Verification of Postprocessed Summer Temperature Forecasts at U.K. Sites Using Observed Climate-Based Thresholds

MICHAEL SHARPE

Weather Science, Met Office, Exeter, Devon, United Kingdom

EMILY STEWART

University of St Andrews, St Andrews, Scotland

(Manuscript received 11 September 2018, in final form 22 February 2019)

ABSTRACT

The full range of ensemble forecast members, expressed in percentile format, can give additional valuable information to users compared to a deterministic solution. This study contains the first rigorous assessment of the operational U.K. site-specific percentile forecast generated by the Met Office. Maximum meteorological summer daytime temperature forecasts issued between 2014 and 2016 are analyzed using the ranked probability score (RPS), ranked probability skill score (RPSS), categorized mean squared error (MSE), quantile skill score (QSS), and relative economic value (REV). Site-specific observed climatology is used to define the temperature threshold for each category (where applicable) thereby ensuring identical categorical event base rates across all 99 sites considered. Forecast ranges between 6 and 120 h are assessed, with the RPS decomposition indicating no perceivable change in the reliability yet an almost linear decrease with forecast range solely due to a near-linear increase in the resolution. Using the categorized MSE (the deterministic equivalent to the RPS), the probabilistic forecast is shown to possess more skill than its deterministic counterpart and the disparity between these scores increases with the forecast range. This finding is reinforced by a REV assessment; this indicates that the economic value associated with the probabilistic envelope is greater than that associated with the deterministic solution at the majority of cost–loss ratios. The QSS appears to be well correlated with the RPSS ($r_s = 0.852$ at $T + 24$) and identifies the outlying quantiles of the probabilistic forecast as being the least skillful.

1. Introduction

Subjective use of climatology to contextualize weather conditions has been the recent focus of the meteorological community particularly in the context of warnings of events that are likely to cause impacts (Neal et al. 2014). Objective use of site-specific climatological information to identify thresholds has also been the basis of a paper by Magnusson et al. (2014) and more recently by Sharpe et al. (2018). The present study extends this methodology to the everyday summer daytime maximum temperatures reported by the U.K. surface observing network and forecast daily by postprocessed probabilistic and deterministic site-specific numerical weather prediction models developed by the Met Office. Four temperature thresholds

are used to align the continuous output from these models with public perceptions, thereby categorizing each forecast as a very cool, cool, typical, warm, or very warm summer day. Therefore, the focus of this study is an attempt to adapt established verification metrics to reflect the perceptions of forecast users in an objective way, since a member of the public is (arguably) more likely to interpret the forecast in this way. At each site, the numerical value of each threshold is obtained by selecting a percentile from the cumulative distribution function formed by the long-term climatology at that site. The use of climatological information in the assessment of numerical weather prediction models is not a new concept. NCEP has used equally likely climate-based bins for the evaluation of its models since 1995 (Zhu et al. 1996). However, in the context of gridded models, reanalysis data are often used for the climatology (Zhu and Toth 2008); whereas, restricting the analysis to station sites allows the use of observed climatology

Corresponding author: Michael Sharpe, michael.sharpe@metoffice.gov.uk

instead. Site-specific data are also more appropriate to members of the public who access data for their location using the Met Office app or website. The frequency of occurrence (base rate) of each category will be the same at every site because the same percentile is chosen for each of the four thresholds. However, the resulting numerical values will depend upon the site-specific climatology, so sites with distinctly different climatologies will have very different threshold values. Although the Met Office does not currently convey its operational forecasts in terms of site-specific climatological categories, such expression would be possible and of potential benefit to those who crave contextualization of weather information. Currently temperatures are presented in terms of a number (degrees Celsius in the United Kingdom) and a single all-site color range. However, both of these can be misleading to members of the public. For instance, some individuals may only appreciate temperature in degrees Fahrenheit. Similarly, an all-site color range is not an adequate representation since U.K. site climatologies vary significantly. For example, when an all-site color range is adopted the blue (commonly used to represent cold) is more likely to represent typical conditions in northerly or high-altitude sites.

In this study three measures are used to assess the site-specific probabilistic forecasts: the ranked probability score (RPS) (Epstein 1969), the quantile score (QS) (Ben Bouallègue et al. 2015; Bentzien and Friederichs 2014), and the relative economic value (REV) (Richardson 2000). Both the REV and RPS are used to compare the skill associated with both the deterministic and the probabilistic forecast, where comparison with the latter is made using a categorized version of the mean squared error (MSE). The QS has been adopted in preference to the arguably more commonly used continuous ranked probability score (CRPS) because the site-specific, postprocessed, probabilistic forecast is expressed in terms of quantiles (rather than ensemble members). An assessment of the effect on performance of using climate-based categories is achieved through a comparison of the QS and the RPS.

Section 2 contains a discussion of the site-specific climatology generation; section 3 describes the methodologies used to verify the forecast; section 4 presents the results of analyzing forecasts issued during the meteorological summers of 2014, 2015, and 2016; and section 5 gives brief conclusions.

2. Site-specific climatology

Following the rationale outlined in Magnusson et al. (2014) temperature data from the quality controlled

U.K. observing network (Met Office 2012) rather than from model analyses were used to generate a climatological cumulative distribution function CDF_i^c at site i .

With the help of the National Climate Information Centre a CDF_i^c was generated at each U.K. observing site using the 30-yr time series of quality controlled observational data collected between 1983 and 2012. However, 45 of the 144 sites available in the Met Office meteorological database were excluded from the analysis because less than 20 years of summer daytime maximum temperature data were available during this period and this was deemed insufficient to adequately represent the climatology. At the remaining 99 sites categories are used to assess the forecast. To aid forecast communication to users (who typically struggle to appreciate climatological percentiles) category thresholds were chosen based on a whole number of days. Therefore, the daily meteorological summer (June, July, August) maximum temperature condition was defined as follows:

- Very cool if it was one of the 10 coolest days of a typical meteorological summer;
- Cool if it corresponds to a temperature between the 10th and 20th coolest day of a typical meteorological summer;
- Typical if it corresponds to a temperature between the 20th coolest and 20th warmest day of a typical meteorological summer;
- Warm if it corresponds to a temperature between the 10th and 20th warmest day of a typical meteorological summer; and
- Very warm if it was one of the 10 warmest days of a typical meteorological summer.

