New operational measure to assess extreme events using site-specific climatology

Michael Sharpe, Clare Bysouth, Philip Gill. Operational Verification Systems and Products,

Met Office UK

IVMW-O, 12th November 2020

January Temperatures

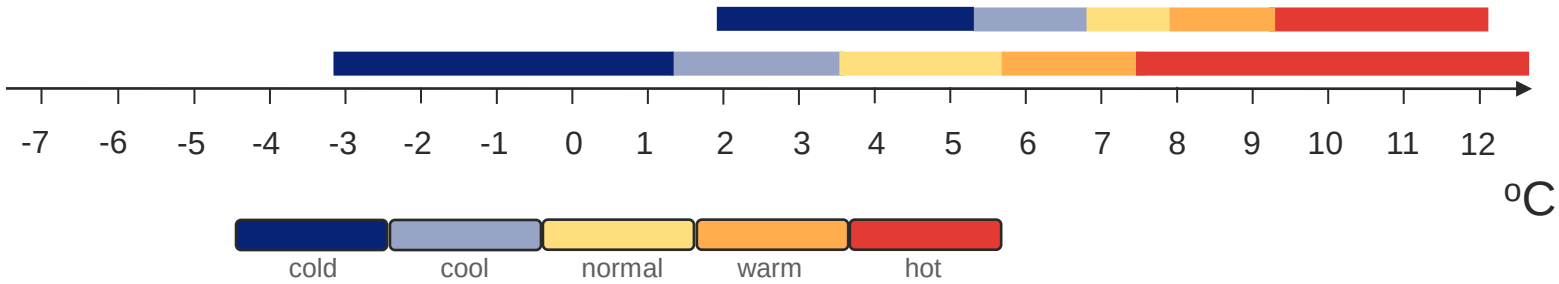
coldest 40%

warmest 40%

coldest 20%

warmest 20%

Scilly
Culdrose



Met Office

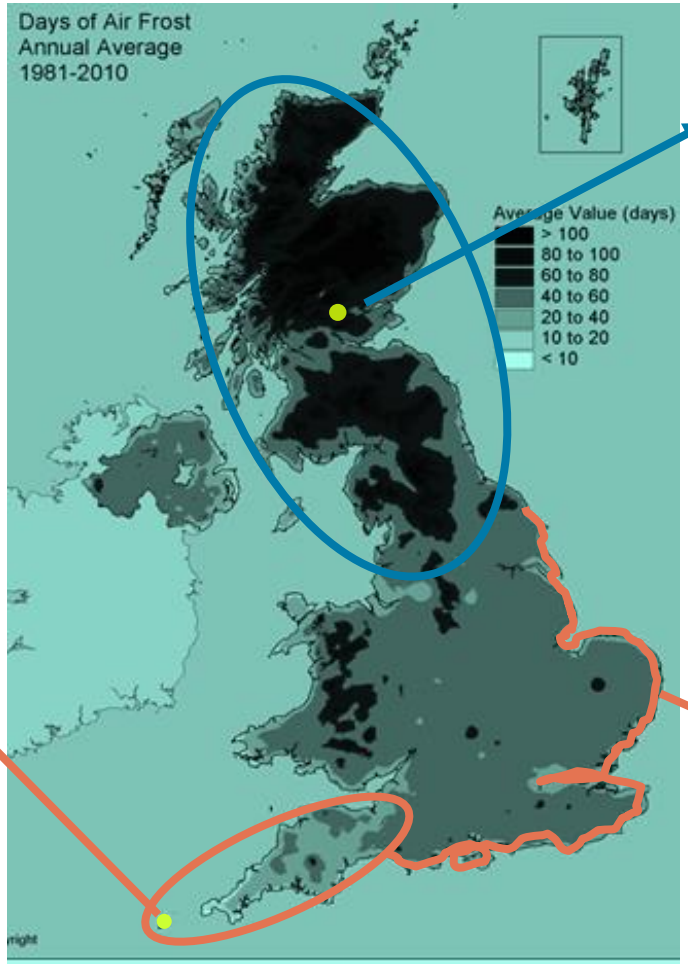
Relative-extremes:
define extreme events relative to site-specific climatology

Absolute-extremes:
occur more frequently in some places, and less frequently in others...

...whereas frosts on Scilly are big news...

- burst pipes
- dead plants
- ill pets

...are all likely



Frosts in Gleneagles are common...
...so they have little impact...

...instead it tends to be dominated by the skill in cold areas.

... performance is virtually unaffected by the skill in warmer areas...

Met Office

Why am I interested in all this…

Public Weather Service requested a

‘*Key Performance Indicator*’

for how well we forecast extreme events.

They want this measure to:

- be easy to explain
- be simple to communicate
- be robust (i.e. no sudden changes)
- directly compare deterministic & probabilistic forecast performance

Any measure of extremes should use local climatology to define an event

Advantages of using local climatology to define extreme events...

