Review of:

Subseasonal-to-seasonal forecasts of Heat waves in West African cities, NHESS (Langue et al., 2023)

The present study addresses the predictability of different heat wave types at the sub-seasonal time scale in West African cities over the period 2001-2020. The authors evaluate heatwave predictability in 2 state of the art sub-seasonal forecast models (46 forecast days), using two of the best reanalysis data sets available.

Overall, I find that the authors demonstrate their findings with a very nice sequence of Figures. However, sometimes important conclusions are derived about Tmax without demonstrating the Tmax plots, even though they can easily be added together with the respective Tmin in the Figures that exist in the main manuscript, or simply shown in the appendix.

Methodologically, I find very appropriate that the authors use several variables and several skill metrics in their analysis. Specifically, the authors proceeded with the evaluation of 3 variables of high interest and usability for heatwave definitions and heatwave research (Tmin, Tmax, Twet_bulb). The skill of the models to detect heat extremes is evaluated using the Brier score and the CRPS, while the predictability of heat waves in the forecast models is assessed by calculating categorical metrics such as the hit-rate, the Gilbert score, and the false alarm ratio (FAR). The results of this study (after Figure 4) are based on the percentile threshold selected for detecting extremes, as the authors create a 0-1 vector containing ones for days indicating extremes.

Unfortunately, I find that the percentile threshold selected for the detection of extremes is subjective and does not consider the model drift increase with lead time. The model drift is investigated in the following study and nicely shown in Figure 1: https://doi.org/10.1029/2019MS001751

The 90th percentile threshold the authors select is calculated over every calendar date and then the minimum percentile out of those is considered as a threshold for the full 46-day ECMWF forecast. The authors claim that they keep the selection of threshold constant, due to the relation of the study to a project investigating human impacts of climate extremes. However, this choice is still not justified, since this is a percentile threshold that the authors calculated. Normally human health impact studies do not use percentile thresholds, but actual temperature in degrees, e.g., 28 C. Moreover, this study is still a model evaluation study and, as the authors mention in the conclusions, the results will be used to investigate in detail the origins of the differences observed in the two forecast models over the different regions.

Normally, in sub-seasonal time scales, another 90th percentile threshold will be calculated for, e.g., August 1st for an initialization on July 31st and another 90th percentile threshold will be calculated for August 1st for an initialization on July 15th. However, a technique like that is not followed, so the threshold selection of the authors leads in many cases to better model skill in terms of CRPS and Brier score at lead week-5 instead of lead week-2. Moreover, the authors claim that the results of lead week 1 and lead week 2 are similar, which is a discrepancy to other studies. The reason for this discrepancy might again be the threshold selected. Moreover, if the authors want to support such argument, they could at least show some figures in the appendix.

This study presents important research in the field of sub-seasonal prediction for a region that lacks evaluation studies. However, the authors' conclusions cannot be supported by the current analysis. My recommendation is to reconsider the manuscript after major revisions focused on a correct estimation of thresholds. A more detailed review per section is provided bellow.

Abstract:

1. The conclusions given in lines 10-15 are not supported by the current results. Even though the brier score is lower than 0.1 in many cases, from the metrics shown Figure 10 we can deduce that the model shows no skill in detecting heat waves. A general comment to that is the distinction made by the authors between heatwave detection and heatwave prediction. Isn't for a forecast model the heatwave prediction a synonym to a heatwave "detection"? All metrics used are valuable for the evaluation of the model, as this study shows that an apparently low CRPS and Brier score is not a synonym to the model's ability to separate extreme heat from non-extreme heat.
2. Please replace the expression "the model shows skills" throughout the manuscript.

Introduction:

1. I was very confused by the fact that the title and the introduction have a lot of material about the seasonal time scale and that the authors claim that they will evaluate this time scale as well. However, the authors use sub-seasonal forecasts going maximum to 6 weeks lead time, so 1.5 months. The authors should better define time scales in the introduction. The sub-seasonal time scale covers 2 weeks to 2 months. The definition can be found in the S2S project here:

https://public.wmo.int/en/resources/bulletin/subseasonal-seasonal-prediction-project-bridging-gap-between-weather-and-climate

Following this comment, I recommend that the authors should not refer on their intro or anywhere else in the manuscript to the seasonal time scale.