Although these categories correspond to unusual percentile choices (viz., CDF_i^c values of 0.1087, 0.2174, 0.7826, and 0.8913) they allow the forecast to be communicated in a natural and simple way, without the need to mention percentiles of the climatological distribution. Figure 1 displays the distribution formed by the upper threshold (which also serves as the lower threshold of the higher category) of the lowest four categories at each valid site. A negative skew is immediately obvious with far more sites lying toward the warmer end of the scale. This reflects the unequal geographical distribution of sites within the United Kingdom, perhaps in line with population proportions. The site located on Cairngorm Summit (57.117°N, 3.633°W; altitude 1245 m) stands out as the only outlier; it is interesting to note that the warmest 10-day threshold at this site is cooler than the coolest 10-day threshold at 86 of the remaining 98 sites. None of the other sites considered in this study have an

FIG. 1. Box-and-whisker plots showing the climatological distribution of summer maximum temperature category maximums for U.K. synoptic sites for which at least 20 years of climatology are available between 1983 and 2012. The upper temperature bounds for the 10th coolest day, 20th coolest day, 20th warmest day, and 10th warmest day are expressed by blue, green, orange, and red, respectively.

altitude > 500 m. There is a tendency for the maximum daily summer temperature to decrease with increasing latitude; however, Cairngorm Summit remains an outlier (of the distribution formed by the coolest three categories) even when the sites are restricted to those only located in Scotland. Fair Isle (59.533°N , 1.633°W ; altitude 68 m) is the only other outlier for the distributions formed by the coolest 10- and 20-day category thresholds at Scottish sites.

The subfigures of Fig. 2 show the geographical distribution of the threshold values in Fig. 1. Two temperature scales have been used to display this data thereby facilitating a clearer reflection of the spread of temperatures at sites across the United Kingdom and enabling intersite comparison. The warmest site thresholds are located in the southeast (as expected) and the temperatures appear to decrease with increasing distance from London with the coolest (blue) stations located in Scotland and the north of England, west Wales, and the far southwest of England. There also appears to be an indication that at similar latitudes, coastal and mountainous sites tend to be cooler than inland sites at lower altitudes. The warmest site is St. James Park, where

- Very cool is a temperature less than 18.2°C ;
- Cool is a temperature greater than or equal to 18.2°C and less than 19.5°C ;
- Typical is a temperature greater than or equal to 19.5°C and less than 25.5°C ;
- Warm is a temperature greater than or equal to 25.5°C and less than 27.6°C ; and

- Very warm is a temperature greater than or equal to 27.6°C .

For comparison, at the coolest site (the Cairngorm Mountains):

- Very cool is a temperature less than 11.0°C ;
- Cool is a temperature greater than or equal to 11.0°C and less than 11.9°C ;
- Typical is a temperature greater than or equal to 11.9°C and less than 14.7°C ;
- Warm is a temperature greater than or equal to 14.7°C and less than 15.5°C ; and
- Very warm is a temperature greater than or equal to 15.5°C .

The values in Fig. 2 are used to generate the results in section 4.

3. Verification methodology

a. Deterministic and probabilistic data

Postprocessed, site-specific, Met Office forecasts are derived by blending model data from various deterministic and ensemble model forecasts (Moseley 2011); this procedure produces a spread of possible scenarios, each of which is accompanied by a probability of occurrence and a deterministic forecast. The concept is to create an optimal blend of all available observations and models to produce a most accurate site-specific forecast. Gridded nowcasts, regional (high resolution), and global (lower resolution) numerical weather prediction deterministic and ensemble models (generated by both Met Office and ECMWF) are used for this purpose. Where site-specific observations are available, Kalman filtering is also applied and then the different forecasts are combined using lagging and blending techniques.

Deterministic forecast data are available via the online Met Office forecast database DataPoint (Met Office 2018; however, although probabilistic model data are currently unpublished it is available on request from the Met Office via a Freedom of Information request to enquiries@metoffice.gov.uk, citing this paper). The postprocessed, probabilistic, site-specific, cumulative forecast distribution function CDF_i^c , which the Met Office routinely produces generates the following 15 percentiles: 0.025, 0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.90, 0.95, and 0.975. These quantiles are denoted by z_1 to z_{15} . A probabilistic forecast provides more information about the spread of possible outcomes, enabling more informed decisions and potentially providing a more reliable service. In contrast, a

FIG. 2. Geographical distribution of the threshold boundaries used to define site-specific temperature categories at each observing site where at least 20 years of climatology are available between 1983 and 2012. (top left) The upper boundary of the very cool category (10th coolest typical summers day), (top right) the upper boundary of the cool category (20th coolest typical summers day), (bottom left) the upper boundary of the typical category (20th warmest typical summers day), and (bottom right) the upper boundary of the warm category (10th warmest typical summers day).

deterministic forecast consists of just one value—a blend of the ensemble means and deterministic model solutions. Although for the forecast user deterministic data may appear more precise and useful (since little interpretation is required) this simplicity can be misleading because it can give a false impression of certainty.

