NHESS: Subseasonal forecasts of heat wave in west African cities Revisions 2 / review 3

While I have always found the topic of this study interesting and important, and I find this version improved over previous versions, I still have some concerns regarding the methodology and results, and structure of the paper. Some aspects of the methodology and data descriptions are still unclear, and the structure, while improved, is still at times confusing. For example, the skill analysis could do with being split into different parts, each clearly stating the aim, scores chosen, and results, as there seem to be multiple parts to the analysis but it is currently confusing to disentangle them. (e.g. there is one part discussing turning probabilistic forecast to deterministic, and then a description of probabilistic skill scores, with no real clear distinction between the different parts of the analysis and which parts of the results/discussion refer to the different aspects of the study). I have provided some more detailed comments and suggestions below which I hope can be of help to the authors in revising the paper.

Thanks to the reviewer for taking his time to review this work once more. We improved the structure of the document by following the recommendations of the reviewer.

Abstract

The abstract is now very long following the additions. While it is great to see more results highlighted in the abstract, it needs to be reduced in length overall. The authors should also consider what information is clear without the context of the full paper. It is not clear what the 15% / 30% and week 5 / week 2 refer to exactly, so this could be omitted as you really need the context of the full results. For example, I would recommend to rewrite the newly added part (end of line 14 onwards) as follows: “The results suggest that at subseasonal timescales, the forecast models provide a better forecast than climatology, but the hit rate and false alarm rate are sub-optimal and the forecasts may be overestimating the duration of heatwaves, while under-predicting the intensity. Nevertheless, the use of subseasonal forecasts in west African cities can be recommended for prediction of heatwave onset up to two weeks in advance.”

We followed the suggestion of the reviewer.

We replaced : “Heat waves are one of the most dangerous climatic hazards for human and ecosystem health worldwide. Accurate forecasts of these dramatic events are useful for policy makers and climate services to anticipate risks and develop appropriate responses. Subseasonal forecasts are of great importance for actions to mitigate the human and health consequences of extreme heat. In this perspective, the present study addresses the predictability of heat waves at subseasonal time scales in West African cities over the period 2001-2020. The cities were grouped in three climatic regions based on their climate variability: the continental region ’CO’, the Atlantic region ’AT’ and the Guinea region ’GU’. Two types of heat waves were analyzed : dry heat waves using maximum values of 2-meter temperature (T2m_max), and wet heat waves using minimum values of 2-meter temperature (T2m_min) and wet bulb temperature (Tw) respectively. Two models that are part of the subseasonal to seasonal (S2S) forecasting project, namely the European Centre for Medium-Range Weather Forecasts (ECMWF) and the United Kingdom Meteorological Office (UKMO) models, were evaluated using two state-of-the-art reanalysis products, namely the fifth generation ECMWF reanalysis (ERA5) and the Modern-Era Retrospective analysis for Research and Application. The performance of the models in predicting heat waves is assessed through the computation of categorical metrics such as the hit-rate, the Gilbert Skill Score (GSS) and the false alarm ratio (FAR). The forecast models show low skills in predicting heat wave days especially for medium-range forecasts (two weeks ahead) in the three climatic regions. The hit-rate and GSS values are very weak, while the FAR are higher. On average, only 15% / 30% of the predicted heat wave days are actually observed for Week 5 / Week 2, respectively. This suggests that the models overestimate the duration of heat waves with respect to ERA5. Nevertheless, the hit rate and GSS values are superior to the climatology and significant for predicting heat waves in tropical regions, which remains a complex task. Regarding these results, we can recommend the use of subseasonal forecasts in African cities to predict the onset or some days during the heat waves period up to two weeks in advance. Such informations are very useful for the population, hospitals and decision-makers in order to develop some adaptation strategies to reduce the impacts of heat waves in the region." by

