Review of Buzan et al. “Implementation and comparison of a suite of heat stress metrics within the Community Land Model version 4.5”

This paper describes an implementation of several heat-stress metrics into the land component of a climate model. This implementation could be very useful and so the effort is applauded, because it is not possible to reconstruct peak (or average) heat stress from standard model outputs such as peak temperature or from daily-mean meteorological quantities. The description and implementation is detailed which is useful, although the paper is wordy at times and its 50-page length should be shortened if possible without losing key information. One thing that would help is to get rid of the Appendix (see point 42 below), and I recommend eliminating Figure 1 (point 44) and redoing or reconsidering nearly all the other figures.

I find that there are many problems with the manuscript which will require major revisions before it is acceptable for publication. The principle problems are:

1. The figures are unintelligible and require better explanations in the captions and text. Ideally one should be able to understand the figures in a paper just based on reading the captions, but that would certainly not be possible here. Even after reading the text I was often unable to figure out what exactly they were showing. The description of the statistics computed is inadequate and/or incoherent.
2. A bigger point around the figures and indeed the results in general, is that I question whether the authors have chosen the right things to show. Their main point (if I have understood correctly what they have done, which might not be the case) seems to be that in humid regions, heat stress thresholds are reached at lower temperatures and heat stress is less variable than in arid regions. These points are obvious (and have been quantified in several of the papers cited), and you don’t need a GCM, let alone online diagnostics, to show it. A secondary finding is that formulas used in GCMs for things like saturation water vapor pressure are not perfectly accurate, but i do not believe this point is significant (see 6 below).
   What I really want to see in this paper is a demonstration of the value added by what the authors have done. How do online, time-step level metrics improve on what we could have done with archived model output? If we had daily or 3-hourly model output would that be enough to get the same results? Are the results they have shown, also evident in daily or even monthly mean fields? How independent are the different metrics, are there some that can be discarded as having no additional information content beyond others (some information is given on this but with very little discussion)? Or, another interesting question would be how well does the model do at reproducing observed values of these metrics? This could be contrasted with its skill in quantities more likely to be looked at such as peak temperature, daily and monthly means. Do these diagnostics reveal previously hidden problems?
3. I think the abstract is too long, has too many details (e.g. names of parts of the model code) and does not adequately explain the motivation and objectives of the study. The Summary section, 6, is a better abstract (in some ways anyway) than what the authors have written for the abstract.
4. I am discouraged that the authors (apparently) did not incorporate heat-stress metrics that account for wind and radiation variations, since this seems to be one place where they can truly innovate by assessing how overall patterns of heat stress are (or not) altered by taking these into account. Currently every heat stress study has to basically hand-wave that these factors aren’t important—this study could make a real contribution by finally tackling the issue, since all the needed inputs are right there in the model.
5. There are numerous instances, which will appear in the detailed comments, where the authors have used terminology inaccurately or confusingly.
6. The paper implies in a very misleading calculation (lower panels of Fig. 1) that there are significant problems with the way saturation vapor pressure is calculated in GCMs, when the errors are tiny and well-documented in the existing literature.

7. The writing of the paper is quite poor, with badly organized paragraph structure, and many unclear sentences, which makes it hard to read. I have not tried to point all of these out individually but would request that the authors read carefully and try to find ways of rearranging or rewriting text that is unclear or jumps from one topic to another within a paragraph. A good rule of thumb is that everything said in a paragraph should usually relate back to and expand on the opening sentence of the paragraph.

Minor comments.