b. Ranked probability score decomposition

The negatively orientated ranked probability score (Epstein 1969) is a proper score for probabilistic forecasts based on ranked categories. The RPS ranges from 1 to 0. A site-specific value of the RPS is evaluated using Eq. (1):

$$\text{RPS}_i = \frac{1}{K_i} \sum_{m=1}^M \frac{1}{M-1} \sum_{k=1}^{K_i} [\text{CDF}_i^{f_k}(m) - H(o_i^k - t_i^m)]^2, \quad (1)$$

where m denotes the temperature category threshold, K_i is the number of forecasts at site i , $\text{CDF}_i^{f_k}(m)$ is the value of the cumulative probability density function at site i for forecast k , t_i^m denotes the upper boundary to the temperature category m at site i , o_i^k is the observed value at site i during forecast k , and

$$H(o_i^k - t_i^m) = \begin{cases} 0, & \text{if } o_i^k \geq t_i^m \\ 1, & \text{if } o_i^k < t_i^m. \end{cases}$$

Then an all-site value for the RPS is evaluated using

$$\text{RPS} = \frac{\sum_i K_i \times \text{RPS}_i}{\sum_i K_i}. \quad (2)$$

The RPS is the mean squared error of a probabilistic multicategory forecast; therefore, a comparison between probabilistic and deterministic forecasts is achieved by making a direct comparison between Eq. (1) for the RPS and the categorized version of the MSE expressed by

$$\text{MSE}_i = \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{1}{M-1} \sum_{m=1}^M [H(f_i^k - t_i^m) - H(o_i^k - t_i^m)]^2, \quad (3)$$

where f_i^k denotes deterministic forecast k at site i .

The Brier score (BS) (Brier 1950) compares forecast probabilities with binary observations in relation to a single threshold instead of multiple categories. Equation (4) gives an expression for the BS at site i

$$\text{BS}_i(m) = \frac{1}{K_i} \sum_{k=1}^{K_i} [\text{CDF}_i^{f_k}(m) - H(o_i^k - t_i^m)]^2; \quad (4)$$

therefore, RPS_i is simply the mean of $\text{BS}_i(m)$ over each category, that is,

$$\text{RPS}_i = \frac{1}{M-1} \sum_{m=1}^M \text{BS}_i(m). \quad (5)$$

Sanders (1963) and Murphy (1973) describe how the BS decomposes into reliability (REL), uncertainty (UNC), and resolution (RES) terms; while UNC is forecast independent, both RES and REL are important indicators of forecast quality. Reliable forecasts contain probabilities that are close to the observed frequency of occurrence, whereas well-resolved forecasts contain probabilities that differ significantly from the event base rate. The BS decomposition is expressed by

$$\text{BS}_i(m) = \text{REL}_i(m) - \text{RES}_i(m) + \text{UNC}_i(m), \quad (6a)$$

where

$$\text{REL}_i(m) = \sum_{j=1}^J \frac{n_i^j(m)}{K_i} \left[\frac{o_i^j(m)}{n_i^j(m)} - p^j \right]^2, \quad (6b)$$

$$\text{RES}_i(m) = \sum_{j=1}^J \frac{n_i^j(m)}{K_i} \left[\frac{o_i^j(m)}{n_i^j(m)} - \bar{o}_i(m) \right]^2, \quad \text{and} \quad (6c)$$

$$\text{UNC}_i(m) = \bar{o}_i(m)[1 - \bar{o}_i(m)]. \quad (6d)$$

In these equations J is the number of probability bins, p^j is the probability associated with bin j , $n_i^j(m)$ is the number of times the threshold corresponding to category m at site i was predicted with a probability within bin j , $o_i^j(m)$ is the number of times threshold m was exceeded at site i given that it was predicted with a probability within bin j , and $\bar{o}_i(m)$ is the proportion of forecast times during which threshold m was exceeded at site i . The RES, REL, and UNC terms for the RPS are easily obtained by combining Eqs. (5) and (6), giving

$$\text{RPS}_{\text{REL}_i} = \frac{1}{K_i(M-1)} \sum_{m=1}^M \sum_{j=1}^J n_i^j(m) \left[\frac{o_i^j(m)}{n_i^j(m)} - p^j \right]^2, \quad (7a)$$

$$\text{RPS}_{\text{RES}_i} = \frac{1}{K_i(M-1)} \sum_{m=1}^M \sum_{j=1}^J n_i^j(m) \left[\frac{o_i^j(m)}{n_i^j(m)} - \bar{o}_i(m) \right]^2, \quad (7b)$$

and

$$\text{RPS}_{\text{UNC}_i} = \frac{1}{(M-1)} \sum_{m=1}^M \bar{o}_i(m)[1 - \bar{o}_i(m)], \quad (7c)$$

respectively.

The decomposition in Eqs. (6a)–(6d) is based on a system of forecast probability binning, and it is common practice to use equally sized probability bins, for which deciles are often chosen. The postprocessed multimodel probabilistic forecast is expressed in terms of 15 quantiles; therefore, interpolation will usually be necessary to obtain an exact forecast probability associated with each temperature threshold m . The value of the probabilities within each bin will vary from forecast to forecast and so [Stephenson et al. \(2008\)](#) devised a new (more complicated) decomposition of the BS that accounts for the within-bin variance. Generalizing this result to the RPS will add an additional level of complexity; therefore, in this study, the probability associated with each threshold m has been chosen as the minimum forecast quantile that has a value less than m . This choice significantly reduces the complexity because no interquantile interpolation is required and the within-bin variance reduces to zero. The impact of this simplification is likely to be minimal, whereas the added complexity of accounting for within-bin variance in Eqs. (7a)–(7c) would be complex.