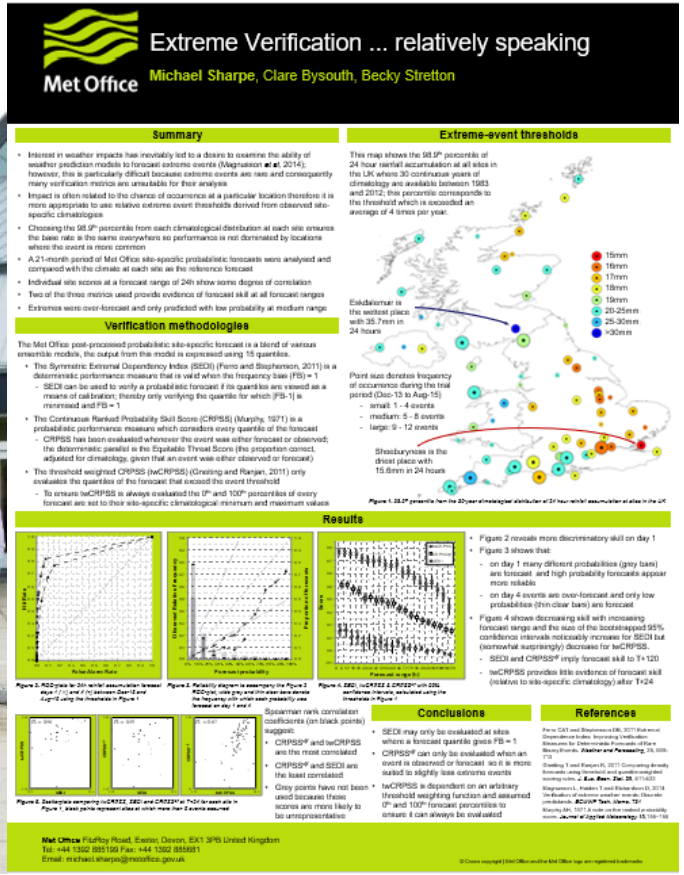
a) *Base rate is the same everywhere*
b) *Simple aggregation gives all-site performance*
c) *Similar impact everywhere*
d) *Event defined in terms of a return period*

Disadvantages...

a) *Can't verify where climatology is unavailable*
b) *Assimilation affects model climatology (use observed)*
c) *multiple thresholds → technically more challenging*
d) *Different to the norm!*

Met Office

7th IVMW – Berlin



Extreme Verification ... relatively speaking

Michael Sharpe, Clare Bysouth, Becky Stretton

Met Office

Summary

- Interest in weather impacts has inevitably led to a desire to examine the ability of weather prediction models to forecast extreme events (Magnusson **et al.** 2014); however, this is particularly difficult because extreme events are rare and consequently many verification metrics are unsuitable for their analysis
- Impact is often related to the chance of occurrence at a particular location therefore it is more appropriate to use relative extreme event thresholds derived from observed site-specific climatologies
- Choosing the 98.9th percentile from each climatological distribution at each site ensures the base rate is the same everywhere so performance is not dominated by locations where the event is more common
- A 21-month period of Met Office site-specific probabilistic forecasts were analysed and compared with the climate at each site as the reference forecast
- Individual site scores at a forecast range of 24h show some degree of correlation
- Two of the three metrics used provide evidence of forecast skill at all forecast ranges
- Extremes were over-forecast and only predicted with low probability at medium range

Extreme-event thresholds

This map shows the 98.9th percentile of 24 hour rainfall accumulation at all sites in the UK where 30 continuous years of climatology are available between 1983 and 2012; this percentile corresponds to the threshold which is exceeded an average of 4 times per year.

Eskdalemuir is the wettest place with 35.7mm in 24 hours

Shoeburyness is the driest place with 15.6mm in 24 hours

Point size denotes frequency of occurrence during the trial period (Dec-13 to Aug-15)
- small: 1 - 4 events
- medium: 5 - 8 events
- large: 9 - 12 events

15mm
16mm
17mm
18mm
19mm
20-25mm
25-30mm
>30mm

Figure 1. 98.9th percentile from the 30-year climatological distribution of 24 hour rainfall accumulation at sites in the UK

Verification methodologies

The Met Office post-processed probabilistic site-specific forecast is a blend of various ensemble models, the output from this model is expressed using 15 quantiles.

- The Symmetric Extremal Dependency Index (SEDI) (Ferro and Stephenson, 2011) is a deterministic performance measure that is valid when the frequency bias (FB) = 1
  - SEDI can be used to verify a probabilistic forecast if its quantiles are viewed as a means of calibration; thereby only verifying the quantile for which [FB-1] is minimised and FB = 1
- The Continuous Ranked Probability Skill Score (CRPSS) (Murphy, 1971) is a probabilistic performance measure which considers every quantile of the forecast
  - CRPSS has been evaluated whenever the event was either forecast or observed; the deterministic parallel is the Equitable Threat Score (the proportion correct, adjusted for climatology, given that an event was either observed or forecast)
- The threshold weighted CRPSS (twCRPSS) (Gneiting and Ranjan, 2011) only evaluates the quantiles of the forecast that exceed the event threshold
  - To ensure twCRPSS is always evaluated the 0th and 100th percentiles of every forecast are set to their site-specific climatological minimum and maximum values

Results

Figure 2. ROCplots for 24h rainfall accumulation forecast ranges T+[1,7] and T+[8,14] between Dec-13 and Aug-15 using the thresholds in Figure 1

Figure 3. Reliability diagram is exemplary the Figure 2 ROCplot; wide grey and thin clear bars denote the frequency with which each probability was forecast on day 1 and 4

Figure 4. SEDI, twCRPSS & CRPSS* with 95% confidence intervals, extracted using the thresholds in Figure 1

- Figure 2 reveals more discriminatory skill on day 1
- Figure 3 shows that:
  - on day 1 many different probabilities (grey bars) are forecast and high probability forecasts appear more reliable
  - on day 4 events are over-forecast and only low probabilities (thin clear bars) are forecast
- Figure 4 shows decreasing skill with increasing forecast range and the size of the bootstrapped 95% confidence intervals noticeably increases for SEDI but (somewhat surprisingly) decreases for twCRPSS.
  - SEDI and CRPSS* imply forecast skill to T+120
  - twCRPSS provides little evidence of forecast skill (relative to site-specific climatology) after T+24

