Review of nhess-2023-144: **Subseasonal-to-seasonal forecasts of Heat waves in West African cities**

Review overview

The concept of this study is certainly of interest, and evaluation of extreme events such as heatwaves, which are increasing in frequency, intensity and duration, is important at a range of timescales. I read this paper with interest. This study examines the extended-range/subseasonal timescale, which can be useful for providing early indications of potentially hazardous events, ahead of more detailed forecasts that shorter-range forecasting systems are capable of predicting.

While I find the concept useful and interesting, and I like the range of skill scores used, I did find that several aspects of the methodology are not described clearly and there are some questions around the datasets and methodology used. Much more clarification is required around the forecasts used for this evaluation, and discussion of the potential drawbacks. The authors consistently refer to 'subseasonal to seasonal', which, while catchy, is not completely covered here, as the seasonal time frame is not included. The descriptions should probably be changed to subseasonal throughout. I also had some questions around the identification of heatwaves and the thresholds used, and the method of dealing with ensemble members. The abstract and introduction mention both wet and dry heatwaves, and daytime and nighttime, but these distinctions are not clearly defined and discussed, and appear to be mostly lacking from the rest of the paper and the results and conclusions.

While the research questions and results are interesting, I feel that the structure of the writing could be significantly improved throughout the paper, as it is currently challenging in places to follow the work, and to fully understand the somewhat contrasting conclusions. For a decision-maker, what are the takeaways to help understand how these forecasts could/could not be used in heatwave forecasting and anticipatory action?

Thank you to the reviewer for taking the time to revise this document. The comments made by the reviewer are very insightful. The predictability of heat waves is assessed using two sub-seasonal forecasting models involved in the S2S project, namely "ECMWF" and "UKMO". We totally agree with the reviewer that the forecast products do not cover the seasonal range, and we have replaced the term "sub-seasonal to seasonal" with "sub-seasonal" throughout the manuscript. Heat waves are evaluated in the models using dichotomous measures such as success rate, false alarm ratio and Gilbert skill score. Two types of heat waves are studied: wet heat waves are those resulting from a combination of humidity and air temperature, while dry heat waves are associated with maximum air temperatures. On the basis of the reviewer's comments, we have clarified some points in the main document and improved its quality. The results of the present study show predictive skills in sub-seasonal forecasting models up to two weeks in most cities of the region of study, however they overestimated the occurrence of heat waves. The results are useful for policy makers to develop early warning systems to prevent the population from potential heat waves. We added this aspect of the results to the main document.

I have provided some more specific comments on the text and some of the figures below. I hope these can be useful as the authors consider the revisions and next stages of the manuscript.

Detailed comments

Abstract

- Line 13: Short-term forecasts typically refers to those of <4 days – 2 weeks lead time would typically be classed as medium-range forecasting

  We replaced "short-term forecasts" in the manuscript by "medium range forecasts"

- Line 15: Fail is a strong word, without context?

We replaced "they fail in predicting the intensity of heat waves." by " the accurate forecasts of the intensity of heat waves remains challenging by the models"

Introduction

- Line 30-35: It could be worth mentioning that often, national meteorological services have a definition of a heatwave used to provide warnings? (unless it is not the case in the study region, but otherwise, there is also a WMO recommended heatwave definition (https://www.un-spider.org/category/disaster-type/extreme-temperature).

  We added the definition of heatwave given by WMO in the introduction "A period of marked unusual hot weather (maximum, minimum and daily average 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."

- Same comment at line 41, research paper authors are not the only ones / the authoritative ones to define heatwaves, particularly in a forecasting perspective. Is there a definition used most often by the forecasting services based in the study region?

  We did not find a definition of heat wave provided by the forecasting services in the study area. They used the same definition provided by the literature review.

- Line 39: 'min, min or max' – is there a typo here? Min seems repeated

  That was a mistake. We replaced 'min, min or max' by " min, mean or max"

- Line 39: heat stress indices are mentioned, but not really defined anywhere? (check and come back to) – it may be useful to define here what a heat stress index is and how it differs from the other metrics listed

We added "The heat stress indices 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) (McGregor et al., 2015)".

- Line 45: It is of course of crucial importance for early warning systems to provide information on the occurrence of heatwaves. However, early warning systems are not usually done using seasonal weather forecast models, which often lack the skill and resolution to accurately predict individual extreme events. Typically, an early warning system would refer to a shorter/medium-range lead time, supplemented with advanced information on the potential for hazardous weather using S2S forecasts. The authors go on to make this point about seasonal forecasts providing early indications, which I completely agree with, but early warning systems require a range of lead times, including shorter timescales to account for the fact that forecasts get much more accurate at shorter lead times.

  Thanks to the reviewer for this clarification. We changed: " This is usually done using seasonal weather forecast models" by :

  " In general, early warning systems integrated shorter and medium-range forecasts of potential weather hazards. This type of forecast window refers to sub-seasonal time scale from 2 up to 6 weeks. The sub-seasonal range is highly relevant for actions aimed at mitigating the human and health consequences of extreme heat [e.g. \citet{white2017potential,moron2018sub,tompkins2019predicting,osman2023sub}]. Sub-seasonal forecasts are used to monitor the evolution of specific weather patterns that have been identified in advance with seasonal forecasts. "

- Line 49-50: citation?