2. I really liked the part where the evaluation is done at the city scale. We do not normally see that is sub-seasonal prediction studies and it adds novelty to this study.

Methodology:

Line 214: Do the authors mean that they calculate the daily climatological 90th percentile threshold? So, this threshold should vary depending on the time of the year, right? Is the 90-percentile calculated separately for forecast model, reanalysis data, and station data?

The authors return to the explanation of the 90-percentile definition in the line 233 and state "…daily exceedances of daily values of indicators to the climatological daily threshold " which kind of agrees with the statement above but then in lines 236-238 they state: "Therefore, the climatological daily threshold is chosen to be constant over the whole period; and it is defined as the minimum of the daily climatology thresholds over the study period. This approach allows us to properly assess the severity of a heat wave and its potential human impacts." At the end what is it exactly that the authors do?

Also, as previously mentioned, a lead time dependent percentile should be considered.

Line 226: I thought the authors mentioned before in their manuscript that they assess separately wet and dry heatwaves. Why then your binary vector contains data from extreme values of all temperature variables?

Major comment for methodology: The predictability of a model should not be assessed only by comparing to a random chance, as this would not make a strong argument into using this forecast model. The authors should assess predictability by comparing to a reference forecast. For example, the Brier skill score could be calculated separately for two common reference forecasts, being the climatological forecast and the persistence forecast.

Results:

Section 3.1: Calculating forecast climatologies for a sub-seasonal forecasting system that provides forecast over sub-seasonal lead times (maximum 6 weeks) does not mean that the evaluation done here is an evaluation of seasonal forecasts (also the title of the manuscript states that).

The authors do not evaluate the seasonal predictability of the forecast system, as this model cannot provide a seasonal prediction. The authors basically provide the climatological biases of the sub-seasonal forecasting systems over the different seasons.

Seasonal forecasting means that the forecasting system provides at lead-zero more than 8 weeks forecast, which is completely different with calculating climatologies. Other than that, I find this section very interesting, as we can conclude during which seasons the sub-seasonal forecasts have the largest biases over west Africa.

Section 3.2: As seen in the supplementary material of the previous section, the climatology is lead time dependent and therefore crucial for understanding whether the model predicts an extreme or not. I think that this would significantly change the results on predictability of this study and the authors would also see important differences between lead week 1 and lead week 2. For example, the outcome stated in lines 335-336 ("We have noticed that the skill of the models does not improve necessarily with decreasing lead time. ") is related to the choice of climatological distribution.

Here are some presentations from the ECMWF where they explain the calculation of the lead time dependent climatology:

Example for seasonal forecasts:

https://confluence.ecmwf.int/pages/viewpage.action?pageId=174864039

Model climate calculation in page 33:
https://resources.eumetrain.org/data/7/711/high_latitudes_ew_2023_s1c.pdf

Demonstration on how lead time affects EFI verification in page 34. Have a look on known issues on page 35, that is also why in the current study the Thursday initializations should not have been removed as it drastically affects sample size:
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiT2_GgwKeBAxWUh_0HHZtzD-oQFnoECA4QAQ&url=https%3A%2F%2Fconfluence.ecmwf.int%2Fdownload%2Fattachments%2F70951731%2FForecasting_Extremes_Oct2017.pdf%3Fapi%3Dv2&usg=AOvVaw3YVlL5ZgSZsAzDf-LKfIzI&opi=89978449

Line 338: Could the authors explain why the forecast models show better predictive skill in the AT region?

Section 3.3.1:

Lines 354-355: The authors state “This approach based on a relative threshold (see section 2.4.3) will contribute to partially correct the biases previously found in the models.” How does a variable daily threshold correct bias? A daily threshold simply accounts for differences between seasons.

408-409: The authors state “The forecast models show skills above the reference both for short- and long- term forecasts (Week2, Week5)”

I am wondering which figure shows that there is skill above the reference. Moreover, when the authors say skill, they should specify the metric to which they refer to. Taking into account the pairs of FAR and Hit_rate, Figure 10 shows no skill in detecting/predicting heat waves for none of the seasons/ models/ lead weeks.

410: Do the authors here mean instead of “more skills”, “higher skill scores”?