Figure 5. Scatterplots comparing twCRPSS, SEDI and CRPSS* at T+24 for each site in Figure 1, each point represent sites at which more than 2 events occurred

Spearman rank correlation coefficients (on black points) suggest:

- CRPSS* and twCRPSS are the most correlated
- CRPSS* and SEDI are the least correlated
- Grey points have not been used because these scores are more likely to be unrepresentative

Conclusions

- SEDI may only be evaluated at sites where a forecast quantile gives FB = 1
- CRPSS* can only be evaluated when an event is observed or forecast so it is more suited to slightly less extreme events
- twCRPSS is dependent on an arbitrary threshold weighting function and assumed 0th and 100th forecast percentiles to ensure it can always be evaluated

References

Ferro CAT and Stephenson DB, 2011 Extremal Dependence Indices: Improving Verification Measures for Deterministic Forecasts of Rare Binary Events. *Weather and Forecasting*, 26, 699-713

Gneiting T and Ranjan R, 2011 Comparing density forecasts using threshold and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* 29, 411-422

Magnusson L, Haiden T and Richardson D, 2014 Verification of extreme weather events: Discrete predictands. *ECMWF Tech. Memo.* 727

Murphy AH, 1971 A note on the ranked probability score. *Journal of Applied Meteorology* 10, 155–156

Met Office FitzRoy Road, Exeter, Devon, EX1 3PB United Kingdom
Tel: +44 1392 885199 Fax: +44 1392 885681
Email: michael.sharpe@metoffice.gov.uk

Met Office

METEOROLOGICAL APPLICATIONS
Meteorol. Appl. 25: 29–32 (2018)
Published online 21 September 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1665

RMetS Royal Meteorological Society

How well do Met Office post-processed site-specific probabilistic forecasts predict relative-extreme events?

Michael A. Sharpe,* Clare E. Bysouth and Rebecca L. Stretton
Weather Science, Met Office, Exeter, UK

ABSTRACT: The Met Office routinely generates post-processed forecasts at sites throughout the United Kingdom; both deterministic and probabilistic products exist and deterministic data populate the publicly available website. In recent years, providers of weather information have focused upon the impact of events; impact is often related to the frequency of occurrence of an event at a site which is determined by its climatology. The ability with which a site-specific forecast predicts relative-extremes may be investigated by examining the skill with which these events (defined in terms of a percentile chosen from the climatology at each site) are predicted. The blended, deterministic, website forecast is less likely to forecast extreme events; therefore, the probabilistic forecast product (which does not currently appear on the Met Office website) was evaluated for its ability to predict heavy rainfall (RI^{1}_{24}), maximum summer day time temperature (T_{max}), minimum winter night time temperature (T_{min}) and strong winds (WS_{10}) over a 21 month period between December 2013 and August 2015. To this end, four methods of verification are considered: the Symmetric Extremal Dependency Index (SEDI), a threshold weighted version of the continuous ranked probability skill score (CRPSS) and a conditioned version of the CRPSS together with an analysis of the discrimination and reliability. Each method indicates forecast skill, with T_{max} and RI^{1}_{24} identified as the most and least skilful respectively and WS_{10} identified as the most reliable. Site-specific values of both versions of the CRPSS appear relatively well correlated and these scores also show correlation with SEDI for WS_{10}.

KEY WORDS extremes; forecasting; verification

Received 14 September 2016; Revised 20 January 2017; Accepted 4 February 2017

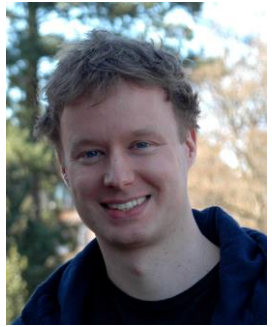
1. Introduction

In recent years the general public and meteorological community have become increasingly interested in extreme weather events because of their impact on society and/or infrastructure. This has inevitably led to a desire to examine the ability of weather prediction models to forecast extremes. However, such an examination is particularly difficult because, by their very nature, extreme events are rare.

Before proceeding any further it is necessary to define an extreme event carefully. At most UK locations a rainfall accumulation of 50 mm in 24 h would be considered an extreme event and in some instances this type of 'absolute' definition is appropriate (e.g. road, rail or air transport applications); however, it is often difficult to assess the performance accurately using this type of absolute extreme event threshold because it will be exceeded frequently at some locations but almost never exceeded at others. Consequently, the overall performance can be dominated by locations where the event occurs most frequently and usually has less impact. The impact of an event is often related to its chance of occurrence and, although absolute extreme events are important, when it comes to relating weather events to impacts it is often more appropriate to adopt a relative definition, where event thresholds are derived from the distribution formed by the climate at each location. This may be achieved by choosing the same percentile from the cumulative distribution function of the climatology (CDF^{c}) at each site; consequently, the event base rate is the same at every location, so the overall performance is not dominated by locations where the event is more common. Another advantage is the ability to express events in terms of their expected frequency of occurrence (e.g. the wettest day of the year); such definitions are arguably more meaningful than thresholds such as 50 mm in 24 h. A similar approach was independently taken at the European Centre for Medium-Range Weather Forecasts by Magnusson *et al.* (2014); these authors used the 98th percentile of the observed climatology to define event thresholds which were used to assess three different forecasts using the Symmetric Extremal Dependency Index (SEDI) (Ferro and Stephenson, 2011) and relative economic value (Richardson, 2000). An alternative method of analysis which Magnusson *et al.* (2014) suggest (but do not implement) is what they refer to as a 'modified version of the continuous ranked probability score (CRPS), where a function is applied to give more weight to extreme events'. It is likely that this is a reference to the threshold weighted continuous ranked probability score devised by Gneiting and Ranjan (2011), a methodology which examines the skill of any forecast quantile that predicts an extreme event. One method not considered by the present study is relative economic value because it has been considered in some detail by Magnusson *et al.* (2014).