c. Ranked probability skill score

The ranked probability skill score (RPSS) is positively orientated, ranges from $-\infty$ to 1 and is evaluated using

$$\text{RPSS}_i = \frac{\sum_{m=1}^M \sum_{j=1}^J [\bar{o}_i(m) - p^j] \{ [\bar{o}_i(m) + p^j] n_i^j(m) - 2o_i^j(m) \}}{K_i \sum_{m=1}^M \bar{o}_i(m) [\bar{o}_i(m) - 1]}, \quad (11)$$

obtained via the application of Eqs. (5), (6a), (7), and (8). Clearly, at sites where the frequency of occurrence of each category during the assessment period is almost identical to its long-term climatological equivalent, the choice of reference forecast is immaterial. However, the RPSS will indicate improved performance at sites where these frequencies differ because climatologically based frequencies for the reference forecast will not produce an optimal $\text{RPS}_{\text{ref}_i}$. Therefore, in the following sections $\text{RPS}_{\text{UNC}_i}$ has been chosen as the reference forecast for any figure that displays site-specific RPSS_i values. However, choosing site-specific climatology is appropriate from the standpoint of the operational environment in which it is necessary to issue a

forecast every day because if, for some technical reason, no forecast was available at the time of broadcast there must be a fallback position. Therefore, all-site RPSS performance figures use the site-specific climatologically based frequencies for the reference forecast.

The RPSS can be expressed as

$$\text{RPSS}_i = 1 - \frac{\text{RPS}_{\text{REL}_i}}{\text{RPS}_{\text{ref}_i}} + \frac{\text{RPS}_{\text{RES}_i}}{\text{RPS}_{\text{ref}_i}} - \frac{\text{RPS}_{\text{UNC}_i}}{\text{RPS}_{\text{ref}_i}}, \quad (9)$$

where $\text{RPS}_{\text{REL}_i}$, $\text{RPS}_{\text{RES}_i}$, and $\text{RPS}_{\text{UNC}_i}$ are given by Eq. (7) and since $\text{RPSS}_i \leq 1$, it follows that

$$\frac{\text{RPS}_{\text{RES}_i}}{\text{RPS}_{\text{ref}_i}} - \frac{\text{RPS}_{\text{REL}_i}}{\text{RPS}_{\text{ref}_i}} \leq \frac{\text{RPS}_{\text{UNC}_i}}{\text{RPS}_{\text{ref}_i}}, \quad (10)$$

where only the left-hand side is dependent on the forecast. This demonstrates how forecast skill is bounded from below by the (forecast independent) ratio of $\text{RPS}_{\text{UNC}_i}$ to $\text{RPS}_{\text{ref}_i}$, an expression which may be different for each site. Consequently, to ensure that the minimum possible value for RPSS_i is the same at every site it is necessary to set $\text{RPS}_{\text{UNC}_i} = \text{RPS}_{\text{ref}_i}$. In other words, when performing intersite comparisons, the sample uncertainty should be used as the reference forecast and this generates the following expression for the skill score,

forecast every day because if, for some technical reason, no forecast was available at the time of broadcast there must be a fallback position. Therefore, all-site RPSS performance figures use the site-specific climatologically based frequencies for the reference forecast.

d. Quantile score

The quantile score ([Gneiting and Raftery 2007](#); [Bentzen and Friederichs 2014](#)) is a proper scoring method that sums over a weighted absolute error between forecasts and observations. Unlike the other scores considered in this study, it does not use categorical thresholds, but instead the QS considers each quantile separately. This is ideal for verifying the

postprocessed probabilistic Met Office forecast because it is expressed in terms of 15 quantiles. Ben Bouallègue et al. (2015) define the QS as

$$QS = \frac{1}{N} \sum_{n=1}^N \rho_{\tau}(o_n - q_{\tau,n}), \quad (12)$$

where τ is the probability level of the quantile, $q_{\tau,n}$ is the n th forecast value for this quantile, ρ_{τ} is a conditional indicator function known as the check loss function (first defined in the context of regression by Koenker and Bassett (1978), o_n is the observation, and $\rho_{\tau}(o_n - q_{\tau,n}) = \tau(o_n - q_{\tau,n})$, if $o_n - q_{\tau,n} \geq 0$, and $(1 - \tau)(q_{\tau,n} - o_n)$ otherwise. Consequently, Eq. (12) becomes

$$QS = \frac{1 - \tau}{N} \sum_{n: o_n < q_{\tau,n}} (q_{\tau,n} - o_n) + \frac{\tau}{N} \sum_{n: o_n \geq q_{\tau,n}} (o_n - q_{\tau,n}), \quad (13)$$

and just like the RPS, it is possible to decompose the QS into reliability, resolution, and uncertainty components. Similarly, the quantile skill score (QSS) is evaluated using

$$QSS = 1 - \frac{QS}{QS_{\text{ref}}}, \quad (14)$$

where (just as for the RPSS) the reference forecast is usually chosen to be either the long-term site-specific quantile climatology or the sample quantile uncertainty term given by

$$E(\text{no forecast}) = \begin{cases} C(h + m + f + \text{cr}), & \text{if } E(\text{never act}) \geq E(\text{always act}), \\ L(h + m), & \text{if } E(\text{never act}) < E(\text{always act}), \end{cases} \quad \text{and} \quad (16c)$$

$$E(\text{perfect forecast}) = C(h + m), \quad (16d)$$

where h , f , m , and cr refer to the number of hits, false alarms, misses, and correct rejections, respectively. Dividing both numerator and denominator by L introduces the C/L ratio into Eqs. (16a)–(16d), thereby facilitating the evaluation of REV for different cost–loss ratios. A particular advantage of this assessment method is that it allows for easy comparison between deterministic and probabilistic forecasts.