  We added the following references "Bazo et al., 2019; Lala et al., 2020"

- Line 51-xx: I saw that Vitart (et al) also studied the Pacific Northwest heatwave of 2021, considering the ECMWF subseasonal forecasts and a more recent version of the ECMWF model, and 9 other S2S models. This may be of interest for the authors to include, as it uses a more recent model version than the Russian heatwave studies. https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL097036
  Other authors also examined other time ranges of the forecasts for this heatwave.

  Thanks to the reviewer, we added this work in the manuscript.

- Line 114: ERA5 is used to initialise the extended-range **reforecasts**, but not all of the different aspects of ECMWF's IFS

  This is a good remark.

  We replaced "Since ERA5 is used to initialize the atmospheric component of the ECMWF model which is one of the forecast models" by

  " Since ERA5 is used to initialize the atmospheric component of the ECMWF extended reforecasts (ENS-ext) "

Section 2

- Section 2.1: it could be interesting in this section to include some description of heatwaves and their impacts in this region – what have been the impacts of significant past heat waves? What kinds of temperatures are reached? During what season do impactful heatwaves occur?

  We added this information to the manuscript :

  "In April 2010, North Africa was affected by a severe heatwave, with daily maximum and minimum temperatures exceeding 40°C and 27°C respectively over a period of 5 days. This event was disastrous for the population and caused significant material damage. In May 2013, the Senegalese city of Matam, experienced an intense heat wave with temperatures sometimes reaching 50°C in the shade. The event was

persistent both during the day and night, and it caused 18 deaths among the elderly people in 10 days. Mauritania also experienced a devastating heatwave in May 2013, with maximum temperatures exceeding 46°C, causing the death of more than 25 elderly people and children."

- Section 2.1: this section refers to the region having a short wet season followed by a long dry season. But the results later are split into winter/spring/summer/autumn – this should be further explained and justified.

  We added this information in the document:

  "The evaluation of the skills of the models to predict heat waves is carried out from January to December to cover the wet and dry seasons in the region. The results were then splitted into four sub-seasons to assess the intra-seasonal variability of the skills of the models."

- Section 2.2: A brief discussion on the potential disadvantages of reanalysis datasets might be warranted, if not included later (e.g. they may not always have the resolution to be able to pick up the very highest temperatures during heatwaves)

  Thanks to the reviewer for this suggestion. We added the following to the manuscript: " Another point to highlight in this work is the use of reanalyses to evaluate heatwave forecasting models. We are aware that reanalysis data have a high resolution compared to observations from local stations. As a result, they are unable to represent the urban heat island effect which exacerbates heat stress during heat waves. The resolution of the reanalyses makes it impossible to detect the highest temperatures at specific locations.

Section 2.2.2: The assessment of dry and wet heat waves seems like it should be a separate section, as at the moment it seems to be that it is only applied to MERRA, as it sits under that subtitle. It is also not clear

how the heatwave identification described later takes into account both wet and dry heatwaves, nor is this clear in the results. Wet and dry heatwaves should also be defined (do they refer to the afore mentioned wet and dry seasons? to the humidity experienced during a heatwave? Or otherwise?)

We clarified this clarification in subsection 2.1 Region of interest

" In this study, two types of heat waves are analyzed : dry and wet heat waves. Dry heat waves are mostly driven by incoming solar radiation and occuring during the day. The detection of dry heat waves is processed using maximum values of T2m as indicator. Wet heat waves are the most lethal for human health. Humidity is an important driver of wet heat waves. The detection of wet heat waves is done by using minimum values of T2m and mean wet bulb temperature as indicators. "

- Section 2.3.1 ECMWF forecasts: there are some errors in this section, and I find some of the description unclear. This isn't helped by the fact that ECMWF very recently upgraded their models so some of this information no longer stands. It may be useful to revisit the description in this section for clarity.

  Thanks to the reviewer for this remark

  More details below:

- The ECMWF IFS has several separate forecasting systems (medium-range (now high-resolution), extended-range, seasonal), and it would be useful to specify which is being used and described here, as other parts of the system have different resolutions, lead times and ensemble members.
- ECMWF provides both extended-range (up to 46 days) and seasonal (up to 6 months) forecasts to the S2S programme. I understand that the authors are using the extended-range forecasts, and it may be useful to refer to the forecasts as this throughout.

- ‘ECMWF ENS’ is often used to refer to the medium-range (up to 15 days and now high-resolution) ensemble, and so could cause confusion – these are not the same exact forecasting system as the extended-range (at least not any more).
- The authors may wish to specify that they are interpolating to a 0.25° grid to match the resolution of ERA5 for evaluation (I presume), with the caveat that this does reduce the resolution of the native forecasts, and that higher resolutions can be beneficial for capturing extremes.
- The IFS is no longer running at CY41r2. If the authors downloaded hindcast data from forecast dates in 2021, the cycles could have been 47r1 (implemented 30 June 2020), 47r2 (implemented 11 May 2021), or 47r3 (implemented 12 October 2021). The authors should confirm which cycle(s) was used. These cycles indeed had 51 ensemble members at 18 and 36km resolution depending on the lead time, as the authors describe. The latest version of the extended-range forecasts (48r1) has 101 members, run at 36km for the full forecast range (days 0 to 46), and is run daily rather than twice a week. (https://www.ecmwf.int/en/about/media-centre/news/2023/model-upgrade-increases-skill-and-unifies-medium-range-resolution)
- A useful description of hindcasts/reforecasts may be ‘Hindcasts are forecasts produced for past dates using the most recent version of the forecasting system, and allow analysis of how the current system would have performed, alongside a consistent dataset covering a longer time period for evaluation’, as a useful use of these data for the authors’ purposes?
- The authors also use ‘hindcasts’ and ‘reforecasts’ interchangeably. ECMWF typically call them reforecasts, and the authors should be clearer if it is themselves calling them ‘hindcasts’ throughout the study.
- I didn’t understand the explanation for not using the Thursday hindcasts, sorry