In the present study, the relative-extreme event definition is adopted to assess the skill with which the probabilistic version of Met Office post-processed site-specific forecasts predict extreme 24 h rainfall accumulations (RI^{1}_{24}), maximum summer day time temperatures (T_{max}), minimum winter night time temperatures (T_{min}) and hourly wind speeds (WS_{10}). Site-specific climatology generation and analysis is discussed in Section 2;

* Correspondence: M. A. Sharpe, Met Office, Fitzroy Road, Exeter, EX1 3PB, UK. E-mail: michael.sharpe@metoffice.gov.uk



We've tried ...

1. *SEDI (deterministic method)*
   a) *Consider each (15) forecast quantile separately*
   b) *Choose one based on its frequency bias*
   c) *use it as the 'deterministic' forecast*

Contrived

2. *CRPS (probabilistic method)*
   a) *Compare the CDF of the forecast with the CDF of the observation (a Heaviside function)*
   b) *Restrict CRPS to only examine extreme events by conditioning or thresholding*
   c) *Integrate numerically over forecast percentiles*

$CRPS^{o|f}$ is improper

for maximum temperature O(t) = { 0 for all t < ob ; 1 for all t ≥ ob }

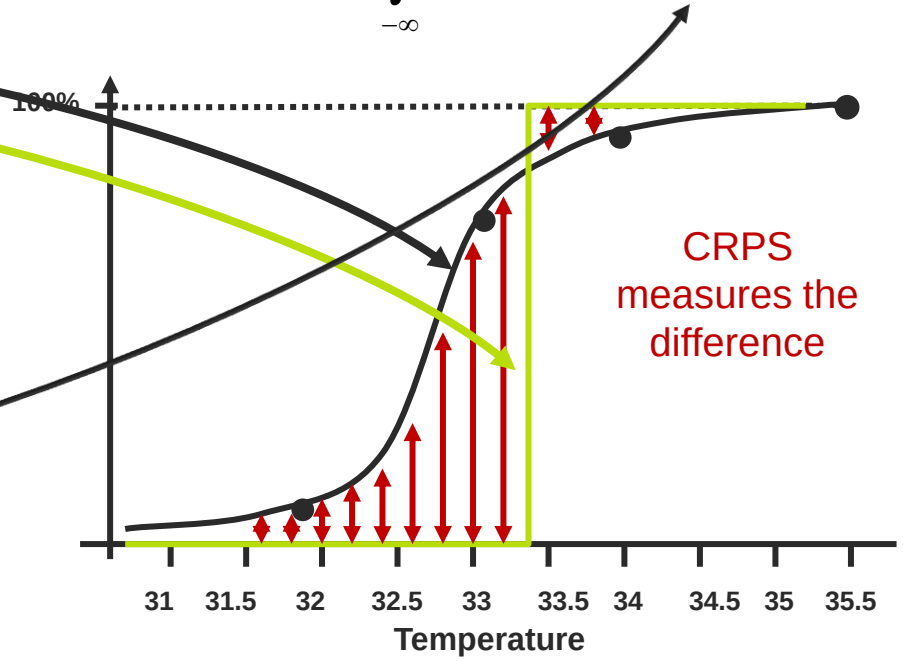
Continuous ranked probability score

Calculates the difference between

the forecast CDF

and the observed CDF

$$twCRPS = \int_{-\infty}^{\infty} \left(F(t) - O(t)\right)^2 dt$$



We are only interested in extremes...

... but the CRPS evaluates everything

Gneiting & Ranjan, 2011:

- Threshold weighted CRPS
- Choose a threshold weighting function that excludes non-extremes

We tried to frame the measure in a way that can be communicated.