4. Results

The present study examines the ability with which the postprocessed site-specific probabilistic and deterministic forecast products issued by the Met Office predict the

$$QS_{\text{UNC}} = \frac{1}{N} \sum_{n=1}^N \rho_{\tau}(o_n - \bar{o}_{\tau}). \quad (15)$$

Here $\bar{o}_{\tau}$ (the so-called unconditional sample quantile) is the τ th quantile of the observed distribution; therefore, QS_{UNC} is a characteristic of the observations and as such it is completely independent of the forecast.

e. Relative economic value

The cost–loss ratio (Thompson 1952; Murphy 1977) is a decision-making analysis tool that considers forecasts in terms of a user's decision to act against an adverse event. Fundamental to this approach is an expected cost C associated with taking action (irrespective of whether the event occurs or not) and an expected loss L associated with a missed event (given that no action is taken). The relative economic value is the ratio of the expected financial gain of taking the actual forecast service $E(\text{forecast})$ to the expected financial gain of taking a perfect forecast service $E(\text{perfect forecast})$. Consequently

$$\text{REV} = \frac{E(\text{forecast}) - E(\text{no forecast})}{E(\text{perfect forecast}) - E(\text{no forecast})}, \quad (16a)$$

where

$$E(\text{forecast}) = C(h + f) + Lm, \quad (16b)$$

maximum summer daytime temperature categories defined by the thresholds displayed in Fig. 2 during the meteorological summers between 2014 and 2016. Although these forecasts are generated by blending various probabilistic and deterministic models, the final probabilistic forecast solution is heavily weighted toward the output from the Met Office Global and Regional Ensemble Prediction System (MOGREPS). During the first 72 h, the PDF generated by the latest model run of the U.K. version of MOGREPS is given a weight of 70% in the probabilistic forecast. Whereas, at forecast ranges greater than 72 h, the global version of MOGREPS is given this weighting with a smaller contribution coming from the medium-range global ensemble forecast produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).

FIG. 3. RPS and its decomposed components as a function of the forecast range, calculated using all the sites and categories displayed in Fig. 2.

Figure 3 displays the RPS together with its decomposed components as a function of increasing forecast range between 6 and 120 h. An increase in the RPS with forecast range is as expected, and what is surprising is the near-linear nature of this increase from 0.04 at $T + 6$ to 0.075 at $T + 120$. This appears to be solely due to a decrease in RES, since both UNC and REL are almost invariant with respect to the forecast range. UNC is forecast independent so its homogeneity is expected; however, the invariance of REL with respect to forecast range is a positive result. A decrease in RES with increasing forecast range indicates that the probabilities within the forecast tend toward the local climatology, as often speculated on a subjective examination of site-specific meteograms (box-and-whisker plots generated by either ECMWF or the Met Office).

Figure 4 displays the all-site RPSS together with its decomposed REL and RES components. Using UNC as the reference forecast ensures that the RPSS is simply the difference between RES/UNC and REL/UNC. Figure 3 showed how UNC and RES were almost invariant to the forecast range; consequently, it is no surprise to observe that both the RPSS and RES/UNC decrease in a near-linear fashion with increasing forecast range while REL/UNC remains almost invariant. Between $T + 6$ and $T + 30$, the most highly weighted model used within the postprocessing at the time was the high-resolution MOGREPS-U.K. model; however, at $T + 36$ an alternative (lower resolution) model replaced it. This substitution manifests itself in Fig. 4 as an interruption of

the steady decrease in both RPSS and RES/UNC—the values of these statistics at $T + 36$ being almost identical to those at $T + 30$. Therefore, it would be interesting to investigate whether an earlier switch (at $T + 30$ or before) from the MOGREPS-U.K. model would result in improved RPSS performance. Similar behavior is observed between the time steps at $T + 42$ and $T + 48$, $T + 66$ and $T + 72$, and $T + 114$ and $T + 120$.

Figure 5 compares the site-specific values of the RPSS at $T + 24$ using the site-specific UNC reference (x axis) and the site-specific long-term climatology reference (y axis). Most points lie below the diagonal, indicating that the RPSS relative to UNC is smaller at the majority of sites. It is no surprise that forecast skill is lower when it is measured relative to sample uncertainty. The sample uncertainty represents forecasting each category according to its frequency of occurrence during the trial period and therefore it is only known after the event. The frequency of occurrence of each category during the trial period may differ (potentially quite significantly) from the long-term frequency of occurrence of these categories prior to the trial. Therefore forecasting the climatological frequency of occurrence will produce a poorer score than forecasting the true frequency during the trial period (i.e., the sample uncertainty). Consequently, the sites that lie farthest from the diagonal in Fig. 5 are those where the frequency of occurrence of each category during the trial period is most different from the 30-yr site-specific climatology. The two sites highlighted in red and blue are Exeter

FIG. 4. RPSS (with the sample uncertainty as the reference forecast) and its decomposed components as a function of forecast range, calculated using all the sites and categories displayed in Fig. 2.

Airport and Bridlington MRSC; these sites are the farthest and closest to the diagonal, respectively. The inset histogram displays the difference in the frequency of occurrence of each category (expressed as a proportion) during the long-term climatology and the equivalent frequencies during the trial period. At Bridlington the frequency of occurrence of each category was within $\pm 5\%$ of its equivalent during the climatological period. At Exeter Airport the largest difference was observed in the typical category which occurred over 20% less during the trial period.