Figure 10: Why is only the evaluation for T_min shown? The authors could add Tmax evaluation in Figures 10 and 11.

The authors state “We can infer from this result that nighttime heat waves are more predictable than daytime heat waves.” Is there a Tmax figure I missed from which the authors infer this? Moreover, if that is indeed shown on a figure, is the reason for the higher predictability the actual 90th percentile metric used which could be more variable in Tmax and that is why it appears less predictable.

Also, why do the authors not provide the evolution of the mean climatological biases between the forecast models and reanalyses for Tmax in the Appendix as provided for Tmin and Tw in S2 and S3?

The authors should avoid the expression “the models show more skills”. It should be rephrased to “the models show higher Brier skill in …

415: What is an inter-day variability? Is it the std calculated over all your samples per region?

416-417 Can the authors explain more this this sentence: The low inter-day variability of T2m in the AT region indicates a more stable signal which will lead to favorable conditions for heat wave detection in the models based on a statistical perspective.

Wouldn’t a stable signal lead to lower probability of extremes, so an even harder prediction of heatwaves?

Line 426: By “significantly decreasing” do the authors mean that they have calculated a level of significance? Do maybe the authors mean that the values are strongly decreasing? Or maybe they mean that the values are strongly decreased. In any case, GSS values that go from 0.2 to 0.1 are not strongly decreased, are low overall.

432: The use of the word “ability” in this sentence is misleading. Is it an ability to predict events that did not occur?

The authors provide the T_min variable in Figure 10 and in Figure S15. What is the difference in the 2 plots? Why there are no Tw and Tmax plots provided in the appendix? In Figure 10 which percentile threshold of ensemble members is used? The authors should add this in the figure caption.

445: How did the authors get to this conclusion: “The forecast models show skills at weekly time scale compared to the baseline climatology.” Calculating the Brier skill score using the climatological forecast as reference could support (or not) this statement.

Figure 10 shows that, according to the evaluation done here, the model is over-forecasting heat waves. This can be even more explicitly shown if the authors plot hit rate against FAR to create the Roc curve. In this curve, the pair of 0.55 hit rate – 0.75 FAR will be below the diagonal. Being bellow the diagonal indicates no skill to discriminate between events and no-events, with the diagonal indicating random value/no-skill. This plot would disagree with the conclusion drawn by GSS.

http://www0.cs.ucl.ac.uk/staff/w.langdon/roc/

In the following ROC curve, I indicate with a yellow dot where your 0.55 hit rate – 0.75 FAR pair would be.

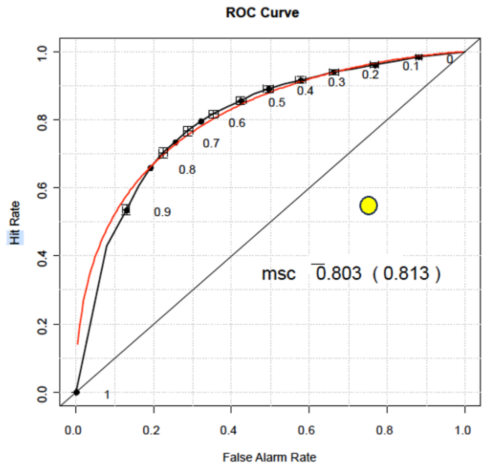

FIGURE 11: The markers have very small size, and so are hard to see. Here the authors could easily plot Tmax values as well, which is a very valuable parameter for heatwaves and its important to see its model biases here as well. I would suggest that the authors create one row for each variable (Tmin, Tmax, Tw) and maybe show both models in every sub-panel.

Minor comments:

Line 24 needs some references.

Line 37: That is a very big list of references, just to reference daily raw temperature as a variable relative for heatwaves. In the case of simple definitions, it would be helpful to keep the refence lists shorter and target to show the refences that are the most relevant.