We modified this section according to the reviewer's remarks. The new section is the following :

" 2.3.1 ECMWF forecasts

The extended-range ECMWF forecast model runs on the Integrated Forecast System (IFS) cycle CY47R3 released on October 10th, 2021. The native spatial resolution of the ECMWF model is Tco639 L137 (about 16 km) up to day 15 and Tco319 (about 32 km) after day 15, but the downloaded data are interpolated to a regular 0.25°x 0.25° latitude/longitude grid to match the resolution of ERA5 for evaluation. It contains 91 sigma levels from the surface to 80 km. ECMWF provides two types of outputs for the S2S program: real-time forecasts and reforecasts called "hindcasts". Real-time forecasts are forecasts for the coming days. Hindcasts are forecasts produced for past dates using the most recent version of the forecasting system, and allow analysis of how the current system would have performed, alongside a consistent dataset covering a longer time period for evaluation. ECMWF extended-range real-time forecasts are run with 51 ensemble members (50 perturbed and 1 unperturbed), while hindcasts are run with 11 members. In this study, we focus on hindcasts only. ECMWF extended-range hindcasts are produced twice a week, on Monday and Thursday at 00Z. This means that for each week a new set of hindcasts is produced to calibrate the real-time ensemble forecasts for Monday and Thursday of the following week using the latest version of the IFS. We only analyzed the hindcasts produced on Thursday. This is because we firstly want to carry out a multi-model analysis. According to a first investigation on the initialization dates of the hindcasts of different models, we found that most of the models were initialized on the same date as ECMWF (Thursday of each week) but did not cover the study period. The 11-member ensemble hindcasts start on the same day and month as the real-time forecast, but covering the last 20 years. In our case, the forecast year is 2021 and we focus on the previous 20 years from

that date, and the hindcasts run from 0-46 days. The variables of interest in the ECMWF S2S are T2m(max,min) over the last 6 hours, daily average T2m and d2m from which the daily average Tw was derived. The data are open access and available on the S2S project website (https://apps.ecmwf.int/datasets/data/s2s-realtime-instantaneous-accum-ecmf/levtype=sfc/type=cf/)."

- Line 83: the authors may wish to acknowledge that when dealing with extreme events, including different extreme events in the analysis may well result in different conclusions regarding the skill.

  Yes, we agree with the reviewer, but we didn't catch the link to the present.

- Section 2.3.2: Parts of the UKMO forecast description are also confusing, for example the transition from discussing 4 members to 7 members. Perhaps a table outlining key aspects of both forecasting systems, and the timeframes to which they apply, would be helpful to provide an overview of the system characteristics?

  We clarified this point in the manuscript by replacing :

  " The UKMO real-time forecast consists of a set of 4 members run daily for a period of 60 days (3 perturbed members and 1 control member). The UKMO hindcasts are produced 4 times per month, on the 1st, 9th, 17th and 25th, and cover a 24-year period from 1993 to 2016. We are aware that these initialization dates are not the same as those of ECMWF, but we are interested in this work on the predictability of heat waves in a broad perspective, not on specific events. The ensemble hindcasts are composed of 7 members per cycle (from the 25 March 2017 hindcasts, prior to that 3 members per cycle)." by

  "The UKMO real-time forecast consists of a set of 4 members (3 perturbed members and 1 control member) run daily for a period of 60 days. The UKMO hindcasts are produced 4 times per month, on the 1st,

9th, 17th and 25th, and cover a 24-year period from 1993 to 2016. We are aware that these initialization dates are not the same as those of ECMWF, but we are interested in this work on the predictability of heat waves in a broad perspective, not on specific events. Prior to 2017, specifically on March 25th, the UKMO ensemble hindcasts were composed of 3 members per cycle (2 perturbed and 1 control). Since 2017, the number of members has increased from 3 to 7 (6 perturbed and 1 control)."

- Section 2.3.2: I believe the description of the concatenation could be simpler. Is an equation necessary, or is it enough to simply state that prior to 2016, hindcasts are used, and after that, the real-time forecasts are used, followed by the details from line 190?

  We agree with the reviewer that the description of the concatenation applied here can be simplified, but through this equation we want to highlight the complexity behind this data processing task. We will keep the equation in the document.

- I am not completely convinced of the decision to reduce the number of ensemble members in this methodology, thus reducing the uncertainty representation of the forecast.

  We added this explanation to the main document :

  "In order to apply the concatenation over time between the re-forecasts and real time forecasts, the coordinates dimensions of the two datasets must be the same. As shown early, the number of ensemble members in UKMO re-forecasts and real time forecasts are completely different. Therefore, to meet this requirement, we reduced the number of ensemble members from 7 to 4 (1-control member and 3-perturbed members)in the re-forecasts to match the number of ensemble members in the real-time forecasts."

  It would also be useful to provide an overview of how other characteristics of the model have changed between the hindcast version and the potentially multiple operational versions used during

the period of the real-time forecasts? This could be covered in the aforementioned table.

We added the following table in the document.