Figure 6 shows the value of the RPSS (calculated using site-specific sample uncertainty as the reference forecast) at $T + 24$ (left) and $T + 120$ (right) for each site in Fig. 2. At $T + 24$ there is a cluster of high-performing (red) sites in a swath of locations stretching from Yeovil in the southwesterly county of Somerset, through inland counties of the south of England, Greater London, and on into East Anglia with another cluster on the coast between North Wales and the Lake District. The two highest-performing sites at $T + 24$ are both aerodromes; the value of the RPSS at the aforementioned Yeovil (51.017°N, 2.63°W) is 0.652 and at Farnborough (51.283°N, 0.767°W) it is 0.656, and both of these sites are located within the south of England. Poor performing sites at $T + 24$ also appear to be somewhat clustered with one group located in the southwest of England and another in the north of Scotland above 58°N. It is also interesting to observe that 16 out of the

20 green/turquoise sites are coastal locations. Perhaps the added complexities of forecasting in the coastal zone adversely affect performance at these locations; however, 7 of the 29 red sites are also located on the coast so coastal effects cannot be the only cause of this effect. The four worst performing sites at $T + 24$ are Bridlington (54.083°N, 0.167°W), Ronaldsway (54.083°N, 4.633°W), Inverbervie (56.850°N, 2.267°W), and St. Mary's Airport on the Isles of Scilly (49.917°N, 6.300°W). At $T + 120$ the picture is more mixed, the seven highest-performing turquoise/green sites are widely distributed, from Portland (50.517°N, 2.450°W) and St. Catherines Point (50.567°N, 1.3°W) on the south coast of England to Fair Isle (59.533°N, 1.633°W) off the north coast of Scotland. The southerly cluster of sites identified as high performing at $T + 24$ remain good (light blue) performers at $T + 120$; however, the northwest coastal locations have become relatively poor (dark blue) performers. By-and-large the cluster of poorly performing sites in the southwest of England remain poor performers at $T + 120$. The poorest performing sites at $T + 120$ are West Freugh (54.867°N, 4.933°W) in Dumfries and Galloway, Eskdalemuir (55.317°N, 3.217°W) also in Dumfries and Galloway, Ronaldsway (54.085°N, 4.632°W) on the Isle of Man, and Dishforth (54.133°N, 1.417°W) in North Yorkshire. Only Ronaldsway (the worst performing site according to this measure) is also in the bottom four sites at $T + 24$ and $T + 120$.

FIG. 5. Site-specific RPSS values at $T + 24$, comparing the effect of choosing site-specific uncertainty as opposed to site-specific climatology as the reference forecast for all sites in Fig. 2. The inset figure displays the difference in frequency of occurrence (expressed as a proportion) of each category during the sample period compared with the 30-year climatology. The histogram colors correspond to the highlighted red and blue sites (the farthest and closest sites to the diagonal, respectively).

The postprocessed model blend produces both a probabilistic and deterministic forecast. The calculations used to create the deterministic solution are entirely separate from those used to create its probabilistic equivalent. The deterministic forecast is a weighted mean of deterministic forecast values and ensemble means. Figures 3–6 have examined the probabilistic version; however, the deterministic equivalent currently populates the Met Office website. The categorized MSE evaluates the difference between the observed category and the category predicted by the deterministic forecast. In the form expressed by Eq. (3) the MSE is directly comparable to the RPS.

Figure 7 plots the categorized MSE (for deterministic forecasts) against the RPS (for corresponding probabilistic forecasts), with every point corresponding to a different site in Fig. 2. The black and red color coding denote a forecast range of 24 and 120 h, respectively. All points lie above the diagonal line of equal skill, since

both scores are negatively oriented this indicates that the probabilistic forecast tends to be more skillful than its deterministic equivalent at every site. At $T + 24$ there are three sites that display almost equal deterministic and probabilistic forecast skill. These sites are Exeter airport (near the coast in southwest England), Manston (near the coastal town of Ramsgate in southeast England) and Ronaldsway (on the South Coast of the Isle of Man), whereas Bridlington (a coastal site in northeast England) stands out as the farthest from the diagonal. Both of the measures displayed in Fig. 7 identify Bridlington as the site with the largest RPS and MSE values at both $T + 24$ and $T + 120$. The $T + 24$ score at this site appears to be less than that associated with the $T + 120$ forecast at a majority of U.K. sites. Out of the remaining 94 sites, the $T + 120$ MSE was lower than that corresponding to the $T + 24$ Bridlington deterministic forecast at 84 sites and the $T + 120$ RPS was lower than that corresponding to the $T + 24$

FIG. 6. Geographical distribution of RPSS (calculated using sample uncertainty for the reference forecast) for sites in Fig. 2 at (left) $T + 24$ and (right) $T + 120$.

Bridlington probabilistic forecast at 85 sites. However, the site that is farthest from the diagonal at $T + 120$ is Weybourne (a coastal site on the North Norfolk coast in East Anglia). The site that is closest to the diagonal at $T + 120$ is Mumbles Head (located on Swansea Bay on the south coast of Wales).

The MSE and RPS are highly correlated at both forecast ranges with correlation coefficients of $R^2 = 0.825$ at $T + 24$ and $R^2 = 0.914$ at $T + 120$, indicating a strong relationship between the sites deterministic and probabilistic forecast skill. It is also interesting to note that when the sites are ranked in terms of these statistics the Spearman ranked correlation coefficient obtained by comparing $T + 24$ with $T + 120$ is 0.784 for the RPS and 0.684 for the MSE. This indicates that a high-performing site at a short forecast range tends to remain a high performer at longer ranges. The cloud

of red points is noticeably farther from the diagonal equal-skill line, suggesting that the probabilistic forecast score decreases more slowly with increasing forecast range compared with that associated with the deterministic solution and therefore gives better guidance to users.

Figure 8 displays the all-site values of the categorized MSE and RPS at forecast ranges from $T + 6$ to $T + 120$. At every forecast range the probabilistic forecast outperforms its deterministic equivalent; however, the difference between these performance statistics noticeably increases with the forecast range, thereby confirming the supposition inferred from Fig. 7. The increase in MSE is almost monotonic (with the exception of $T + 90$) and as discussed in relation to Figs. 3 and 4 the generally monotonic increase in RPS with forecast range is also broken at $T + 36$, $T + 48$, $T + 108$, and $T + 120$.

FIG. 7. RPS and MSE calculated for all $T + 24$ (black) and $T + 120$ (red) forecasts issued at the sites displayed in Fig. 2.

