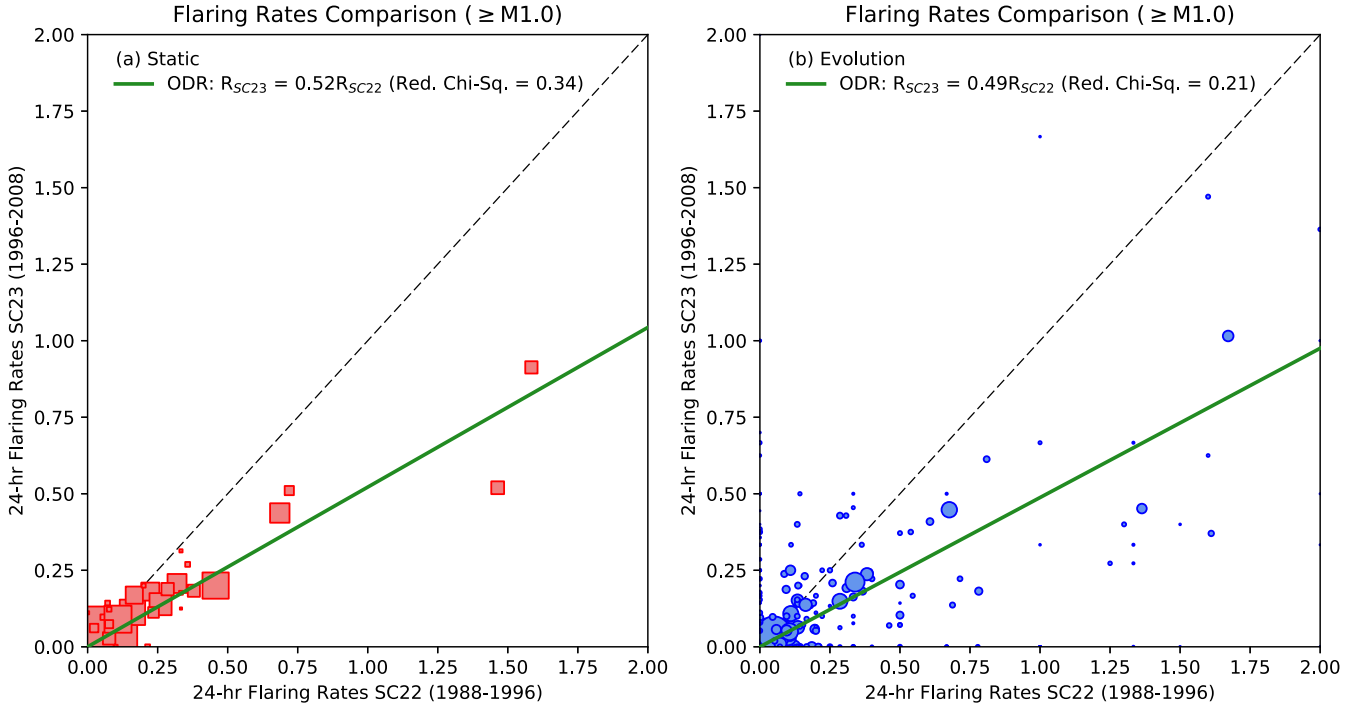

Fig. A.2. Comparison of ≥M1.0 24-h flaring rates between SC22 (1988–1996) and SC23 (1996–2008) for the McIntosh static (panel a) and evolution-dependent (panel b) Poisson forecast methods. Dashed diagonal lines indicate the unity relation, while ODR best-fit linear relations are overlaid as thick lines. Best-fit slopes and reduced chi-squared values are also included.