"Heat waves are one of the most dangerous climatic hazards for human and ecosystem health worldwide. Accurate forecasts of these events are useful for policy makers and climate services to anticipate risks and develop appropriate responses. Subseasonal forecasts are of great importance for actions to mitigate the human and health consequences of extreme heat. In this perspective, the present study addresses the predictability of heat waves at subseasonal timescales in West African cities over the period 2001-2020. The cities were grouped in three climatic regions based on their climate variability: the continental region 'CO', the Atlantic region 'AT' and the Guinea region 'GU'. Two types of heat waves were analyzed : dry heat waves using 2-meter temperature and wet heat waves using average wet bulb temperature. Two models that are part of the subseasonal to seasonal forecasting project, namely the European Centre for Medium-Range Weather Forecasts (ECMWF) and the United Kingdom Meteorological Office models, were evaluated using two state-of-the-art reanalysis products, namely the fifth generation ECMWF reanalysis and the Modern-Era Retrospective analysis for Research and Application. The performance of the models in predicting heat waves is assessed through the computation of categorical metrics such as the hit-rate, the Gilbert Skill Score and the False Alarm Ratio. The results suggest that at subseasonal timescales, the forecast models provide a better forecast than climatology, but the hit rate and false alarm rate are sub-optimal and the forecasts may be overestimating the duration of heatwaves, while under- predicting the intensity. Nevertheless, the use of subseasonal forecasts in west African cities can be recommended for prediction of heatwave onset up to two weeks in advance."

Introduction

Line 75-76: The authors state that Batte showed that the Meteo-France model can predict heatwaves up to one week in advance, in the previous paragraph. Here, they say that this approach can't provide information about the onset and duration of heatwaves. Why is that? It is not clear – if a forecasting system can predict a heatwave up to one week in advance, it implies that the onset could be predicted up to one week in advance. Did Batte et al show specifically that onset and duration are not well predicted? If so, that should be highlighted. Perhaps the authors wish to highlight that while studies have shown that forecast models have skill in predicting heatwaves in the short to medium-range, such information can be well supported by longer-range forecasts which may be able to provide even earlier information about the potential onset of a heatwave event, allowing more time to prepare, and that is what is studied here.

We clarified this point in the document, we replaced :

"Batté et al. (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." by

" Batté et al. (2018) assessed heat waves predictability using T2m and AT anomalies. While this approach provides information about the evolution of T2m and AT for the subsequent days, it is not sufficient to determine heat waves characteristics such as the duration and frequency. "

Section 2

The description of the ECMWF forecasts is still a bit confusing to identify which data exactly have been used by the authors, as several different datasets are described. Additionally, the model version used by the authors is not the most recent version of the ECMWF model, which may be confusing in parts because the authors refer to 'is running', and that hindcasts are using 'the most recent version'. This should be re-worded to indicate that the study uses forecasts from the IFS cycle 47r3, which was operational from October 2021 to June 2023. (for info, the current forecast version is 48r1, released in June 2023. This version consists of a 51-member ensemble forecast with a horizontal resolution of around 9km out to 15 days, and another 51-member ensemble forecast out to 46 days at 36km horizontal resolution. The version will be upgraded again later in 2024)

We clarified this point in the document, we changed :

"The extended-range ECMWF forecast model is running 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 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. It is no longer possible to carry out a multi-model evaluation and we have therefore limited the evaluation to ECMWF and UKMO. 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/).”

to

“*The extended-range ECMWF forecast model used in this work is run on the Integrated Forecast System (IFS) cycle CY47R3 released on October 10th, 2021. The native spatial resolution of the extended-range 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 137 sigma levels from the surface to 80 km. ECMWF provides two types of outputs for the S2S program: real-time forecasts and reforecasts or "hindcasts". Real-time forecasts are forecasts for the coming days. Hindcasts are forecasts produced for past dates and allow analysis of how the current system would have performed, alongside a consistent dataset covering a longer time period for evaluation. They are useful for the calibration of the model and post treatment analyses.*