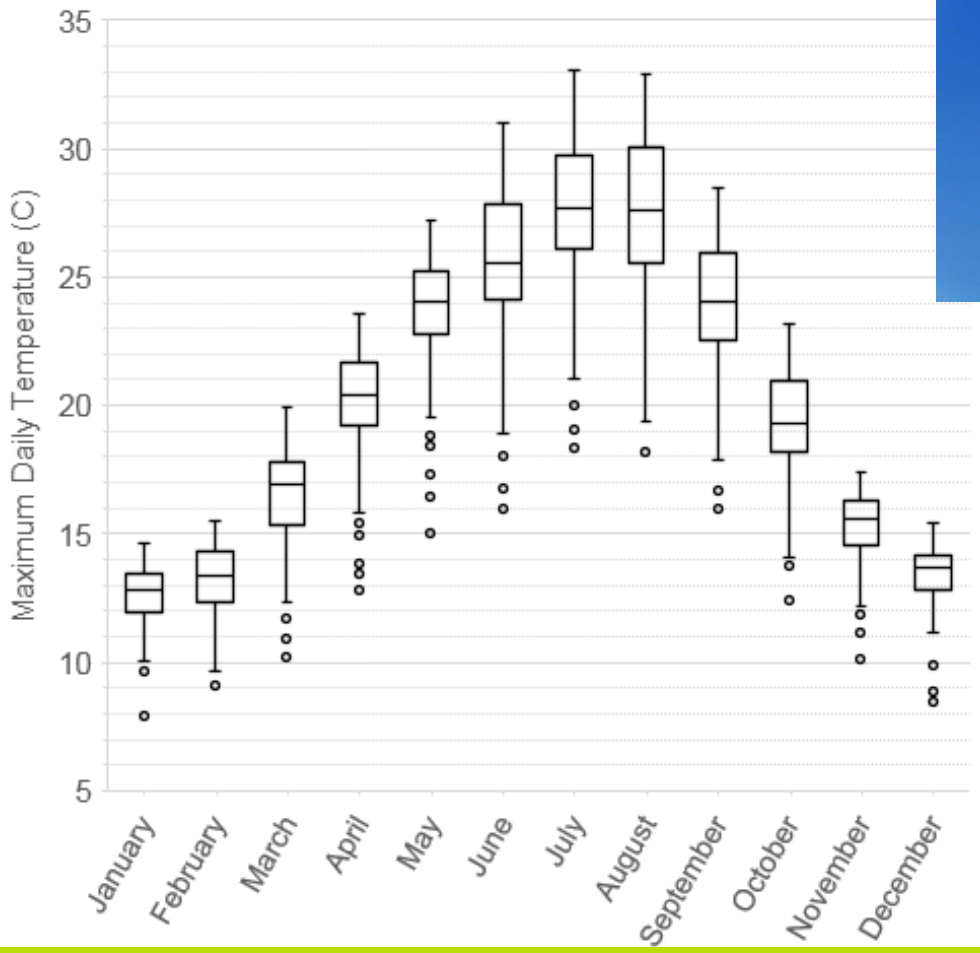
So we asked a common question…

Is today the coldest / warmest / windiest day in 3-years?

To answer that we've looked at:

- all observations at every site from 1987 to 2016
- the most extreme value in 3 randomly chosen years
- maximum temperature, minimum temperature & wind speed

The hottest day you should expect every 3 years at each UK site*



*based on 1987-2016

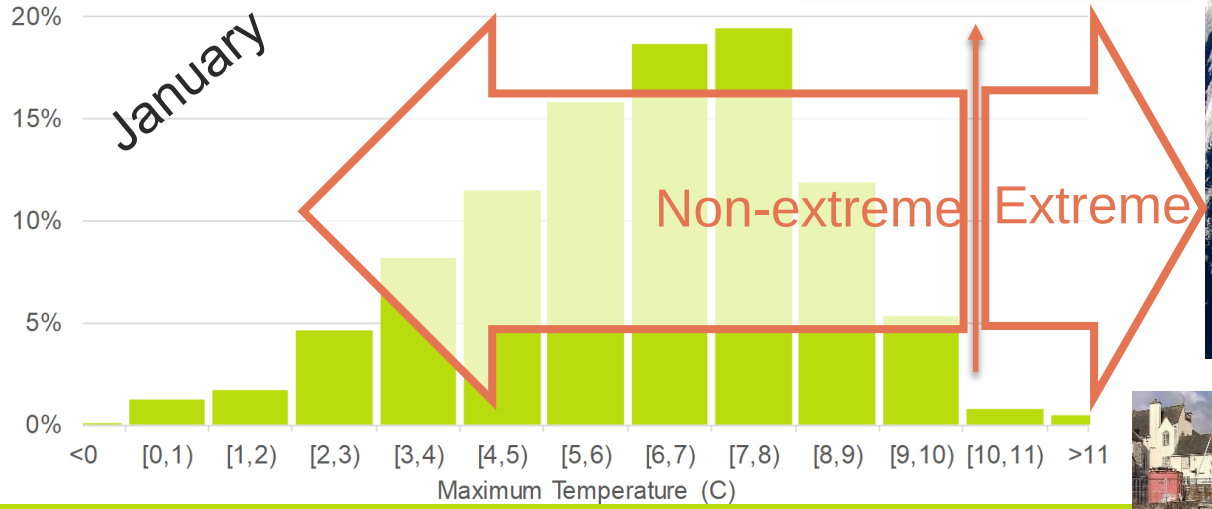
Met Office

New operational measure:

- most extreme 3 year event (98.89th percentile)
- 30-year site-specific monthly climatology

$$twCRPS = \int_{10.09}^{\infty} \left(F(t) - O(t)\right)^2 dt \quad w(t)dt$$

$$tw(t) = \begin{cases} 0 \; \forall t < 10.09\text{C} \\ 1 \; \forall t \geq 10.09\text{C} \end{cases}$$