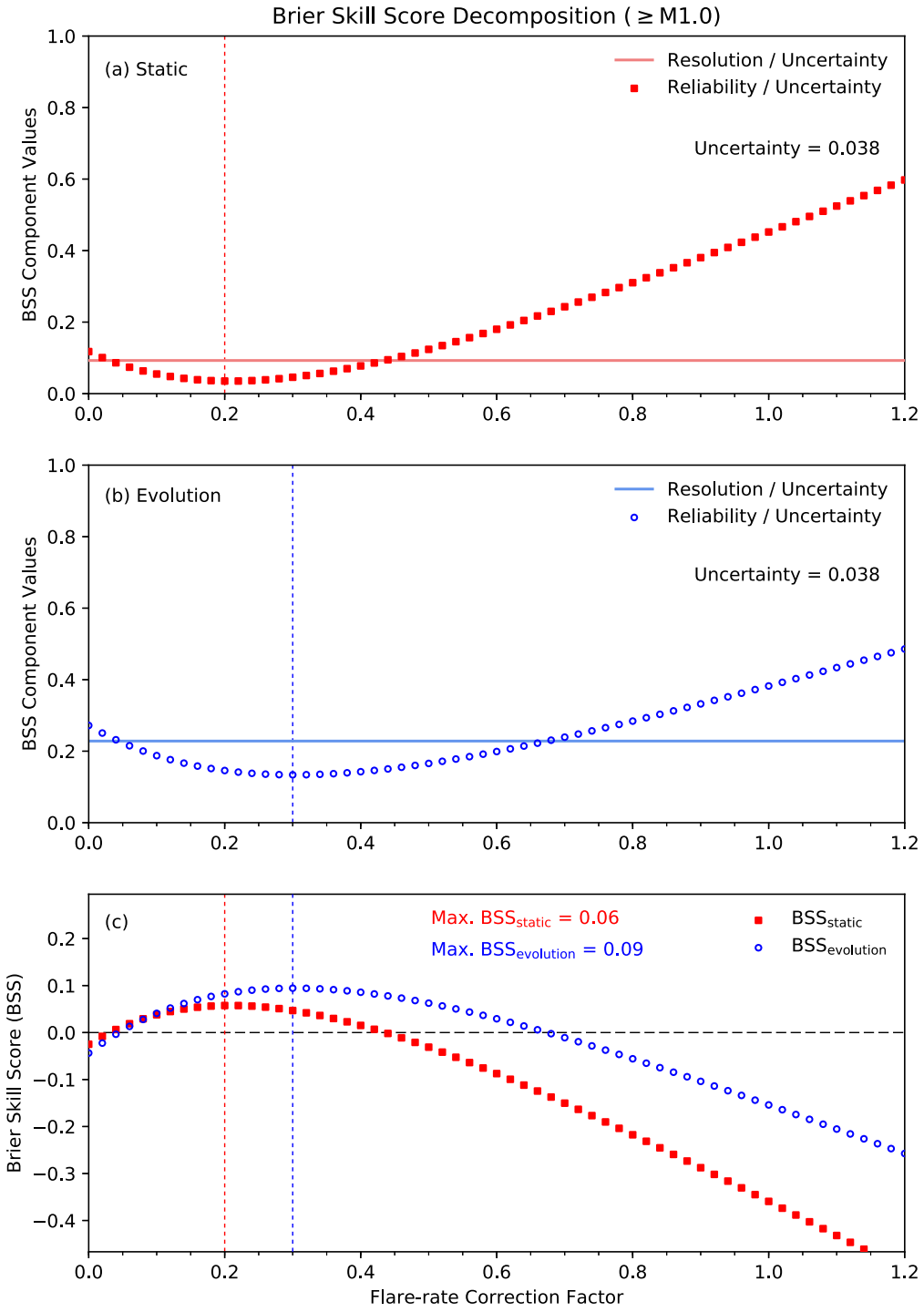

Fig. A.3. Brier skill score (BSS) decomposition for the McIntosh static (panel a) and evolution-dependent (panel b) forecast methods for ≥M1.0 flares. BS components of reliability (data points), resolution (solid horizontal lines), and uncertainty (printed values) are displayed in panels a and b as a function of rate-correction factor (RCF) applied to the SC22 flaring rates. The resulting BSS is presented in panel c, also as a function of RCF applied to the SS22 flaring rates, for the static (red filled squares) and evolution-dependent (blue open circle) methods, with maximum values of BSS indicated by vertical dashed lines for both cases.

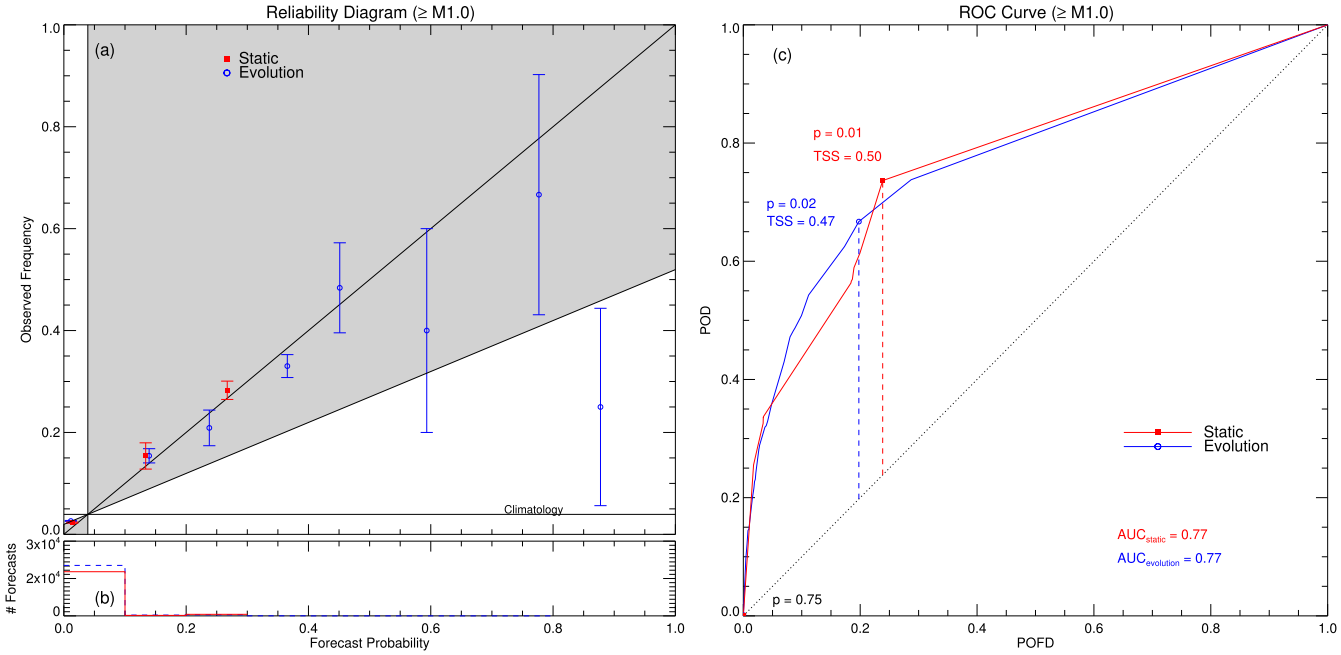

Fig. A.4. As Figure A.1, but using the BSS-optimised RCFs of 0.20 and 0.30 applied to the SC22 ≥M1.0 flaring rates for the McIntosh static and evolution-dependent forecast methods, respectively.