The ECMWF extended-range hindcasts are run with 11 members (10 perturbed and 1 unperturbed). 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. We have only analyzed the hindcasts produced on Thursdays, as a preliminary investigation into the initialization dates of the hindcasts showed that most models were launched on Thursdays. In the database, some models use fixed dates of the month (01st, 09th, 17th for example) and others, specific days of the week (Monday, Thursday for example), which generate some difficulties to handle. Most of the models do not cover the period under study, except for UKMO which is available but uses fixed initialisation dates of the month (see section below for more details). It is therefore no longer possible to carry out a multi-model evaluation as we had *planned and we have limited the evaluation to ECMWF and UKMO. The ECWMF hindcasts produced on Monday cover the same period of the ones produced on Thursday. Thus, using the Monday hindcasts we have no significant changes in the frequency of the events detected. 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/).*"

In the UKMO forecast data description, it states that the initialisation dates are not the same as ECMWF, which does not match with your statement in the ECMWF section that only Thursdays are used to be consistent with the other models. I still do not find a valid reason from the authors for only using the Thursday reforecasts of ECMWF, and not also using the Monday forecasts to increase the sample size. Additionally, the UKMO section is confusing because it states that the hindcasts cover 1993 to 2016, and then starts describing forecasts prior to and since 2017. This needs to be rewritten to be more clear. I would recommend to state the full period of time covered by any UKMO hindcasts, and then break it down into the differences in the forecasts from 2001-2016, and 2017-2021.

We clarified this point in the document, we changed :

"The UKMO model runs on the HadGEM3 GC2.0 model which simulates the uncertainties of the initial conditions using a lagged initialisation and the uncertainties of the model using a stochastic scheme. The native spatial resolution of the UKMO model is N216: 0.83°x0.56° (about 60 km at mid-latitudes). It contains 85 vertical levels from the surface to 85 km and 4 soil levels: level 1 (0 - 0.1 m), level 2 (0.1 - 0.35 m), level 3 (0.35 - 1 m) and level 4 (1- 3 m). Similar to ECMWF, UKMO provides to the S2S program real-time forecasts and hindcasts. 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 1 st , 9 th , 17 th and 25 th , 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 25 th , 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). Our target period is going from January 2001 to February 2021, and as mentioned earlier, the UKMO hindcasts are not available after the year 2016. To solve this problem and get more robust statistical results, we recompose the products to obtain a new composite that covers the whole target period. "

to

“*The UKMO model runs on the HadGEM3 GC2.0 model which simulates the uncertainties of the initial conditions using a lagged initialisation and the uncertainties of the model using a stochastic scheme. The native spatial resolution of the UKMO model is N216: 0.83°x0.56° (about 60 km at mid-latitudes), but the downloaded data are extrapolated to a regular latitude/ longitude grid of 0.25° x 0.25°. It contains 85 vertical levels from the surface to 85 km and 4 soil levels: level 1 (0 - 0.1 m), level 2 (0.1 - 0.35 m), level 3 (0.35 - 1 m) and level 4 (1- 3 m). Similar to ECMWF, UKMO provides to the S2S program real-time forecasts and hindcasts. 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 analyzed here are run using the model version released in 2023 which produces 7 members per cycle (6 perturbed and 1 control) for the period 1993-2016 (no UKMO hindcasts available after 2016). They are produced 4 times per month, on the 1 st , 9 th , 17 th and 25 th. We are aware that these initialisation 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. Our target period is going from January 2001 to February 2021, and as mentioned earlier, the UKMO hindcasts are not available after the year 2016. To solve this problem and get more robust statistical results, we recomposed the products to obtain a new composite that covers the whole target period. ”*

Line 203: The authors may wish to state that the land-sea mask of ERA5 has values on a scale of 0 to 1, and what this means, since it is not necessarily standard practice (many masks provide a binary 0 or 1, so it should be clarified what setting a threshold of 0.5 actually means, because it may not be obvious to the reader if they don’t work with the ERA5 data and mask themselves)