Models	Hindcasts				Real time forecasts			
	dates	***size***	***range***	***period***	***dates***	***size***	***range***	***Model version***
ECMWF	2/week, on Monday and Thursday	11	0-46 days	past 20 years	2/week, on Monday and Thursday	51	0-46 days	CY48R1
UKMO	4/month on the 1st, 9th, 17th, 25th	3 prior 2016 7 from 25/03/2017	0-60 days	1993-2016	4/month 1st, 9th, 17th, 25th	4	0-60 days	GloSea5-GC2-LI

- Line 197: I believe here the authors are referring to a lack of data from local stations to evaluate the forecasts again. The sentence implies that no data is available from this region for weather forecasts to assimilate in their production – are the authors sure this is the case? Particularly since weather forecasts also use various other sources of observations beyond station data.

  We clarified this point in the manuscript as follow:

  We replaced: "Weather forecasts provide the evolution of atmospheric variables on a global scale, which implies the need to have data from local stations to access information on a local scale. This is a major problem in areas where there is a lack of weather stations to collect data, as is the case in African cities." by

  "Weather forecasts provide the evolution of atmospheric variables on a global scale, which implies the need to have data from observation stations to access information on a local scale. This is a major problem in areas where there are not enough weather stations to collect data, as is the case in African cities. Nevertheless, when observation stations are available in the region, access to the data collected remains difficult."

- Section 2.4.2:
    - Are the daily maximum, daily minimum and wet bulb computed from the hourly data? Or otherwise?

      We added this information to text :

      "Daily maximum and minimum temperatures are computed respectively from maximum and minimum temperatures in the last 6 hours. This choice of the computation of the extreme daily values is made according to the forecast models outputs. Daily average wet bulb temperature is computed from hourly dew point temperature."

    - Are nighttime and daytime heatwaves considered separately, or as one continuous heatwave that does/does not provide relief overnight? This can have implications for heat stress and health, but it is not clear how it is factored into the authors' definition of a heatwave. I think it is hinted at, but was not entirely clear to me in the definition.

      We clarified this point in the manuscript:

      "Nighttime and daytime heat waves are considered separately in the study. Nighttime heat waves are detected using minimum values of indicators, while for daytime heat waves, maximum values of the indicators are used. "

    - Is the 90^{th} percentile representative of the health impact of heatwaves on humans / ecosystems?

      Thanks to the reviewer for this interesting remark.

      According to previous studies on heat waves and their impacts on human health, the 90th percentile appears to be a sufficient threshold for heat waves detection (e.g., Fischer and Schär,

2010; Perkins et al., 2012; Perkins and Alexander, 2013; Fontaine et al., 2013; Russo et al., 2016; Lavaysse et al., 2018; Ngoungue et al., 2023).

- What if the 90th percentile does not reach a temperature likely to cause heat stress? Why not use a temperature or wet bulb threshold known to cause health impacts in this region?

  We added more explanations about this point in the manuscript :

  "Heat waves in the Sahel region occur mainly in spring due to the high temperatures in the region at that time \citep{barbier_detection_2018,guigma_characteristics_2020}. In this study, the region of interest was extended to the Guinean region in which heat waves are mainly driven by humidity. Heat wave detection was then carried out using the 90^{th} percentile as a threshold over the January to December season. The 90^{th} percentile appears to be a sufficient threshold for monitoring heat waves affecting human health. Nevertheless, it is useful to calculate the intensity of events in order to determine a classification according to their severity (intensity), from "harmless" to "extremely dangerous", for example. This is what the STEWARD project is doing by developing a database on heat waves and their potential impact on human health."

- Line 213-214 states the 90th percentile is computed over the entire period, and then line 215 says it's calculated for each day of the year. I am left unsure as to which of these is used (or which is used for which analysis, if both are used at different points), and this could be quite impactful for the results.

  We clarified this point by replacing :

  "We defined a heat wave as a consecutive period of at least 3 days during which the daily temperatures exceed the calendar 90th percentile threshold computed over the entire period for

T2m_min, T2m_max or Tw respectively. The 90th percentile is calculated for each calendar day of the year." by :

" We defined a heat wave as a consecutive period of at least 3 days during which the daily temperatures exceed the calendar 90th percentile threshold computed over the entire period for T2m_min, T2m_max or Tw respectively."

- Section 2.4.3; the description of these steps could be simplified and clarified further. The two first points may not really be necessary to spell out, and the third could perhaps be simplified, but it is also not clear over which timeframe this is done. Is it done for each day of the time series?

  We clarified this point by replacing :

  "To determine the occurrence and duration of heat waves, we create boolean files from T2m_min, T2m_max and Tw time series at each grid point following the steps below :

  for days in T2m_min, T2m_max or Tw time series, if days are hot days, we replace in our zero vector the values corresponding to those days by 1. Hot days are days with T2m_min, T2m_max or Tw above the 90th percentile daily thresholds. In order to assess the characteristics of heat waves, only hot days belonging to heat wave sequences are kept. This is applied for all grid points and we obtain boolean files containing 0 or 1 (Ngoungue et al. 2023). These boolean files will be processed both for the reanalyses and the forecasts to assess the representation of heat waves occurrence and duration." by

  "To determine the occurrence and duration of heat waves, we create individual boolean files from the T2m_min, T2m_max and Tw time series at each grid point, containing 1 if hot days and 0 otherwise. This operation is performed on a daily time scale over the study period. Hot days are days on which the values of T2m_min, T2m_max or Tw are above the daily 90th percentile thresholds. In order to assess the characteristics of heat waves, only hot days belonging to heat wave

sequences are considered (Ngoungue et al. 2023). Boolean files are calculated separately for reanalyses and forecasts in order to assess the representation of heat wave occurrence and duration."