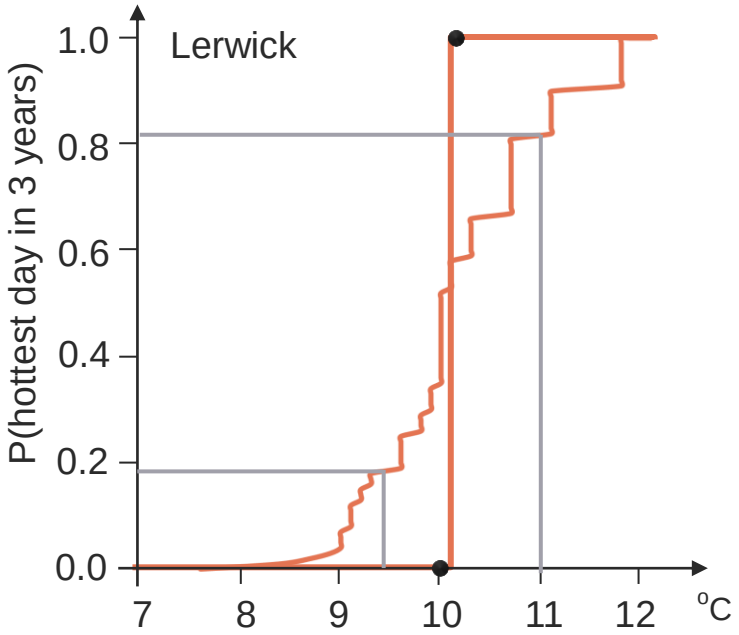


Met Office

New operational measure:

- most extreme 3 year event (98.89th percentile)
- 30-year site-specific monthly climatology
- Monthly sampling gives the probability that t is the hottest day in 3 years

$$twCRPS = \int_{10.09}^{\infty} \left(F(t) - O(t)\right)^2 w(t)\,dt$$



So, for example

- 9.4°C has a 19% chance and…
- 11°C has a 81% chance…

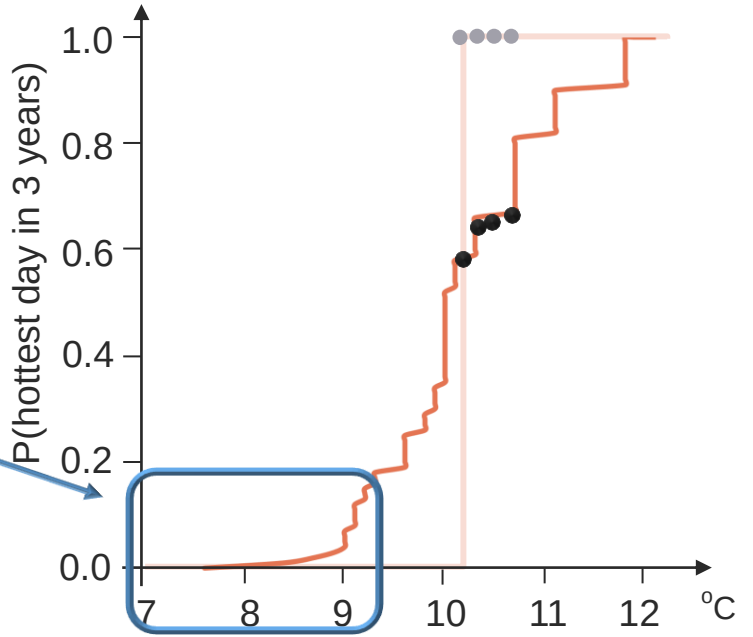
…of being the hottest January day in Lerwick in 3 years

Met Office

$$twCRPS = \int_{-\infty}^{\infty} \left(F(t) - O(t)\right)^2 tw(t)dt$$

New operational measure:

- most extreme 3 year event (98.89th percentile)
- 30-year site-specific monthly climatology
- Monthly sampling gives the probability that t is the hottest day in 3 years



Advantages:

- Has a real meaning to the public
- Can examine more extreme thresholds due to tail
- Score is less sensitive to small distributional changes
- So annual climate updates unlikely to affect robustness

Met Office

Probabilistic measure	Deterministic equivalent
$CRPS = \int_{-\infty}^{\infty} \left(F(t) - O(t)\right)^2 dt$	$MAE = \lvert f - o \rvert = \int_{\min(f,o)}^{\max(f,o)} 1 \, dt$
$twCRPS = \int_{-\infty}^{\infty} \left(F(t) - O(t)\right)^2 tw(t) dt$	$twMAE = \int_{\min(f,o)}^{\max(f,o)} tw(t) dt$
Score range: 0 to ∞	**where 0 is a perfect forecast**

Met Office

$$twMAE = \int_{\min(f,o)}^{\max(f,o)} tw(t)dt$$

f

o

twMAE≠ MAE

twMAE= MAE

P(hottest day in 3 years)

1.0

0.8

0.6

0.4

0.2

0.0

5 6 7 8 9 10 11 12 13 14 15 °C

Met Office

But what do twCRPS and twMAE values mean?

Skill Scores are more intuitive…

$$\text{Skill Score} = \frac{100 \times (\text{score}_{\text{climate}} - \text{score}_{\text{forecast}})}{\text{score}_{\text{forecast}}}$$