We clarified this point in the text, by adding the following in section 2.4.1 :

“ Following the same approach as in Ngoungue Langue et al. (2023), local temperatures over the cities were derived from the reanalysis using the reanalysis grid point closest to the station that satisfies a land-sea mask (lsm) of at least 0.5 ([Table2] shows the lsm values for all the cities considered in this study, the same technique was applied for the forecast models). The lsm indicates the proportion of land contained in a grid point. If the lsm is less than 0.5, this means that the grid point is mainly covered by the ocean, while a lsm greater than 0.5 implies more land coverage. ”

Heat wave detection – is it a regional characteristic, and documented, that humidity is typically lower during the day? Is the humidity always dry enough that it doesn’t have an impact on heat stress during the day? This should be clarified, because it is not the case everywhere – typically, it is important also to consider the humidity during the day in combination with the peak temperatures, in order to truly capture the risk of heat stress. The authors may also want to comment on the impact of both together – it can read as though dry and wet heat waves are two different types of event, but really they are just the daytime and nighttime components of one heatwave event, they are not completely distinct, the authors are just disregarding humidity during the day. The authors may wish to comment on the importance of nighttime temperatures and humidity following extreme daytime temperatures, due to the chance (or lack of) for recovery overnight.

We clarified this point in the text, we changed:

"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 (Steadman, 1979a, b). 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." by

"In the present study, two types of heat waves are investigated : dry and wet heat waves. Dry heat waves are associated with high temperatures and low humidity conditions. The detection of dry heat waves is processed using 2-meter temperature (T2m) as an indicator. We distinguished two categories of dry heat waves : those that occur during the daytime and are detected using maximum values of T2m (T2m_max), and those that occur during the night and are detected using minimum values of T2m (T2m_min). Concomitant heat waves, those that occur during daytime and the night, are extremely dangerous because the body does not have the time to recover from the daytime heat waves during the night (Li et al.,2017; Wang et al., 2020). The most lethal heat waves are due not only to high temperatures but also to the effect of humidity (Steadman,1979a, b ; Heo et al.,2019 ; Yu et al., 2021). Humidity is an important driver of wet heat waves. The combination of high heat and humidity can compromise the human body's main cooling mechanism: transpiration. The evaporation of sweat from skin cools our bodies, but higher humidity levels limit evaporative cooling. As a result, we can suffer heat stress and illness, and the consequences can even be fatal. Wet heat waves are detected here using average Tw as an indicator (Yu et al. 2021)".

What is the reference temperature in the relative humidity formula?
We added this information in the text : "the reference temperature is the thermodynamic temperature of the triple point of water"

Line 240: "The 90th percentile appears to be a sufficient threshold for monitoring heat waves affecting human health." How did the authors determine this? Can the authors cite the literature where this is determined? Other studies I have seen use the 95th percentile as a relevant threshold for human health.
We added some references supporting the choice the 90th percentile for the monitoring of heat waves affecting human health : " Perkins and Alexander 2003 ; Déqué et al., 2007; Fontaine et al., 2013; Russo et al., 2014 ; Oueslati et al., 2017; Barbier et al.,2018; Lavaysse et al., 2019 "

Line 251 – 253: is ‘hot day’ an accurate description, when T2m_min deals with nighttime temperatures? Or is a ‘hot day’ one where all three of the variables are all exceeding their respective 90th percentiles during a 24-hour period?

We aware that this can lead to confusion, but to simplify text, we have defined it as follows :

“Hot days are days belonging to heat waves with the values of one of the 3 indicators (T2m_min, T2m_max, Tw) above their 90th percentile. ”

Lines 256 – 264: I am completely confused by this paragraph, sorry – I don’t understand any aspects surrounding the definition of heatwave intensity or how it relates to other aspects. This paragraph needs to be rewritten, but I’m afraid I don’t understand well enough to offer any advice on it.

