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.

We improved the quality of the document by taking into account the reviewer's comments. Below are some changes to the introduction and other parts of the document.

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.

We changed the introduction according to the reviewer's comments. The new introduction is the following :

" The impact of heat waves on different sectors, in particular the economy and health, makes them one of the most dangerous climate hazards globally \citep{perkins_review_2015}. Heat waves pose a significant threat to human health, as they cause discomfort and stress to body temperature regulation. \textcolor{red}{In some cases, heat waves can lead to various cardiovascular and respiratory diseases, increasing the risk of morbidity and mortality [e.g. \citet{anderson2009weather,gasparrini2011impact,kovats2008heat,huynen2001impact}]}. The combination of heat and humid atmospheric conditions} can exacerbate heat stress and lead to deaths, particularly among vulnerable populations such as children and the elderly \citep{russo2017humid}. The

damages of heat waves on human health are amplified in urban regions due to urban heat islands. Over the last decades, West Africa cities were affected by several heat extreme events. In May 2013, the Senegalese town of Matam experienced a severe heatwave, with temperatures reaching 50°C in the shade. This event, which persisted day and night, caused the death of 18 elderly in 10 days. At the same period, Mauritania was affected by a devastating heatwave with maximum temperatures exceeding 46°C, causing the death of more than 25 elderly people and children. Recently in April 2024, most of African countries experienced extreme heat conditions. In Mali, for example, temperatures exceeded 48°C, causing the death of more than 100 people.

Climatic projections show an increase in the frequency, intensity and duration of extreme temperatures over the next century and beyond [e.g. \citet{kharin2007changes,fischer2010consistent,perkins_increasing_2012}]. Under these warming conditions, the frequency of extreme events such as heat waves will increase. In the latest Intergovernmental Panel on Climate Change report (IPCC report 2023), the authors show that equatorial regions will be more affected by climate change than mid- and high- latitudes. This result from the IPCC is a warning bell, as the predictability of heatwaves remains poorly documented in some equatorial regions, such as sub-Saharan Africa.

The impact of heat waves on human activities and health increases the need for skillful and reliable climate forecasts on subseasonal to seasonal time scales in order to anticipate risks and develop appropriate responses \citep{lowe2016evaluation}. Therefore, early warning systems are of crucial importance to provide information on the occurrence of such events. In general, early warning systems integrated shorter and medium-range forecasts of potential weather hazards. This type of forecast window refers to subseasonal time scale from 2 up to 6 weeks. The subseasonal range is highly relevant for actions aimed at mitigating the consequences of extreme heat [e.g.\citet{white2017potential,moron2018sub,tompkins2019predicting,osman2023sub}]. Subseasonal forecasts are used to monitor the evolution of specific weather patterns that have been identified in advance with seasonal forecasts.

Subseasonal and seasonal forecasts (S2S) are of great importance for humanitarian services in order to build up a "Ready-Set-Go" early warning concept that allows early actions to be taken before a potential disaster [e.g., \citet{bazo2019pilot,lala2022evaluating,domeisen2022advances}].

Heat waves are often associated with extreme heat and wet/dry temperatures. A heat wave is defined as a period of unusual hot temperature over a region persisting at least three consecutive days during the warm period of the year based on local (station-based) climatological conditions, with thermal conditions recorded above given thresholds [e.g. \citet{perkins_measurement_2013,deque2017multi,barbier_detection_2018,batte_forecasting_2018,ngoungue2023heat}]. Many factors can affect the definition of a heat wave, including the end-user sectors (human health, infrastructures, transport, agriculture) and also the climatic conditions of the regions \citep{perkins_measurement_2013}. Heat waves can be defined from daily meteorological variables such as daily raw temperature [e.g. \citet{batte_forecasting_2018,lavaysse_towards_2018,engdaw2022changes,ngoungue2023heat}], \textcolor{red}{minimum, mean, maximum} daily wet bulb temperature [e.g. \citet{yu_changes_2021,ngoungue2023heat}] or heat stress indices [e.g. \cite{robinson_definition_2001,fischer2010consistent,perkins_increasing_2012,guigma_characteristics_2020,ngoungue2023heat}] using relative or absolute thresholds. It refers to indices resulting from a combination of some atmospheric variables useful to assess the human body comfort (wind speed, relative humidity, and incoming solar radiation) such as apparent temperature, Universal Thermal Comfort index, excess heat factor (EHF) and excess heat index (EHI) \citep{mcgregor2015heatwaves}.