Line 45-46 “This is usually done using seasonal weather forecast models.” This sentence needs references. Also, since the study is also for sub-seasonal time scales it would be great if the authors motivate the sub-seasonal time scale and add references with studies connecting early warnings with the sub-seasonal time scale.
One reference I have in mind is:
Osman et al., 2023: Sub-seasonal to decadal predictions in support of climate services
DOI: 10.1016/j.cliser.2023.100397

Line 51: There is a published study on the predictability of extreme events across the globe:
Advances in the Sub-seasonal Prediction of Extreme Events: Relevant Case Studies across the Globe
Domeisen et al., 2022, DOI: https://doi.org/10.1175/BAMS-D-20-0221.1
Also, overall, for the onset/intensity/duration of European heatwaves at sub-seasonal time scales, where the calculation of lead-time dependent climatology is also explained:
Subseasonal predictability of onset, duration, and intensity of European heat extremes
Pyrina et al., 2022, https://doi.org/10.1002/qj.4394

Line 66: "Expert Team on Climate Change Detection and Indices (ETCCDI) database," also needs a reference.

Line 67: What is apparent temperature?

65-71: I find this paragraph very long and confusing especially because it is not going to be related to the method that will be used in the current paper. I would either remove it or keep a few sentences about it. Also, the paragraph goes on about what other studies did but there is no connection to what will be done here. The method of Lavaysse et al., 2019 is mentioned later, but it would be nice to also mention a few sentences about their method and why is good and you have followed it here.

Also, the Lavaysse et al., 2019 study could be mentioned before, when talking about the heatwave studies, so that the reader connects the study already with heatwave evaluation.

Line144: Please provide a citation from a publication or a book. Citations of webpages are not proper for research papers.

Lines 167-169: I do not understand the meaning of this sentence, was there a problem with the ECMWF output? Why would you choose to not evaluate all available initializations and reduce so much your sample size? A more accurate approach would be to evaluate the hindcasts for another operational model version, such as the model version of 2021, 2022, or even 2019.

Line 183: Here it is stated that the authors are "interested in the predictability of heat waves in a global perspective", which is confusing as it may be understood that the evaluation will be done for the whole globe. Also, what is stated in the sentence is not a good argument on why using different weekly initializations are comparable. The authors could just say that using ECMWF initializations on Thursdays leads to 4 initializations per month, making the sample size comparable to the UKMO model.

Figure 3: The authors cannot name a) for T2m_min and a) as well for winter. You could do i) t2m_min, ii) tm_max.

Figures 3,4: It should be mentioned in the figure that the authors consider all available lead times for this figure.

Figures 5,6:

1. For each of the variables investigated (Tmin, Tmax, Tw) when are there the stronger climatological biases? How much do they change if we consider a lead time dependent percentile threshold?

2. In many cases the skill increases with lead time, which is not common at all especially comparing forecast weeks 2 and 5. Is the change in skill driven by some particularly well

predicted period of extremes at lead week 5? Or maybe it comes by the fact that the authors define a common 90 percentile threshold for all lead times?

According to the error metrics of figures 5c and 6c, a conclusion would be that the users should trust the summer prediction over the CO region at lead week 5 more than at lead week 2!? This result is even more striking when looking at the winter season and Tw in figure S6.

Figure 7,8: The authors should explain what the grey values represent in the plots.

Line 322: Actually, there are systematic decreases and increases in biases with lead time in several subplots. For example, the bias is especially pronounced for ERA5 in the region CO (S2-a,f,h,e and in S3-everywhere). Why do some of the biases decrease with lead time? Can the authors explain some of these results or at least mention them?

Figures 7, 8 Why would the authors indicate a colorbar without units? Here it should be Bias (%)

Figure 9 Again here why having a colorbar without units? Especially with your duration definition that is very important. Please change to: Bias (days per year)

--The long text in each section is hard to read and makes it hard to return to a specific point when needed. Please separate the long text of each section in paragraphs. See here some tips: https://www.uvic.ca/learningandteaching/assets/docs/instructors/for-review/Information%20for%20Students/science%20paragraphs.DVG.FINAL.pdf

--Some figures have: “FAR ratio”, but ratio is inside the word FAR anyway. Change to: “FAR”.

479: How do the authors now that the models have issues with the “spatial evolution of heat waves”? Do they mean spatial variability?

Conclusion section:

This section should be rewritten after the revision of this study.

Typos:

Line 99: The references need brackets: Moron et al. (2016); Ngoungue Langue et al. (2023)

Line 182: Change “init dates” to “initialization dates”

Line 251: Change in the “The skill of the probabilistic models … are assessed...” to “…is assessed…”

Line 236: remove ;