Thank you for your comment and sorry for the confusion. We clarified this point by changing :

“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. We would like to point out that the threshold values used to compute heat waves intensity are not the same for heat waves detection. In the scope of this study, part 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), we are interested in heat waves, which can be harmful to human health. As mentionned previously and based on previous works, the 90 th percentile is a suitable threshold to detect heat waves affecting human health. In order to assess the severity of the events using the same reference, we have chosen a constant daily threshold for the computation of heat waves intensity. It is defined as the minimum of the daily climatology 90 th percentile over the study period. Therefore, most dangerous heat waves will have higher intensity values.”

to

“The intensity of a heat wave was defined as the sum of the daily exceedances of the indicators (T2m_min, T2m_max and Tw) values to a constant threshold for the duration of the event. It should be emphasized that this constant threshold is not used for heat waves detection, but only to compute their intensity. In this study, which is part 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), we are interested in heat waves, which can be harmful to human health. Therefore, the constant threshold mentioned above is defined as the minimum of the daily climatological 90th percentile over the study period. The choice of a constant threshold for the computation of heat waves intensity is very important because it takes into account the seasonal cycle. This makes it possible to assess the severity of the events using the same reference. In fact, the most dangerous heat waves will have higher intensity values.“

Line 279: “the probability of a member predicting a day as being part of a heatwave” – one ensemble member cannot have a probability attached, as the probability itself comes from the full set of ensemble members; I’m not sure what the authors are trying to say here.
We clarified this point in the text by replacing :

“ Given a threshold of 20% (i.e. at least 20% of ensemble members predict heat wave days), our methodology consists of computing the probability of a member predicting a day as being part of a heat wave. If the probability is greater than the threshold value, the day will be counted as "1" and "0" otherwise” by

“Assuming a threshold of 20% for example (i.e. that at least 20% of ensemble system members are forecasting heatwave days), we determined the probability of the ensemble forecast systems predicting a day as being part of a heatwave. If the probability is greater than the threshold value, the day will be counted as "1" and "0" otherwise.”

Line 278: why is the threshold of 20% used? The authors state it is inspired by another paper, but a prediction of just a 20% probability of a heatwave could be seen as quite low. It would be important to provide a reason for why 20% is decided to be the optimal threshold and how a different threshold may impact the study.

We clarified it in the previous comment (“ line 279”), at this stage in the study the “20%” threshold is an example of threshold value to explain the methodology. Nevertheless, in the document, we provided some analyses obtained using different threshold values.

Lines 283 – 289: what is “the whole season” – what part of the year?
We replaced the “whole season” by “ from January to December”

Section 2.4.5: it is confusing to read in the previous section that the authors turn the probabilistic forecast into a deterministic one, because it’s too complex to assess probabilistic forecasts, and then immediately the next section introduces probabilistic skill scores.

We reorganized the sections 2.4.4 and 2.4.5 in the document as follows :

“2.4.4 Evaluation of probabilistic forecasts

2.4.5 From probabilistic to deterministic forecasts

2.4.6 Evaluation of deterministic forecasts
”

Results
Line 413-414: “the main differences… are found with ECMWF” – what are the differences between T2m and Tw that are found?

We added these differences in the text :
"In winter, for example, there is an underestimation of the duration of heat waves associated with Tw in the ECMWF over the Guinea region, while an overestimation is observed with T2m_min and T2m_max."

Discussion
Lines 499-503: While the UKMO may place special emphasis on data relevant to the UK and surrounding regions, it does also assimilate a wide range of global and regional data from satellites, ground-based measurements etc. It is unclear how this would influence the surface-ocean interactions. Do the authors mean to discuss differences in ocean coupling between ECMWF and UKMO?
We clarified this point in the text, we changed :

"Secondly, the models used the same data assimilation methods (4D-Var) for control analyses but the data and initial conditions are completely different. 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. These differences may influence the representation of surface-ocean interactions in the models. The differences observed in the representation of T2m_min over the Atlantic ocean [Fig.3] can result from these types configurations." to

"Secondly, the models use the same data assimilation methods (4D-Var) for control analyses, but the data and initial conditions are completely different. 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 atmospheric model is coupling with the NEMO3.4.1 ocean model, while UKMO uses the NEMO3.6 ocean model. The two systems are not using the same atmospheric and ocean models, which implies different parameterisations. The differences observed in the representation of T2m_min over the Atlantic ocean [Fig.3] could result from representation of surface-ocean interactions in the models."