A few studies on heat wave forecasting have been carried out in West African regions. \citet{batte_forecasting_2018} evaluated the predictability of heat waves during spring at subseasonal time scale using the Météo-France model as part of the S2S project. To assess the skills of the models, they used the apparent temperature and T2m anomalies as indicators for heat wave detection.

Apparent temperature (AT) represents the temperature actually felt by humans, caused by the combined effects of air temperature, relative humidity and wind speed. The results show that the Météo-France model is able to predict heat waves up to one week in advance. \citet{guigma2021prediction} assessed the predictability of Sahelian heat waves during spring at subseasonal time scale using the ECMWF extended long-range forecast system (ENS-ext), ERA5 and BEST gridded data for the evaluation. Their approach is based on the prediction of the probability of heat waves occurrence using T2m and heat index (HI) as indicators. They show that ENS-ext is able to forecast Sahelian heat waves with significant skill up to 2 weeks in advance; and with increasing lead time, wet heat waves are more predictable than dry heat waves.

\citet{batte_forecasting_2018} assessed heat waves predictability using T2m and AT anomalies. While this approach provides information about the weather situation for the future days, it cannot provide useful information about the onset and duration of heat waves. In this study, we will adapt the methodology proposed by \citet{lavaysse2019predictability} when assessing the predictability of heat waves over Europe. \textcolor{red}{This method offers a complete evaluation of heatwave characteristics including not only the evaluation of heatwave intensity, but also of heatwave onset and duration.} It involves the computation of evaluation metrics to assess the skills of the models (see Section 2.4.6).

The present study assesses the predictability of heat wave frequency and characteristics in West African cities over the period 2001-2020 using two models part of the S2S project namely, ECMWF and UKMO. \textcolor{red}{To the author's knowledge, this work is the first of its kind in the region and represents a benchmark for future studies.} To achieve our goal, we first analyze the representation of T2m and wet bulb temperature (Tw) in the forecast models with respect to the reanalysis data used as references (see Section 2). Secondly, we evaluate the models on the representation of extreme heat events. Finally, the skill of the models in predicting heat waves is evaluated.

The remainder of this article is organized as follows: in section2, we present the region of study and the data used for this work; the description of the methodology is also provided. Section3 contains the main results of this study following the methodology presented in section 2."

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.
We followed the suggestion of the reviewer, and we splitted the long text in the document into small paragraphs.

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. 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?!

Thanks to the reviewer for this good remark and demonstration.

We have changed the conclusion on the Brier score in the document to :
" The Brier score values obtained using ERA5 reanalysis as reference, are very low between 0.05 to 0.175. We could think that the models show skills in forecasting hot days but this is quite difficult to affirm because the 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 (https://www.ecmwf.int/sites/default/files/elibrary/2007/15489-verification-probability-forecasts.pdf)."

I will not comment again on the conclusions related to FAR, GSS, and hit rate.
We are aware that the hit-rate and GSS values are very low, but this is not surprising given that heat waves are extremely rare and difficult to predict because the persistence factor comes into play. Furthermore, we also know that the forecasting models underperform in tropical regions due to a poor representation of convective processes in their physical parameterization. Consequently, these scores, which are low but greater than the climatology, are significant for assessing the skill of the models in predicting heat waves in tropical regions which remains a complex task

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.
We added the following to the manuscript :
"We used for this specific analysis, the UKMO forecasts initialized on the 1^{st} of each month". We also added this information to the other legends in the document where it was useful.

