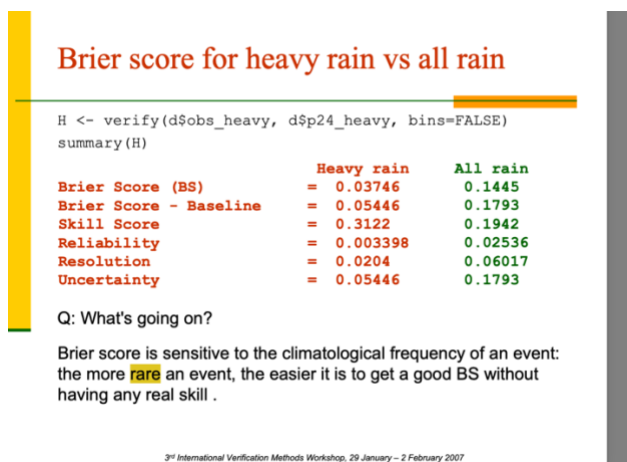


The authors followed several of the recommendations of the first comments provided for their manuscript. However, still, the manuscript is not in a form for publication as the results should be presented alongside accurate conclusions and in a well-written manner.

Some simple examples include the fact that the introduction does not provide a proper introduction on the topic (a matter raised in the previous review). The authors go on with unnecessary information leaving the reader confused and having to deduct alone what will be relevant for the study. For example, lines 58-66. Why are they relevant to the study? Why do we need to know what Katsafados 2014 found on seasonal time scales and related to blocking over Russia? The current study is on heatwave predictability for the subseasonal timescale over Africa and related to extreme temperature skill assessment. Later there is again an expansion on the method by Omondi et al., 2014 without any relevance. Also, even though there was a previous comment and links provided, still the whole manuscript is full of huge blobs of text without being separated into paragraphs. Finally, the title still contains the word “seasonal” forecasts, which are not included in the current study.

Some major concerns were also raised in the previous review, but I guess they were not convincing. The authors state as their main finding in the abstract, text, and conclusions that the model shows a very good Brier Score, it is therefore able to detect heatwaves for lead weeks 2 to 5. This is completely wrong. The Brier score is biased when the categories evaluated are unbalanced, which is the case here with 10% of your sample size being heatwaves. The rarer an event, the easier it is for the Brier Score to get low values, simply because the forecast model predicts well the majority category ----and not the heatwave category---. See here a screenshot of a presentation by a verification workshop by the ECMWF where they explicitly state that the rarer an event the better can the BS get... which is not surprising, as if you look at the equation it contains no sensitivity regarding the climatological frequency of the event.



Brier score for heavy rain vs all rain

```
H <- verify(d$obs_heavy, d$p24_heavy, bins=FALSE)
summary(H)
```

	Heavy rain	All rain
Brier Score (BS)	= 0.03746	0.1445
Brier Score - Baseline	= 0.05446	0.1793
Skill Score	= 0.3122	0.1942
Reliability	= 0.003398	0.02536
Resolution	= 0.0204	0.06017
Uncertainty	= 0.05446	0.1793

Q: What's going on?

Brier score is sensitive to the climatological frequency of an event: the more **rare** an event, the easier it is to get a good BS without having any real skill .

3rd International Verification Methods Workshop, 29 January – 2 February 2007

Here is the link to the presentation:

<https://www.ecmwf.int/sites/default/files/elibrary/2007/15489-verification-probability-forecasts.pdf>

I also run a very simple code and I get the same values as your study's Brier score. The model in this code just predicts no heatwaves. Does this mean that my model is able to detect heatwaves?!

```
import numpy as np
from sklearn.metrics import brier_score_loss

# Create an array that has the ground truth, so 10% is heatwaves which is denoted with 1
# Create an array with 10 zeros
truth_array_prob = np.zeros(10)
# Modify one element to be 1
truth_array_prob[0] = 1
print(truth_array_prob)

# Create an array that represents the values predicted by a model which no ensemble members predicting the HW
predicted_array_prob = np.zeros(10)
print(predicted_array_prob)

# Calculate Brier skill
bs = brier_score_loss(truth_array_prob, predicted_array_prob)
print('Brier score', bs)
print('Low Brier score means nothing for unbalanced categories')
```

[1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Brier score 0.1
Low Brier score means nothing for unbalanced categories

I will not comment again on the conclusions related to FAR, GSS, and hit rate.