- And then if the number of hot days in a row is not >=3, the value is returned to 1?

  No, as we define a heat wave as at least 3 consecutive hot days, if the number of hot days in a row is not >=3, the value is returned to 0.

- Line 232-233 isn't clear to me, apologies.

  We clarified this point in the manuscript, by replacing:

  "The intensity of heat waves was defined as the cumulative sum of the daily exceedances of daily values of indicators to the climatological daily threshold in a sequence of hot days " by

  "The intensity of a heat wave was defined as the sum of the daily exceedances of the indicators values to the climatological threshold during the event."

- Line 236-238: the reasoning behind this, and how this is applied in the methodology, isn't clear to me. Why the minimum of the daily thresholds? Does this correspond to a value that is certain to have an impact on human health? How does this allow proper assessment of the severity? Please expand on this. This relates to a previous point about using percentile thresholds, when using set values corresponding to heat stress may be both simpler and more effective.

  We clarified this point by adding this information in the document :

  " This study is in the framework of the project Agence National de la Recherche STEWARd (STatistical Early WArning systems of weather-related Risks from probabilistic forecasts, over cities in West Africa) project which focuses on the human impacts of climate extremes. We are therefore interested in heat waves, which can be harmful to human health. To do so, the climatological daily threshold is chosen to be constant over the whole period for the computation of

heat waves intensity. It is defined as the minimum of the daily climatology 90th percentile over the study period. This approach allows us to properly assess the severity of a heat wave and its potential human impacts, therefore, most dangerous heat waves will have higher intensity values."

- Line 243: ensemble forecasting does not only account for uncertainties in the physical component of the model, but also uncertainty arising from the chaotic nature of the atmosphere, and from an imperfect observation network and therefore imperfect initial conditions of the forecast.

  Thanks to the reviewer for the suggestion, we added this information to the manuscript.

- Line 247-249: By considering the mean, medium, warmest, coolest, 1st and 3rd ensemble members, you have identified 6 'members'. The Met Office forecasts only have 2 or 7 members, and the ECMWF forecasts 11 members, so I am unsure as to why it is less challenging to use these 6 'members' chosen by the authors, rather than more usefully examining the entire ensemble and therefore the full range of uncertainty represented by the ensemble? It should also be considered that the mean (and quartiles, depending on how these are produced) do not represent an actual forecast scenario or physically likely state of the atmosphere, produced by the model, and so caution is required in assessing this both as a forecast and in evaluating it.

  Here, we don't compute any of these statistics, we just want to show how it is difficult to evaluate ensemble forecast systems based on the amount of information they provide (mean, median, warmest, coolest, 1st and 3rd quartiles of the ensemble members, ensemble members).

Section 3

- some paragraphs would be helpful for readability in sections 3,4,5

  We added some text to facilitate the comprehension of sections 3,4 and 5

- Section 3.1: the use of 'hot bias' and 'cold bias' is quite strong wording, as opposed to positive and negative. How large are the biases? It is not mentioned in the text,but some of these 'hot' biases may only be a small fraction of a degree, so hot might not be the most appropriate choice of wording?

  The bias found in the study varies with the variables, with T2m the bias is around -4 and 4 K, while it is more important with Tw between -12 and 0 K. We replaced the terms 'hot bias' and 'cold bias' in the manuscript by 'positive bias' and 'negative bias' respectively.

- Given that the authors state that the results comparing to MERRA are significantly different to those using ERA5, I am surprised not to see some figures included in the main text. How does this discrepancy impact the evaluation results, if the two verification datasets are so different?

  We have noticed some differences between MERRA and ERA5 when carrying out some analyses on the forecasting models, for example the estimation of the bias in the evolution of T2m and Tw, and the spatial variability of heat wave duration. However, the assessment of the CRPS and the Brier remains similar for both reanalyses. Therefore, for the evaluation of the predictability of heat waves in the models, we used ERA5 as a reference. We agree that the heat wave predictability results for MERRA will not be exactly the same as those obtained with ERA5, but the evaluation metrics will be of the same order. Some figures with MERRA are added to the manuscript.

  Line 321: the plots are shown in °C, but the text uses K – why refer to it differently between the texts and figures?

  That's a good point. It's a mistake that we've corrected.

- Section 3.2: Why are the Tw results only shown in supplementary material if they make up an important part of the research question and results?

That's a good point. We added some results on the Brier score, spatial variability of heat waves duration and the metrics evaluation with Tw in the main text.

- Section 3.3.2: Why is the mean duration the sum divided by the number of affected years, rather than divided by the number of heatwaves? (what if there is more than one heatwave per year?)

  We are interested here in the average characteristics of heat waves over the period in which they occurred. As a result, the average duration of heat waves corresponds to the sum of heatwave days divided by the number of years concerned.

- Line 375: can the authors comment on the representation of convection in both models?

  "The representation of the convective activity in ECMWF, is done using the Tiedtke scheme (Tiedtke, 1989) and UKMO, the Met Office convective scheme (Hagelin et al., 2017)."

- Section 3.4: could the authors explain further the reasoning behind the 20%, 40% and 60% percentile thresholds? I did not follow the aim and reasoning here. The text and Figure 10 seem to refer to a section of the methods that I was unable to find. Perhaps it refers to the last sentences of section 2.4.4, but I did not follow the link, and further explanation may be required.