1. **5198**,11 etc. The word "covariance" is misused; this specifically means <uv> averaged over an ensemble of joint measurements {u,v}. I think "covariation" is what they are after.
2. 15-16. This statement is too vague; in principle one can calculate T_W from r and T which are prognostic variables in the model.
3. 28-end: quite confusing to first assert two regimes (arid and non-arid) and then begin talking about others "strong convection". Also confusing to have "strong convection" and "equatorial" as separate regimes (isn't there strong convection near the equator?)
4. **5199**,2 "Heat death" -> "Heat related conditions" ?
5. **5200**,7. It is implied that these studies do not consider the diurnal cycle but in some cases they did look at daytime high and nighttime values, even if not characterizing the full diurnal cycle.
6. 12: The text should be clearer about what type of "inaccuracies" are being considered here. Are the authors claiming that inaccurate formulas for well-defined quantities such as dewpoint or wet-bulb temperature have been carelessly applied? Or are they referring to heat stress measures such as sWBGT that are widely-used approximations to other measures? It is not clear from the papers listed as offenders here; for example Dunne et al. used a standard WBGT approximation, while Sherwood and Huber 2010 considered only T_W, for which they used the same Davies-Jones formulas used in this paper unless I am missing something (since Huber is an author on both papers this at least should be cleared up!). This distinction matters because even if sWBGT departs significantly from WBGT, it is still a well-defined index that is no more or less relevant a priori than most of the other indices (AT, HI etc.). WBGT itself is based on a very approximate physical analog of the human body, and is not a standard of truth.
7. 23-end: The goals of the study are not clear (problem also in Abstract). If the goal is to compare a large suite of metrics, why do we need a land surface model? Why not just use station observations? If the objective is to make a bunch of new metrics available in a GCM then that should be more clearly stated, and the reasons why this is a good thing should be more clearly explained.
8. **5201**,7: Do you need a citation here? Also, is all the discussion of biomes etc. necessary for interpreting the results of your study? It seems tangential.
9. **5202**,6: Subtitle is strange—these are not water calculations but moist thermodynamic quantities. We need to know how much water is in the air already before we can do these calculations. Also, I think it would make more sense to move this discussion together with the current section 3.1 where all is revisited and much is restated.
10. 10-19: This text is confusing two issues. One is the accurate calculation of the saturation humidity, and the other is how to calculate T_W efficiently. Please write more clearly. GCMs may use imperfect approximations to do the former; do they calculate the latter at all? I do not believe there is a case for updating calculations of saturation, unless you can show that the errors are physically significant. People have been well aware of this situation for years and compared to other modeling uncertainties like boundary layer and cloud parameterization, the small errors in approximate formulas for e_s or r_s seem trivial.
11. 20-21. This sentence is confusing because Bolton (1980) did not present a T_W calculation. The relation between T_W and theta_e needs to be explained.
12. **5203**,5: by "calculating" do you mean "implementing"?

13. 12: unrealistic compared to what?
14. 22: Don't some of them use radiation and wind speed? Ok, a few lines later you mention this, but the wording implies that you won't look at the ones that use wind speed. Why not? Isn't that a strength of your approach that you can do this?
15. **5204**, 4: I do not understand how winds can be implicitly included. Either the metric explicitly includes wind speed, or is calculated assuming some average or typical wind speed, or inherently does not depend on wind speed (e.g. T_W). Implicitly including it would mean that it is included indirectly because it correlates with some other variable that is included.
16. **5205**, eq (1-2). Why e_RH and e_sPa? What does RH stand for? The normal notation is to simply use e for the vapor pressure and e_s or e* for the saturation vapor pressure (I see e_s in Table 2, is e_sPA a misprint?). I don't see any other e so there is no need for a subscript except to distinguish actual and saturation.
17. eq (3): citation needed (not clear if this polynomial fit is from Steadman but as written it implies not). Since Fahrenheit was a person, the subscript should be F not f (consistent with C for Celcius)
18. **5206**,13: What warning system? Citation needed.
19. 17: why do you use the calibration for pigs? Does it make much difference?
20. **5207**,12: waht is "radiation temperature"? And why wasn't wind incorporated as an input?
21. 13: in what sense are these required? You just said the inputs were T, T_W, and "radiation temperature" so aren't those what would be required to calculate the index value? Do you mean sweat rates etc. were used to develop the index?
22. 14: In the introduction you implied that the failure to account for wind and radiation was an important shortcoming of past studies, but now you are saying you aren't interested in those either. More discussion is needed, in the context of the rationale for the study. I think to be honest, if you don't intend to deal with metrics that account for these factors, then in the introduction where you raise this issue you should say up front that this study also will avoid them (currently one will guess the opposite). Finally, this paragraph should state more clearly that the authors are *not* using UTCI. This is a big decision, since the UTCI is arguably the most sophisticated index and was designed to be incorporated into models such as this one (as I understand it).
23. **2508**,14: This curious statement requires elaboration. You mean there are three different ways to construct a natural wet bulb that yield different temperatures? That a natural wet bulb exhibits hysteresis and doesn't have a unique equilibrium temperature? That there are three different equations for predicting the natural wet bulb temperature and we don't know which one to use? What does it mean for a metric to have "multiple end members"? Confused.
24. **5209**,7: you mean its accuracy in reproducing WBGT may be questionable. Not just may be, it is guaranteed not to except in particular conditions, since it ignores factors that affect Tg. But it may be OK for diagnosing the effect of a change in T or humidity on human comfort (other things not changing much).
25. **5211**,10-19: Here the authors give some important information on objectives and rationale, that should have been given much earlier. One thing they should make clear is why they are putting these metrics into the land model, rather than the atmosphere model (I know the answer, but readers who are not familiar with GCM construction may be puzzled).
26. 11: you mean the joint distribution of T, P, and Q conditional on high value of metric X?
27. 12: "hottest" means "hottest according to metric X" (not T)?
28. 18: by "of the percentiles" you mean, of the extremes in metric X?
29. 20: "median of the joint distribution" is a non-sequiter, joint distributions do not have medians. What do you actually mean here? Do you mean the median value of X for all points in the top 99%?