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?
We clarified this point in the document by changing to :
" The 90th percentile threshold is computed independently for the reanalyses and the hindcasts. As mentionned previously, the hindcasts are run at least once every week for a 6-week duration. For each initialization date within a month and for each lead time, we computed the daily climatological 90th percentile over the study period (2001-2020). We provided in supplement material an illustration of the computation of the 90th percentile threshold using ECMWF hindcasts initialized in January 04th (see Fig.S1 in supplement material). Heat waves are detected independently for each initialization date within a month, using the threshold values computed from this initialization (see Fig.S1_n in supplement material). For example, to detect heat waves in ECMWF hindcasts run on January 04^{th}, the daily climatological 90^{th} percentile of each day of the January 04^{th} run is applied to each corresponding day of the runs 04.January.2001, 04.January.2002 to 04.January.2020."

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.
We clarified this point in the document by changing to :
"Seasonal evolution of the daily climatological 90th percentile threshold over AT region using T2m_min ECMWF hindcasts run on: the first thursday of the month (e.g. Thursday 04th for January) (i) and the second thursday of the month (e.g. Thursday 11th for January)(ii)."

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?
Yes, that is actually what we did and we clarified it in the previous comment.

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.
We clarified this point by adding this figure in supplement material in the document.

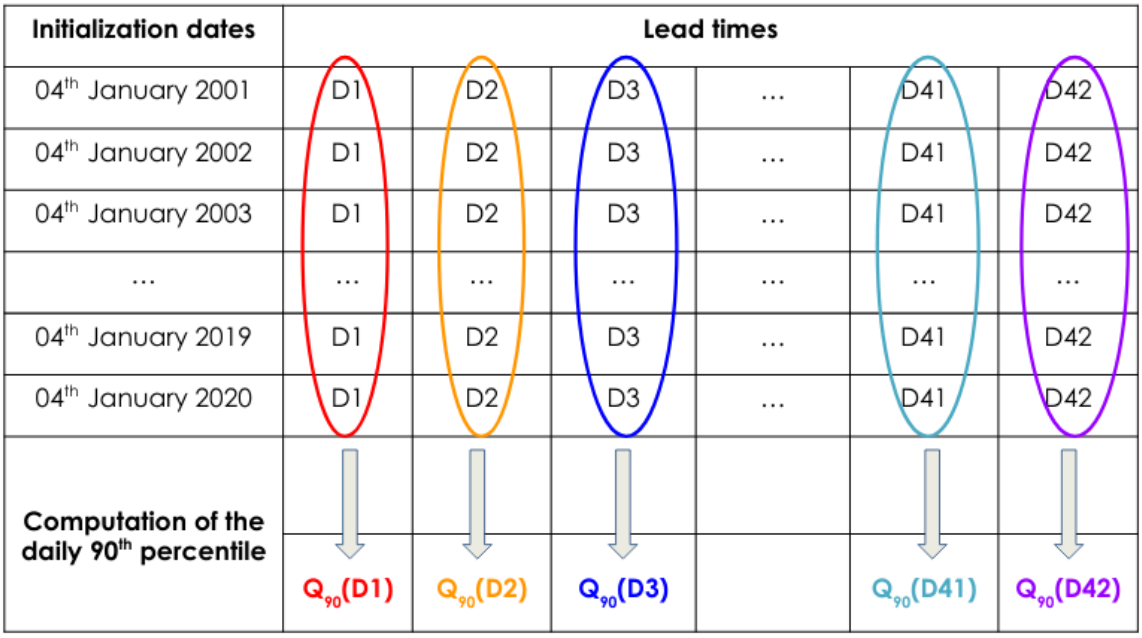

Figure S1: Computation of the daily climatological 90th percentile threshold for hindcasts initialized on January 04th for all lead times using ECMWF model. We used the same approach to compute the threshold for the other initialization dates and the rest of the months.

Some typos and other comments:

Line 31: typo in "A Heat wave"

We replaced "A Heat wave" by "A heat wave" in the manuscript.

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

We replaced "min, mean or max" by "minimum, mean, maximum" in the manuscript.

Line 4:it should be refer "Heat stress indices refers"

We replaced "Heat stress indices refers" by 'it 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.
We replaced “This work is carried out in West Africa” by “This work is carried out for West Africa”.

Line 120: Why is there “;” after stations? I do not think this is correct.
We removed the “;” in the sentence.