  We clarified this point in the manuscript by added :

  "The results presented below are obtained using a 20\% threshold value to optimize the ensemble forecast system (see Section 2.4.4) ."

  An initial evaluation of the forecast models was performed using all ensemble members, and the metrics (hit_rate, FAR, GSS) were calculated for each member. Then, the evaluation metrics are calculated as the average of the metrics for all members. This approach is not suitable for model evaluation. In order to optimize the forecast systems, we transformed the probabilistic forecasts into

deterministic forecasts using threshold values, following the methodology proposed by Lavaysse et al. 2019.

- Regarding seasons, are there seasons where there may technically be heatwaves as the temperature exceeds the 90th percentile for the time of year, but they would not cause heat stress or health impacts? Should these be considered in the same way as those during other seasons? Why are winter/spring/summer/autumn used if the region experiences two seasons (dry/wet) – how do these correspond?

  These questions have been clarified previously (see the second part of the discussion in the manuscript).

  Some context regarding heatwaves themselves and the temperatures reached and impacts in this region could provide interesting further insight (for example in section 2.1 this could be added).

  This has been done previously in section 2.

- From a decision-making perspective, it would be interesting to understand how far in advance these forecasting systems may be able to provide a useful prediction/indication of a heatwave. The results are interesting from a modeling perspective, but I finish reading the results section feeling that I would not really have a confident answer to this question. Could the discussion be expanded to consider the results in this context?

  We added this part to the discussion :

  " This study showed that the forecast models were actually able to predict heat waves occurrence up to two weeks in advance in the different regions. On the other hand, we found that the models overestimated the frequency and duration of events, whatever the lead time. Consequently, it will be necessary to find a good balance between hits and false alarms in order to develop a robust early warning system to prevent populations from heatwaves. "

Section 4

- Line 456-460: the names of the convection schemes unfortunately do not mean much to me – what are the key differences and the implications?

  We added this information to the main document :

  "The Tiedtke convection scheme is one of the first mass-flow convection schemes, which aims to parameterise the effects of deep convection in numerical weather models. It simulates the vertical transport of heat, moisture and momentum associated with convective updrafts and downdrafts. The system takes into account various factors, including atmospheric instability, moisture content and boundary layer conditions to estimate convective processes. UKMO also uses a mass flux convection scheme, but different from the Tiedtke scheme, which takes into account atmospheric instability and moisture content to determine convective activity. The difference between the two convective schemes could lead to a wrong representation of convective activity in the region, and thus limit the predictive skills of the models mostly for wet heat waves."

- Line 461: could the authors expand on 'the data and initial conditions are completely different' ?

  We added this information to the main document :

  "ECMWF assimilates a wide range of global and regional observational data, including satellite, radar and ground-based measurements. The UKMO focuses on observation data relevant to the United Kingdom and surrounding regions. ECMWF uses a 4D-Var assimilation which considers the temporal dimension (four dimensions) in addition to the three spatial dimensions to generate the initial condition. The UKMO

employs two data assimilation techniques : the 4D-Var and Ensemble Variational (En-Var) to estimate the initial state of the atmosphere."

- Line 465: did the authors not reduce the resolution of both forecasts? What impact could this have? Particularly on the discussion of all results relating to the spatial variability and the intensity

  We added this information to the main document :

  "The native resolution of the models has been transformed into a regular 0.25*0.25 grid. Even if we transform the native resolution of the two models into a regular 0.25°x0.25° grid, some local-scale patterns will be found in the new grid. However the impact of the resolution on the skills of the models is not assessed."

- Overall, I find the discussion section raises some interesting points, but does not really expand on why or how they influence the results

  In fact, in this study we have identified some differences between the two models that may explain the differences in predictions in the regions. A more detailed analysis of the influence of each factor on the results is beyond the scope of this paper.

Section 5

- Line 484-485: it was not clear where the key results were that make any distinction between daytime and nighttime heatwaves and how this was handled in the methodology. An interesting aspect of heatwaves is the drop in temperature overnight, and whether this provides any relief from the daytime heat stress, but this doesn't factor into the discussion at all.

  This has been clarified in the previous comment on the definition of daytime and nighttime heat waves(see Section 2.1).

  We added this to the manuscript :

  "The prediction of dry heat waves is slightly better with ECMWF for medium range forecasts, while it is better with UKMO for long-range

forecasts. For wet heat waves, UKMO outperforms ECMWF for both medium- and long-range forecasts."

- What do the authors consider as a nighttime heatwave, one that only occurs at night and not also in the day? It is a little confusing, and more context and insights could probably be included.

  We clarified this point in the previous comment. Yes, nighttime heat waves are those occurring during the night. In the document, we decided to replace "nighttime" and "daytime" heat waves by "wet" and "dry" heat waves to avoid some confusions.

- On a similar note, it is not clearly defined the difference between a wet and a dry heatwave, other than the use of different variables. These terms are only really use in the introduction and conclusions, but the link to the results is missing and the methods are not entirely clear.

  We clarified this point in the previous comment.

- Line 491: what is counted as a failure to predict the intensity? At what lead time? This is a very broad statement.