This figure also appears to indicate that the performance of the probabilistic model at $T + 120$ is approximately the same as the performance of the deterministic model at $T + 54$.

The QSS (Bentzen and Friederichs 2014) is a useful way to evaluate the skill associated with each quantile of a probabilistic forecast; it is particularly useful in the present study because the forecast is expressed solely in terms of quantiles.

Figure 9 displays the all-site QSS for every forecast quantile at $T + 6$, $T + 24$, and $T + 120$, evaluated using site-specific climatology as the reference forecast. As expected the QSS decreases with increasing forecast range; however, at each range considered the quantiles below 20% and above 90% generate the lowest scores. This is an indication that outlying quantiles possess less skill at forecasting summer maximum daytime temperatures.

The QSS and RPSS both assess forecast skill against a reference; however, the (more established) RPSS evaluates skill at predicting multiple categories whereas the QSS requires no such categorization for its evaluation.

Figure 10 examines whether (despite this fundamental difference) these skill scores are correlated by comparing the values they generate for the $T + 24$ forecasts issued at each site. In this figure UNC has been chosen for the reference forecast and the all-quantile QSS has been evaluated by averaging the QSS over every forecast quantile.

The Spearman's rank correlation coefficient value for the points in Fig. 10 is 0.852. This clearly indicates that despite different calculation methodologies there exists a strong correlation between the performance ranking of summer daytime maximum temperature forecasts at each site according to the QSS and the RPSS. Both scores decompose into resolution, reliability, and uncertainty terms; however, its ability to calculate an overall score in addition to scores for individual quantiles places the QSS at a distinct advantage over the RPSS for the assessment of Met Office, probabilistic forecasts since this forecast product is quantile based.

Figure 11 shows the results of a REV assessment of the thresholds corresponding to the ten warmest summer days at every U.K. site as displayed in Fig. 2.

FIG. 8. All site values of the RPS and categorized MSE as a function of forecast range calculated using the sites displayed in Fig. 2.

This category has been chosen because expected losses associated with the lower threshold sets are likely to be minimal. Each subfigure compares the deterministic forecast (red) with each quantile of the probabilistic forecast (gray) and the probabilistic forecast envelope (blue) at forecast ranges of 6, 24, and 120 h. The forecast quantile corresponding to the 50th percentile is displayed in dark gray. Within each subfigure there is an obvious

discontinuity in the first derivative of every curve at C/L ratios of approximately 0.08. This first order discontinuity corresponds to the point at which the cost of always taking action is the same as that of never taking action [i.e., $E(\text{never act}) = E(\text{always act})$ in Eq. (13c)]. At this C/L ratio all light gray curves with higher REV than the dark gray curve correspond to quantiles with percentiles $> 50\%$ and all the remaining light gray curves describe quantiles with percentiles $< 50\%$. The probabilistic forecast envelope outperforms the deterministic solution at all C/L ratios below 0.46 at $T + 6$, 0.46 at $T + 24$, and 0.56 at $T + 120$. This improved performance is due to the higher quantiles of the forecast (those with probabilities $> 50\%$). The deterministic solution marginally outperforms all quantiles of the probabilistic forecast at C/L ratios between 0.46 and 0.66 at $T + 6$, 0.46 and 0.77 at $T + 24$, and 0.56 and 0.67 at $T + 120$; however, in each case the probabilistic forecast again outperforms the determinist solution at C/L ratios that exceed these ranges. Therefore, these figures appear to indicate that

- when the loss associated with an event is many times greater than the cost of the forecast, the probabilistic solution should be preferred (particularly at longer lead times) and should be acted on whenever an event is predicted by a high quantile;
- when the loss associated with an event is approximately between 1.5 and 3 times the cost of the forecast,

FIG. 9. All-site quantile skill score evaluated for each forecast quantile at $T + 6$, $T + 24$, and $T + 120$ for the sites in Fig. 2.

FIG. 10. Scatterplot comparing values for the quantile skill score and ranked probability skill score evaluated using $T + 24$ forecasts at each site in Fig. 2.

the deterministic solution is a suitable choice for action;

- when the loss associated with an event is comparable of the cost of the forecast, the probabilistic solution should again be preferred, acting whenever an event is predicted by a low quantile.

Figure 11 appears to indicate that even if the cost of generating the probabilistic model solution is higher than that associated with the deterministic model, this is justifiable for customers with low C/L ratios, particularly at longer lead times. However, in this instance the deterministic solution is the derived “most likely” solution from the PDF of the probabilistic forecast so there is no significant difference in production cost. The similarity between the deterministic solution and the median of the probabilistic PDF is confirmed in Fig. 11 by the apparent correlation between the REV (as a function of C/L ratio) for the dark gray quantile (50th percentile) and the deterministic solution within each subfigure.

5. Summary

This study examines the skill with which post-processed Met Office numerical models (both deterministic and probabilistic) predicted summer daytime maximum temperatures at U.K. observing sites between 2014 and 2016 for which at least 20 years of observed climatology is available between 1983 and 2012. An RPSS analysis indicates that the probabilistic forecast is

FIG. 11. Relative economic value plots comparing the post-processed deterministic and probabilistic forecasts (including each quantile) at forecast ranges of (top) $T + 6$, (middle) $T + 24$, and (bottom) $T + 120$ using the summer maximum daytime temperature thresholds corresponding to the highest temperature category, as displayed in Fig. 2.

better than the site-specific climatology at all sites, with a significant cluster of high-performing inland sites in the south of the United Kingdom (particularly around London) at a forecast range of 24 h. Sites near the coast, in Northern Ireland, the Isle of Man, and across northern Scotland appeared not to perform as well. By-and-large the sites identified as relatively poor performers at $T + 24$ were also poorly ranked at $T + 120$; however the measured performance at some sites which were identified as relatively good performers at $T + 24$ become relatively poor performers at $T + 120$.