Line 504-505: Is this the correct resolutions for the forecasts that the authors have analysed in this study? The ECMWF extended range is 36km spatial resolution.
We added the correct resolution in the text :
"Thirdly, the spatial resolution of the atmospheric component of the two models : ECMWF has a higher spatial resolution than UKMO (0.32°x0.32° Vs 0.83°x0.56°), which means that it can capture local-scale variability or atmospheric processes and provide more accurate forecasts for specific regions. (https://confluence.ecmwf.int/display/S2S/ECMWF+Model)"

Conclusions
Lines 551-554: If these models can be useful only up to two weeks in advance, can the authors comment on why they should be considered for use in heatwave forecasting, if other studies have shown that medium-range forecasts are also accurate up to 2 weeks ahead for forecasting

heatwaves? If choosing between a 2-week medium-range forecast at high resolution, and a lower-resolution subseasonal forecasts up to 2 weeks ahead, why choose the subseasonal forecast?

In this study, we evaluated the skills of the subseasonal forecasts with a focus on medium (week2) and long (week5) time scales. In general, we found that the subseasonal models perform better at week2 than at week5. This means that sub-seasonal forecasts are able to predict heat waves at least two weeks in advance. Subseasonal models offer a wide forecasting window and make it possible to detect an event long time in advance compared to medium range forecasts. We clarified this point in the text by adding the following :

"Regarding these results, we can recommend the use of subseasonal forecasts in African cities to predict the onset and frequency of heat waves, and some days during the heat waves period at least to two weeks in advance, but as far as their intensity is concerned, it is still challenging."

Figures

Figure 4: the colour scale used in this figure is inappropriate for the data for 2 reasons: (1) the scale shows only negative values from 0 to -14, it is therefore not appropriate to use a diverging colourscale with a centre point at -7, it should use a continuous colour scale, not diverging. (2) the colour scale includes both green and red together, which applying a filter for colour blindness accessibility indicates it is not accessible for those with colour vision deficiencies. Also, are there really not any positive bias values across the entire region, the bias is only negative everywhere? I would find this surprising. I couldn't find any discussion of the results.
We changed the color scale according to the reviewer comment.