  We changed " they fail in predicting the intensity of heat waves; the accurate forecast of heat waves intensity remains a challenging task for the models." by :

  "They underestimate the intensity of heat waves with respect to ERA5 at short, medium and long range forecasts. "

  It implies the forecasts are not useful at all – do the authors conclude in this paper that extended-range forecasts are not useful for predicting heat waves? Can they be used or interpreted at all to complement short-range forecasts? Can some information be provided on the skill of short-range forecasts (could be from other studies), to provide context?

  This does not mean that forecasts are not useful at all; they show skill in capturing some heatwave events at medium and long range. We added this information to the main document :

"Regarding these results, we can recommend the use of subseasonal forecasts to predict the occurrence of heat waves up to two weeks in advance, but as far as their intensity is concerned, it is still challenging."

How do the authors tie in these results, with the earlier statements that based on some skill scores, the models can detect extreme events up to 5 weeks ahead? Detect in what sense?

We have clarified this point in the manuscript, the term "detect" used here is a synonym for "forecast". As heat waves are defined as persistent extreme events, we first assessed the representation of single extreme events in the forecasts. To do this, we calculated the Brier score, and this first assessment does not take into account the persistence of the events. The second assessment concerns the predictability of heat waves in the models, by calculating the hit-rate, FAR and GSS.

Figures

Fig 3: The use of (a) and (b) for both the upper and lower panels and the individual panels is a little confusing at first. Perhaps consider (i) and (ii) for the panels? (or just upper and lower?), or split this into two figures.

We replaced (a) and (b) by (i) and (ii) according to the reviewer comment

Fig. 4: The colour scale here is misleading – it should be adjusted so that the white colour falls at 0, with positive and red and negative in blue, otherwise it is very challenge to properly assess where there is a warm/cold bias, particular with a gradient rather than discrete colour bar. The colour scale should reach the same value at the positive and negative ends.

This has been done according to the reviewer comment and the new figure is the following.

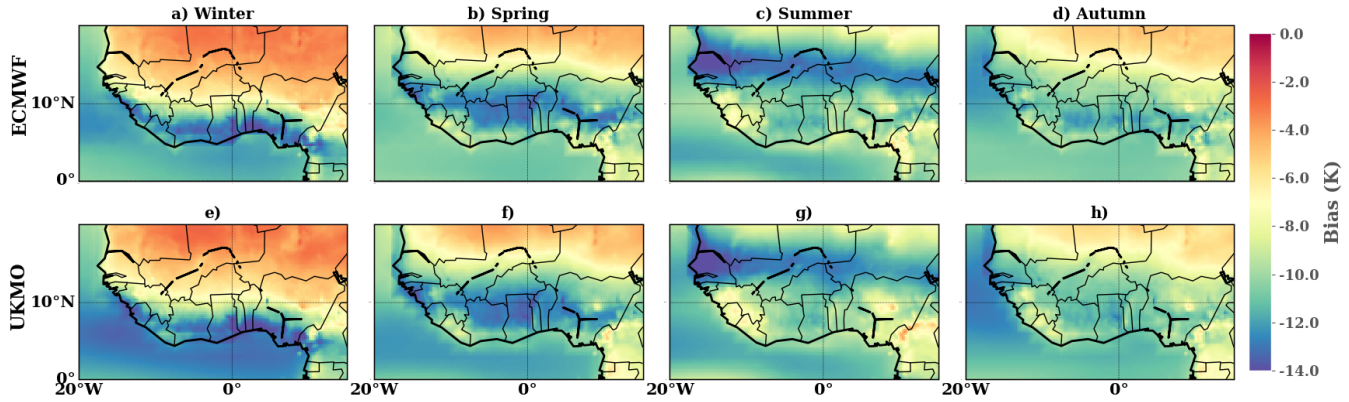
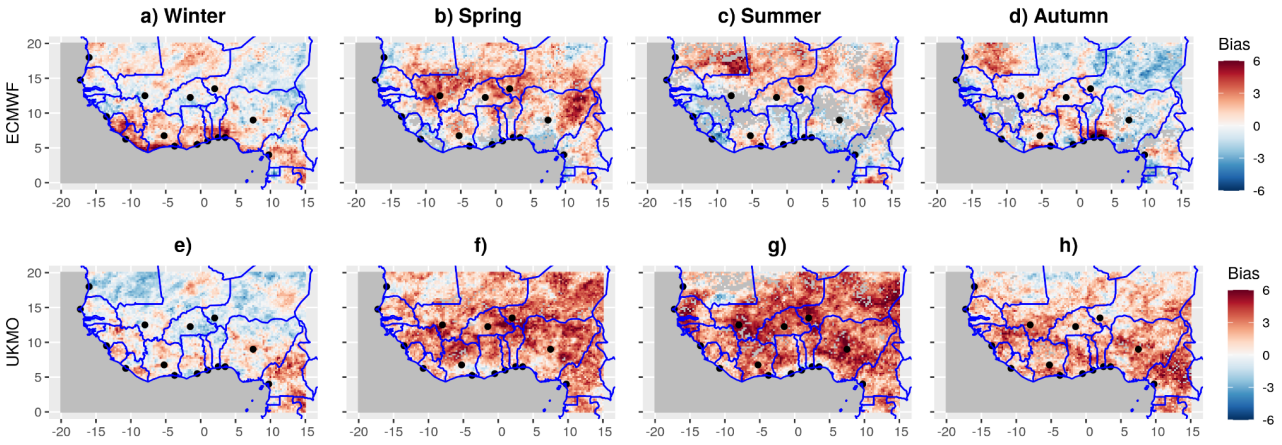