Nevertheless, the addition of figures and discussion about the CRPS score being similar between lead week 2 and 5 adds nicely to the manuscript. However, I am not sure whether the strange CRPS scores have to do with the way it is calculated. It would be good to add which exact initialisations you consider for the calculation of CRPS, for example in January, and add this info in the Appendix. The fact that I am wondering about this, means that the readers will wonder as well. Therefore, being clear with the calculation will add validity to the results.

Regarding the calculation of the 90th percentile, it was great that the authors added an Appendix Figure. However, the procedure of calculations still needs to be clarified. The authors should keep in mind that the readers and reviewers are not supposed to guess the steps followed for such a calculation. Specifically:

Line 243 states: "For example, using ECMWF, the daily climatological 90th percentile is calculated over the study period separately for hindcasts run every Thursday of the month (see [Fig.S1] in supplement material)."

What do the authors mean by separately? Does the word separately refer to the separation between Monday and Thursday runs?

Then we go to the supplementary material and read the caption of Fig S1: “Fig.S1: Evolution of the 90th daily climatological percentile over AT region using T2m_min ECMWF hindcasts for: (i) the first and (ii) the second hindcast initialization dates from January to December.”

1. Do the authors mean by the first and second hindcast initialization dates of each month the first Thursday initialization and the second Thursday initialization of the month? Because one of the 2 first initializations is a Monday and the authors stated in the methods that they do not use Monday runs at all. Here it would help to clarify and to add an example, i.e., for January subpanel (i) refers to e.g. Thursday 03.January and subpanel (ii) refers to e.g. Thursday 10.January.

2. After the authors calculate for each initialization (that is for each model run with 46 days) the DAILY 90th percentile of that run considering all years, then how do they apply the 90th percentile threshold to assess heatwave occurrence? Do they apply the 90th percentile of each day of the 03.January initialization on each corresponding day of the run 03.January.2000, 03.January.2001 etc?

Or do they pool together all Thursday runs initialized in January then calculate a lead time depend climatology based on all daily thresholds and of course separately for lead week 1, lead week 2, etc.? The calculation of percentiles is not straight forward in S2S therefore it would be crucial to provide a schematic explaining clearly and in detail the steps 1 and 2 of the above.

Some typos and other comments:

Line 31: typo in “A Heat wave”

Line 42: I think that here you should replace “min, mean or max” with the full words, being “minimum, mean, maximum”

Line 4: it should be refer “Heat stress indices refers”

Line 101: “This work is carried out in West Africa” The authors should change “in” to “for”, otherwise I think that the current sentence means that they carried out the study while being physically in West Africa.

Line 120: Why is there “;” after stations? I do not think this is correct.

126: Replace this: (NOAA); (in the following, we will use "MERRA" to refer to MERRA-2) as our references for the evaluation

of the forecast models. With this: (NOAA). In the following, we will use "MERRA" to refer to MERRA-2 as our reference for the evaluation of the forecast models.

165-167: grammar is not correct

Line 87: I do not think that the word robust here is correct. The difference is that the Lavaysse method evaluates all heatwave characteristics whereas the other methods are focused on the evaluation of intensity. That does not make the Lavaysse method more robust. I propose to the authors to rephrase into: “This method offers a complete evaluation of heatwave characteristics including not only the evaluation of heatwave intensity, but also of heatwave onset and duration. ”

Line 160 is missing a verb.

171: I do not think that this is a valid reason for what the authors chose: “We only analyzed the hindcasts produced on Thursday. This is because we firstly want to carry out a multi-model analysis.” The authors analyse 2 models... is this a multi-model analysis?

211: Do the authors here mean by “developed approach” the sampling of the closest grid point to a station? If yes, then this is not a developed approach rather a method followed.

217: should be “occur” not occurred

219: are detected and not is detected

Lines: 307-309 do not need to be repeated. There are already in the introduction.

Legend in Figure 2 does not read nicely. The term maxima/minima temperature should be changed to maximum/minimum temperature. Also, this is not grammatically correct: ‘With pool’ refers to the pooling of two (or more) ... Maybe change into: The term “pooling” refers to ...

Mixing tences: line 251 “The predictability of heat waves is assessed ..” versus line 258: “The intensity of a heat wave was defined” That may not seem important, but line 258 could mean that the intensity was defined like that in the cited study and not in the current study.

The naming of section 2.4 as “Metrics” is not proper, as more than half of this section does not describe metrics but: 2.4.1 Estimation of temperatures at the city scale , 2.4.2 Heat wave detection etc..