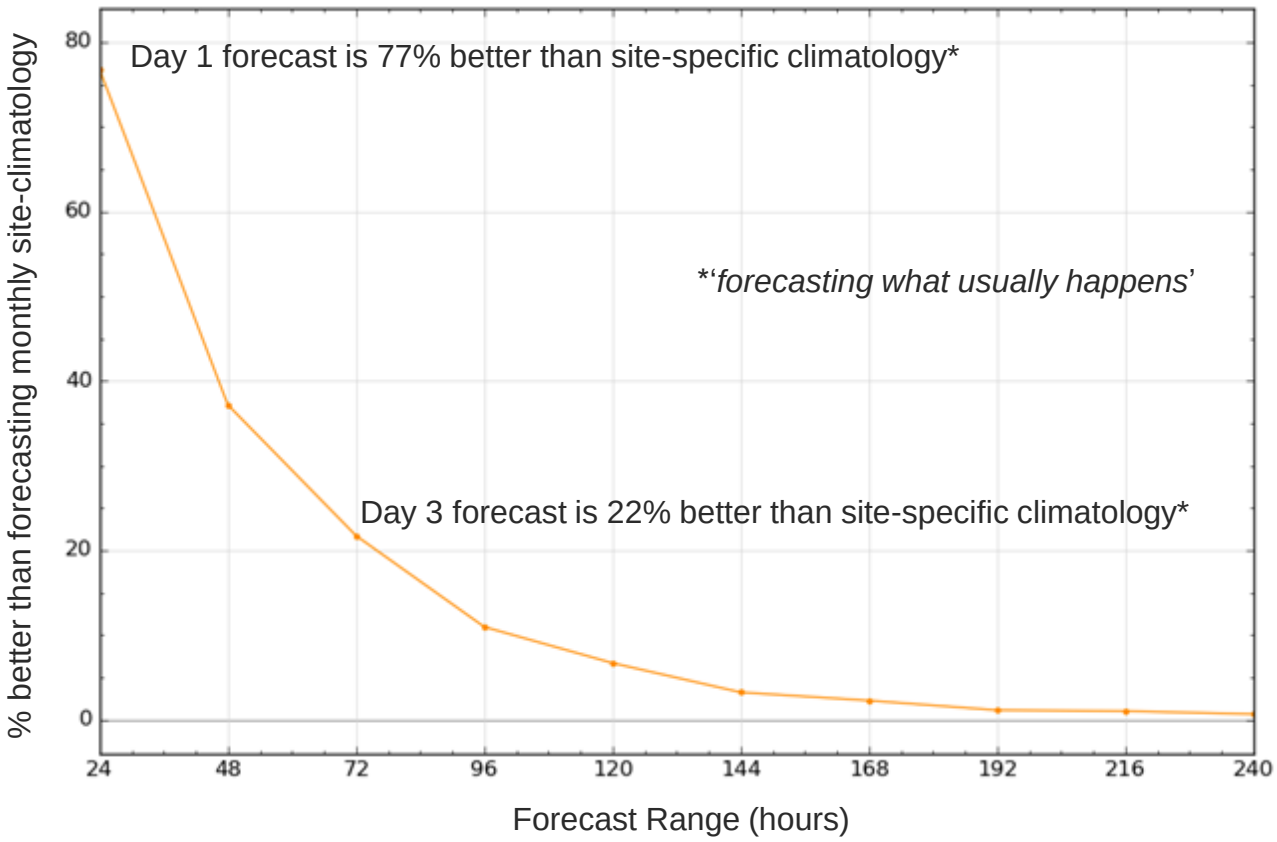
Reference:

- Site-specific climatology

Range: -∞ to 1

Transform to % better than reference

Maximum daily temperature



Forecast: deterministic

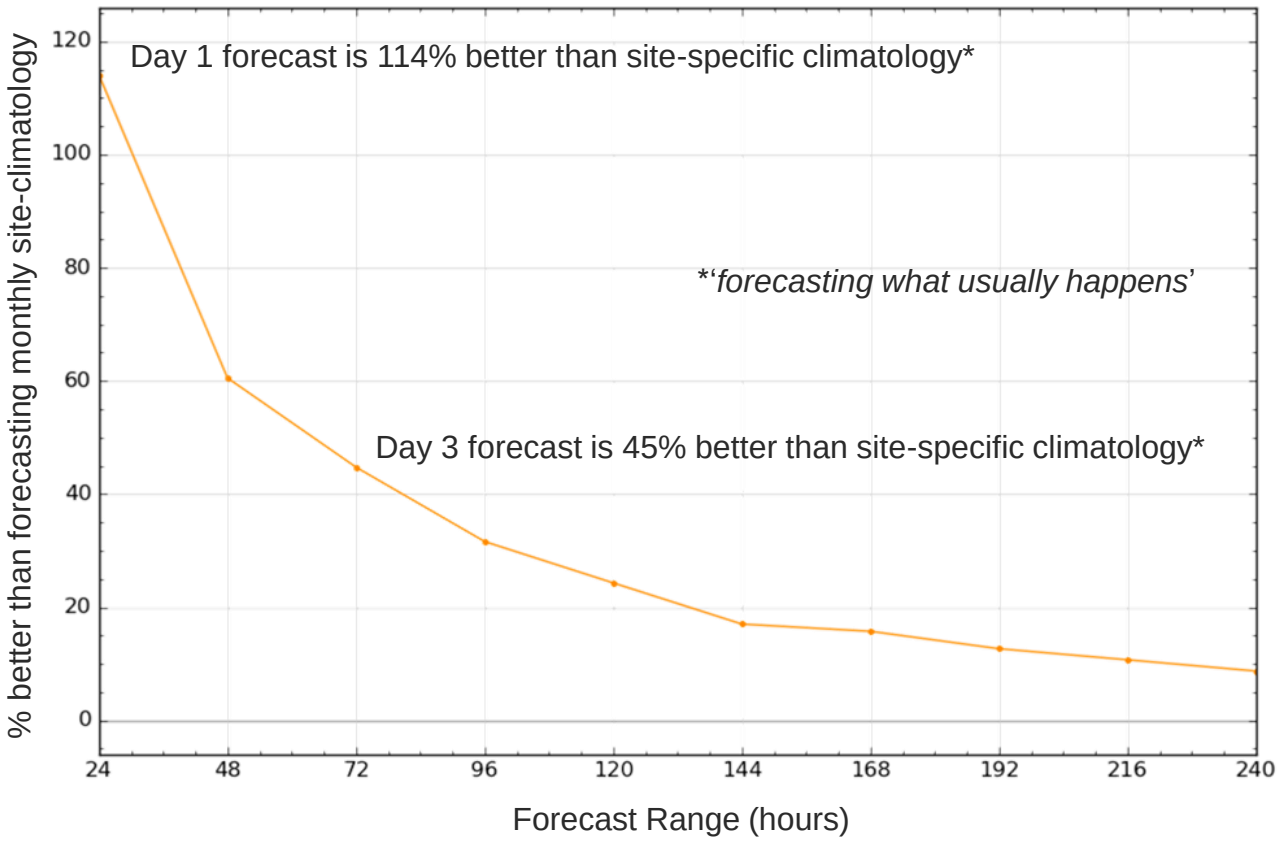
Reference: site climatology

Extreme event: hottest day expected every 3-years*?

Period: Jan-17 to Dec-19

*98.89th percentile

Maximum daily temperature



Forecast: Probabilistic

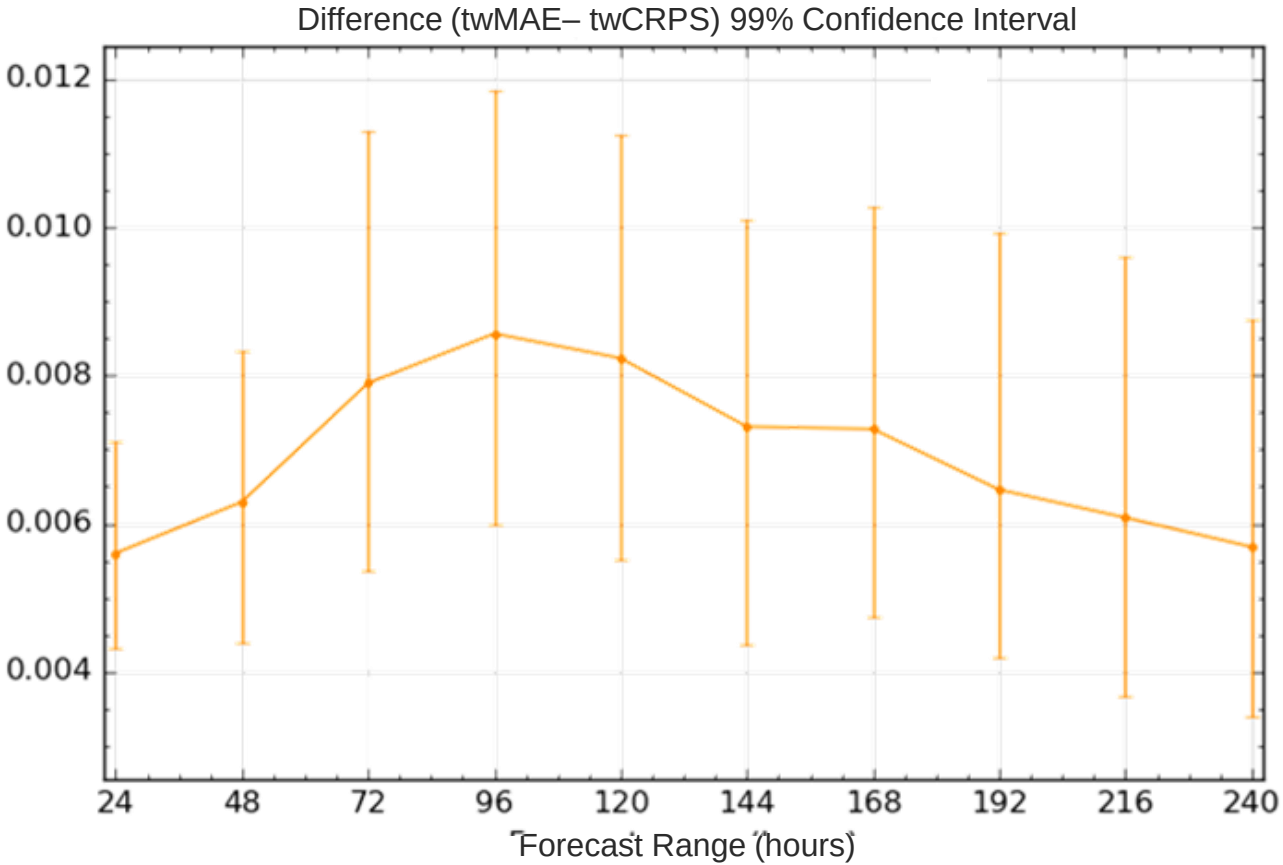
Reference: site climatology

Extreme event: hottest day expected every 3-years*?

Period: Jan-17 to Dec-19

*98.89th percentile

Maximum daily temperature



Strong evidence that probabilistic forecast better than deterministic forecast at predicting the hottest day expected every 3-years*?

*98.89th percentile

Requirements for Public Weather Service

‘*Key Performance Indicator*’

to measure how well we forecast extreme events:

- be easy to explain

  forecast error for the most extreme value that’s likely to occur every 3-years

- be simple to communicate

  how much better is the forecast at predicting extremes than forecasting what usually happens

- be robust (i.e. no sudden changes)

  by calculating an all-site rolling 3-year performance using a rolling 30-year climatological CDF

- directly compare deterministic & probabilistic forecast performance

  yes, because twMAE = twCRPS

Thank you

Michael Sharpe, Clare Bysouth, Philip Gill. Operational Verification Systems and Products,

Met Office UK

IVMW-O, 12th November 2020