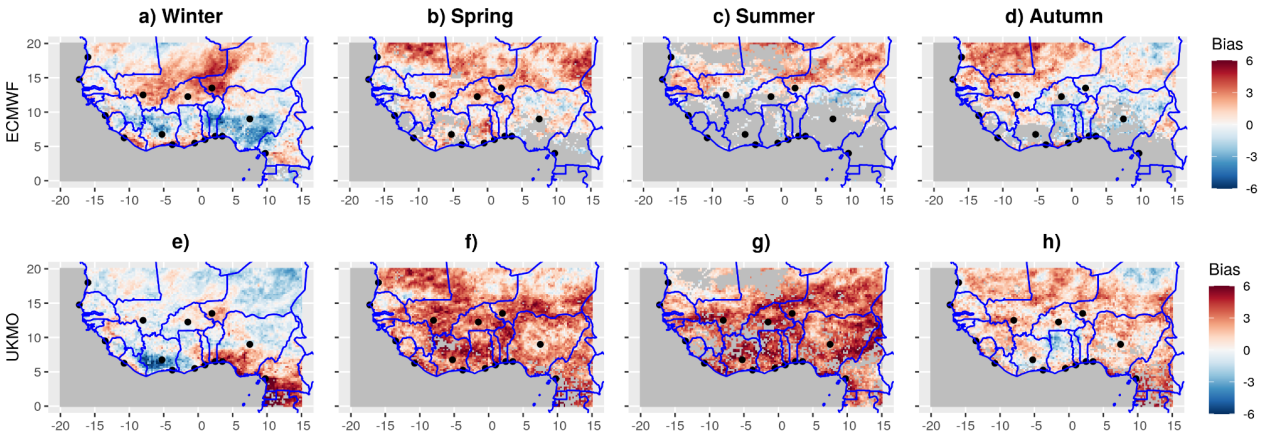
Figure4: Spatial variability of the climatological bias between the forecast models ensemble mean and ERA5 reanalysis over the period 2001-2020 for Tw during the seasons : (a,e) winter; (b,f) spring; (c,g) summer and (d,h) autumn. The bias is computed as the difference between the forecast models and ERA5. The color indicates the bias values in degrees Celsius. The X and Y axes represent the longitude and latitude respectively.

Figures 7, 8: Again, it appears that the colour scales are not covering the same range for the positive and negative ends, and therefore the white colour doesn't represent 0. This can be misleading for the interpretation and should be fixed so that the scale is the same at each end.

We changed figures 7 and 8 according to the reviewer comments.



(i)



(ii)

Figure 7 : Spatial variability of heat wave frequency bias between forecast models and ERA5 over West Africa from 2001 to 2020 for:(i) T2m_min values and (ii) T2m_max values, during: (a,e) winter; (b,f) spring; (c,g) summer and (d,h) autumn. The bias is calculated as the difference in heat wave frequency between the forecast models and ERA5. This analysis is performed using the unperturbed member of the models. The color bar indicates the bias values without units. The X and Y axes represent longitude and latitude respectively. The solid blue lines indicate the borders between countries; the black dots represent the cities of interest for this study (this applies to the rest of the paper).

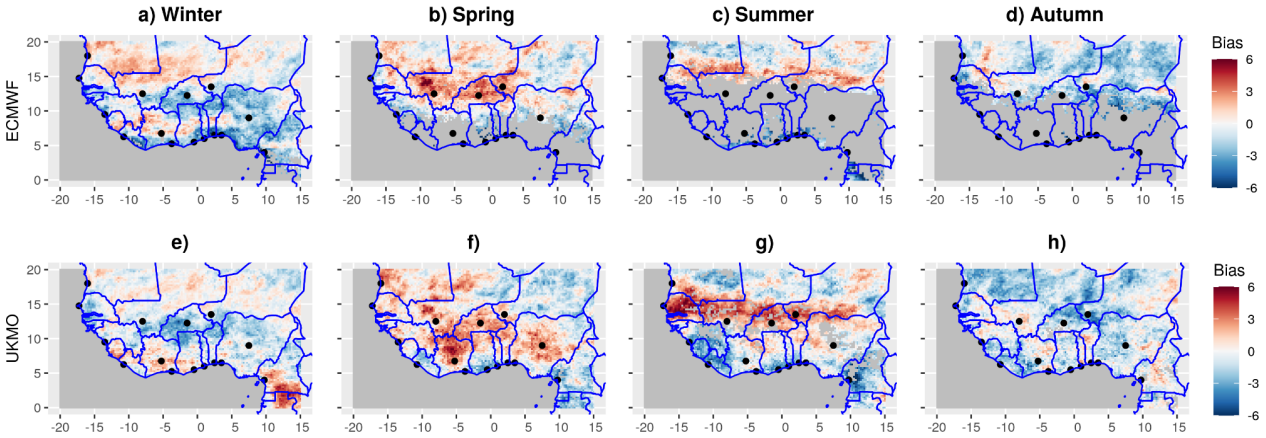


Figure 8 : Spatial variability of heat wave frequency bias between forecast models and ERA5 over West Africa from 2001 to 2020 using Tw during: (a,e) winter; (b,f) spring; (c,g) summer and (d,h) autumn. The bias is calculated as the difference in heat wave frequency between the forecast models and ERA5. This analysis is performed using the unperturbed member of the models. The color bar indicates the bias values without units. The X and Y axes represent longitude and latitude respectively.