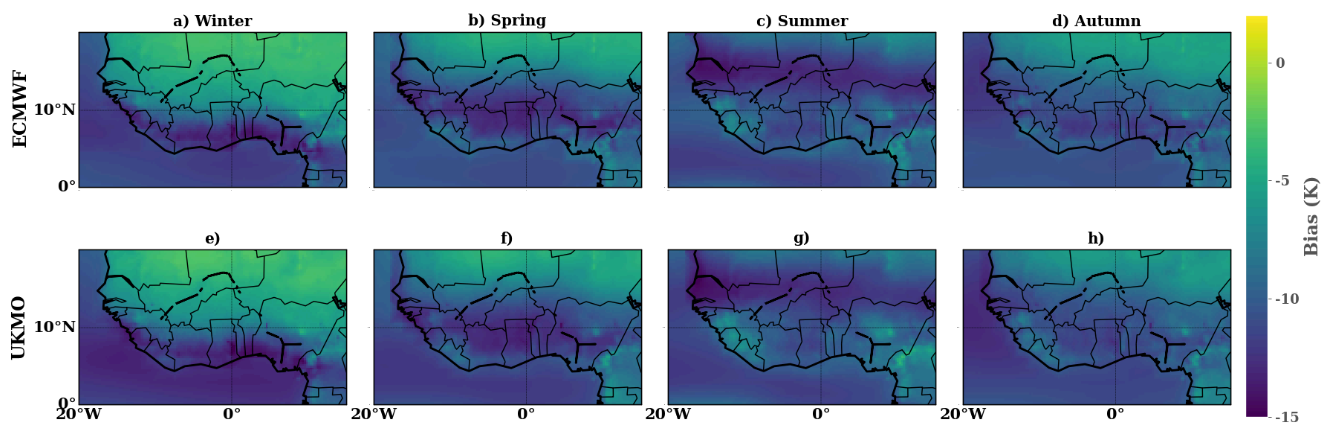

Figure 4 : Spatial variability of the climatological bias between the forecast models ensemble mean and ERA5 reanalysis over the period 2001-2020 for Tw during the seasons : (a,e) winter; (b,f) spring; (c,g) summer and (d,h) autumn. The bias is computed as the difference between the forecast models and ERA5 considering all the lead times. The color indicates the bias values in degrees Kelvin. The X and Y axes represent the longitude and latitude respectively.

Minor comments and typos

Abstract, line 2: remove ‘dramatic’ (not every heatwave is dramatic, and it is an odd choice of adjective for a heatwave)
We removed “dramatic” in the text

Introduction
line 29: change “due to urban heat islands” to “due to the urban heat island effect.”
We followed the recommendation of the reviewer.

line 30: change “by several heat extreme events” to “by several extreme heat events”
We followed the recommendation of the reviewer.

line 31: remove “in the shade” (assuming this is based on an official temperature reading, any such reading should always be in the shade)
We followed the recommendation of the reviewer.

lines 45-46: the authors mention that EWS integrate shorter and medium-range forecasts, then say ‘this window refers to subseasonal timescale from 2 up to 6 weeks’ – but short to medium-range timescales would be defined as up to 2 weeks, and then subseasonal as 2 to 6 weeks. Please rewrite these sentences, for example “Many early warning systems integrate short and medium-range forecasts of potential weather hazards up to two weeks ahead. The subseasonal forecast range, from 2 to 6 weeks ahead, is also highly relevant for actions aimed at mitigating the consequences of extreme heat”
Thanks to the reviewer for this suggestion, we integrated it in the text.

Line 52: “wet/dry temperatures” is not a typical phrase and is confusing. Recommend changing to “Heat waves are often associated with extreme heat, which can be exacerbated by other factors such as humidity levels.”
We followed the recommendation of the reviewer.

Line 61: “It refers to…” should be “Heat stress indices are used to combine relevant atmospheric variables (such as temperature, humidity, solar and thermal radiation, wind speed) to indicate the impact of the environment on the human body. Examples include…”
Thanks to the reviewer for this suggestion, we integrated it in the text.

Line 64: Universal Thermal Comfort Index should be “University Thermal Climate Index”
We followed the recommendation of the reviewer.

Line 71: “ECMWF extended long-range forecasting system” should be “ECMWF extended-range forecasting system”
We followed the recommendation of the reviewer.

Line 74: wet and dry heatwaves are introduced here, but are not defined yet. Consider adding short definitions in brackets e.g. "wet heat waves (those combined with high humidity levels) are more predictable than dry heatwaves (those combined with low humidity levels)."
We clarified this point in the document by adding these short definitions in brackets.

Line 81: add "subseasonal" before predictability to clarify to the reader the forecast range which you will evaluate
We followed the recommendation of the reviewer.

Line 108: "reanalysis data have a high resolution compared to observations" – this is an odd statement, as it is strange to compare a gridded product resolution with the observation network, if sparse. There is also something incorrect here – reanalysis products have too low of a resolution to be able to detect the highest temperatures (potentially) at point locations, and the urban heat island effect.
We clarified this point in the text, we changed :

"We are aware that reanalysis data have a high resolution compared to local observations and therefore cannot 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." by

"We are aware that reanalysis data have a low resolution for detecting the highest temperatures at a point location, as well as the urban heat island effect."