The reliability component to the all-site RPSS (evaluated as being between 0.1 and 0.15) was found to be almost invariant to changes in the forecast range. Using a categorized version of the MSE enables a direct comparison between the probabilistic forecast and its deterministic equivalent. This reveals that the probabilistic model is the better performer at every site at all forecast ranges with the difference increasing with the range.

An all-site REV assessment (using site-specific thresholds corresponding to the very warm category) confirmed this finding, with the probabilistic model outperforming its deterministic equivalent for the vast majority of C/L ratios at $T + 24$ and this advantage increased significantly at $T + 120$. However, the deterministic model solution was the slightly better performer when the financial cost of the service was between approximately one and two thirds of the expected financial loss. The REV analysis also identifies that the performance of the deterministic solution was very similar to that of the 50th percentile of the probabilistic forecast.

A good degree of correlation between site-specific values of the (threshold independent) QSS and the (threshold dependent) RPSS confirms that the choice of thresholds used for the RPSS calculation does not appear to significantly affect the measured skill. In addition, the QSS gives useful information for users who act when a particular forecast percentile crosses a specific activity-related threshold. All-site values of the QSS indicate that the central quantiles (those between the 20th and 90th) are the better performers. At $T + 6$ and $T + 24$ the lowest quantiles (corresponding to the 2.5th and 5th percentiles) display the poorest skill; however, at $T + 120$ the maximum quantile (the 97.5th percentile) becomes the poorest performer.

Categorizing weather forecast values in terms of the climatology at the site for which they correspond is particularly useful to the forecast user who typically wants to know whether warmer/cooler conditions are being forecast at their location relative to the time of year. The analysis undertaken in the present study

examines the forecast skill associated with predicting these categories using either a deterministic or a probabilistic postprocessed numerical model solution at forecast ranges from 6 to 120 h. Perhaps the relatively poor performance of the outlying quantiles is an indication that the probabilistic forecast solution between 2014 and 2016 was underdispersed.

This analysis constitutes the first rigorous assessment of operational percentile forecasts generated by the Met Office, and it is highly applicable to many other weather types for which site-specific forecasts and long-term climatologies exist. Examining weather components during different synoptic conditions will help identify situations in which the model does not perform as well. Another interesting extension would be to attempt to use impact data to choose the thresholds from site-specific climatological CDFs, thereby enabling an investigation of the extent with which these categories influence health or asset management across the United Kingdom. A further extension could also include the use of another model (such as the ECMWF medium range ensemble) as the reference forecast. Additional benefit could also be obtained by assessing forecast performance at each stage of the postprocessing procedure.

Acknowledgments. The authors would particularly like to thank the National Climate Information Centre, the Public Weather Service, Clare Bysouth, and Rebecca Stretton for their assistance and the helpful suggestions of anonymous reviewers. The probabilistic model forecast data used in this study are unpublished; however, they are archived at the Met Office and are available for research use via a freedom of information request. For access, please contact enquiries@metoffice.gov.uk citing this paper.

REFERENCES

- Ben Bouallègue, Z., P. Pinson, and P. Friederichs, 2015: Quantile forecast discrimination ability and value. *Quart. J. Roy. Meteor. Soc.*, **141**, 3415–3424, <https://doi.org/10.1002/qj.2624>.
- Bentzien, S., and P. Friederichs, 2014: Decomposition and graphical portrayal of the quantile score. *Quart. J. Roy. Meteor. Soc.*, **140**, 1924–1934, <https://doi.org/10.1002/qj.2284>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, [https://doi.org/10.1175/1520-0450\(1969\)008<0985:ASSFPF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2).
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- Koenker, R., and G. Bassett, 1978: Regression quantiles. *Econometrica*, **46**, 33–50, <https://doi.org/10.2307/1913643>.

- Magnusson, L., T. Haiden, and D. Richardson, 2014: Verification of extreme weather events: Discrete predictands. ECMWF Tech. Memo. 731, 29 pp., <https://www.ecmwf.int/en/elibrary/10909-verification-extreme-weather-events-discrete-predictands>.
- Met Office, 2012: Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current). NCAS British Atmospheric Data Centre, accessed 1 September 2017, <http://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0>.
- , 2018: Met Office DataPoint. Met Office, accessed 1 September 2017, <https://www.metoffice.gov.uk/datapoint>.
- Moseley, S., 2011: From observations to forecasts—Part 12: Getting the most out of model data. *Weather*, **66**, 272–276, <https://doi.org/10.1002/wea.844>.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- , 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816, [https://doi.org/10.1175/1520-0493\(1977\)105<0803:TVOCCA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2).
- Neal, R. A., P. Boyle, N. Grahame, K. Mylne, and M. A. Sharpe, 2014: Ensemble based first guess support towards a risk-based severe weather warning service. *Meteor. Appl.*, **21**, 563–577, <https://doi.org/10.1002/met.1377>.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–668, <https://doi.org/10.1002/qj.49712656313>.
- Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201, [https://doi.org/10.1175/1520-0450\(1963\)002<0191:OSPF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2).
- Sharpe, M. A., C. E. Bysouth, and R. Stretton, 2018: How well do Met Office post-processed site-specific probabilistic forecasts predict relative-extreme events? *Meteor. Appl.*, **25**, 23–32, <https://doi.org/10.1002/met.1665>.
- Stephenson, D. B., C. A. Coelho, and I. T. Jolliffe, 2008: Two extra components in the Brier score decomposition. *Wea. Forecasting*, **23**, 752–757, <https://doi.org/10.1175/2007WAF2006116.1>.
- Thompson, J. C., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.*, **33**, 223–226, <https://doi.org/10.1175/1520-0477-33.6.223>.
- Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 2.2, <http://ams.confex.com/ams/pdfpapers/131645.pdf>.
- , G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.