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.
We replaced the previous sentence by :
“In the following, we will use "MERRA" to refer to MERRA-2 which, with ERA5, are the references for the evaluation of the forecast models.”

165-167: grammar is not correct
We replaced the previous sentence by the following:
“Hindcasts are forecasts produced for past dates using the most recent version of the forecasting system. They are useful for evaluating the performance of the current version of the model over a past period.”

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. ”
We followed the suggestion of the reviewer.

Line 160 is missing a verb.

We replaced the previous sentence by :
The extended-range ECMWF forecast model is running on the Integrated Forecast System (IFS) cycle CY47R3 released on October 10th, 2021.

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?
We clarified this point in the document. We added the following :
"We only analyzed the hindcasts produced on Thursdays. In fact, an initial survey of the initialization dates of the hindcasts revealed that most of the models were initialized on the same date as ECMWF (Thursday of each week). Subsequently, we realized that all those models, with the exception of UKMO, did not cover the study period."

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.
We replaced by :
"Following the same approach as in \citet{ngoungue2023heat}, local temperatures ……. "

217: should be "occur" not occurred
We followed the suggestion of the reviewer.

219: are detected and not is detected
We followed the suggestion of the reviewer.

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

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 …
We followed the suggestion of the reviewer.

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.
We clarified this point in the document by adding the following:
"The intensity of a heat wave was defined as the sum of the daily exceedances of the indicator values to a daily threshold during the event".

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..
We changed the section name "2.4 Metrics" by "2.4 Methods"
In addition to the points of reviewer 1, I add hereby the comments from the second reviewer:

While I can see that many of my comments were addressed, some of the questions requesting more explanation, e.g., the definitions of the heatwaves themselves and any links to impacts, are not answered. For example, the definition of wet/dry heatwaves – it seems now, day/night heatwaves are changed to dry/wet, and it seems odd that these would be interchangeable.
We clarified the definition of heat wave in the document by adding the following:
" In the present study, two types of heat waves are investigated : dry and wet heat waves. Dry heat waves are mostly driven by incoming solar radiation and occur during the day. The detection of dry heat waves is processed using

maximum values of T2m (T2m_max) as indicator. The most lethal heat waves are due not only to high temperatures but also to the effect of humidity \citep{steadman_assessment_1979,steadman_assessment_1979-1}. Humidity is an important driver of wet heat waves. Wet heat waves are detected using minimum values of T2m (T2m_min) and mean Tw as indicators. T2m_min is also chosen for wet heat waves because relative humidity is higher at night and decreases during the day due the changes in temperature."

The split into the four seasons is not well explained, given that the authors describe the area as having one dry season and one wet season. I can see that effort has been made to include additional information, but still work is needed to improve the structure, narrative, and methodology of the paper.

We clarified this explanation by changing the old paragraph to :

"Heat waves in the Sahel region occur mainly in spring due to the high temperatures in the region at that time (Barbier et al.,2018; Guigma et al., 2020). In this study, the region of interest was extended to the Guinean region in which heat waves are mainly driven by humidity coming from the Atlantic ocean. This advection of humidity over the Guinean region is active during the season. Therefore, the detection of heat waves is performed over the whole season and not just in spring, to cover the wet and dry seasons in the region. "

I am still unconvinced on several points raised in the first review, around the heatwave definitions, the treatment of ensemble members, and the other review articulates very well the concerns and confusion around the verification aspects. The authors appear to also include results regarding heatwave intensity, while also commenting that their threshold approach focuses not intensity but on number of events.

We added some information to clarify the treatment of ensemble members in the text :

"Probabilistic scores exist for evaluating ensemble forecasting systems, but when it comes to the evaluation of specific events such as heat waves, the task

becomes more complex. For example, how do you evaluate a model for which some ensemble members are forecasting a heat wave? To solve this issue, ensemble forecasts are converted into boolean files using threshold values based on the percentage of members correctly forecasting heat wave days. Three threshold values were tested to find the optimal boolean files (see [Table\ref{tab:Table2}])"

For the verification aspects mainly the Brier score, we addressed it in the reviewer 1 comments.