30. **5215**,10: These maps do not present joint distributions - a joint distribution is the probability density of a multidimensional state vector. All that is presented here is a single statistic of the distribution (at each location). We are not told what statistic, so I do not know what these maps actually show. I also don't know what "metric B given metric A" means - does this mean the value of B conditioned on a globally fixed value of A according to a fit (say, multivariate linear or Gaussian) to the sample joint distribution, or a subsampling of all values within some tolerance of A? If so then what value of A is used? There is way too little information here, and what information is provided doesn't make sense.
31. 16: the plural of maximum is maxima
32. **5216**,19: These four categorizations are not on all fours. Equatorial regions are convective, some arid regions are found in mid-latitudes. I cannot make sense out of this classification.
33. Figure captions. The captions of Figures 2-7 each repeat the same unimportant information "1901–2010 CLM4.5 forced by CRUNCEP" - this only needs to be stated in the text (once). But the captions do *not* tell us what is plotted except "joint distribution" (which is incorrect). Please tell us enough so we can figure out how to read the plot. Also, these plots are a bit small and hard to read.
34. Figs. 4-7: These are getting closer to being actual joint distributions but not quite - they are conditional distributions of T for various values of each metric X and "regional association" (try to be consistent between the terminology "regime", "category" and "regional association"). They are not joint distributions because they don't show the distribution of X (except its limits), only the conditional distributions of T given X.
    Why are there many points on the bottom axis?
35. **5218**,4-5: Shouldn't the criterion for saying these are unreasonable be because they disagree with observations (how do we know a priori what T_W should be there)? Why don't you look at some station data or HadCRUH to see what observed humidity is there? Also, does this problem originate in unrealistic CRUNCEP fields or poor behavior of the land model? The first "reason" given for the error does not make sense, since over bare ground any flux at the surface will match that at 2m as long as you are averaging over more than a few minutes (little water vapor or heat can accumulate in 2 meters of air). The second reason should be checked by looking at the distribution of soil moisture values; on its face, it also seems an unlikely explanation since in arid regions there should rarely be rainfall so this problem presumably would not occur very often? And there is no "sand parameterization", there is a soil parameterization that assumes some physical characteristics for sand.
36. **5219**,1: which features?
37. 2-3: This paper has not shown that implementing their metrics reduces uncertainty in anything, let alone justifying such a sweeping statement.
38. 4-6: I don't understand what this sentence means, and don't recall T_E being defined.
39. 18-22: these statements require qualification if they apply to the top 1% of events. Indices that are similar at this extreme mibht be different at lower temperatures.
40. 22-30: the paper swerves into reopening the discussion of assumptions in the metrics. That should all be done earlier, unless these assumptions are key to interpreting the results or future uses of the software package presented (if so that isn't clear at all from what is written).
41. **5220**,1-8: A limitation of the approach the authors use is that many heat stress impacts seem to depend on multi-day exposure duration. Come to think of it, is the authors' plan to write out all their indices multiple times per day? If so then they don't seem to offer any advantage over just writing out T, r, and p and calculating from output. If not then what summary quantities would one output? Peak heat stress during the month, for monthly mean output, for example? Or number of days above a few heat-stress thresholds? There are more issues that must be confronted before these metrics will have practical value, it seems.
42. The Appendix, as far as I can tell, just goes through the contents of the Davies-Jones (2008) paper. What is the point of this? Why not just cite Davies-Jones. If there are key formulas that need to be invoked in the text, put those where they are invoked.

43. Table 3. This caption needs more information - what does “modern” and “future” mean? Are you just describing past work here or does this refer to your calculations? Some of those studies may have used more than what is listed (e.g. Sherwood and Huber used reanalysis data not just CCSM3).

44. Figure 1. The variable “q” should I guess be “Q”, the specific humidity? The plot for q is un-useful and indeed misleading because it implies a very large error in q which is not true. What is actually happening is that you the computation is being done at fixed total pressure p, but as e approaches p this becomes impossible and implies a vanishing (and then negative) dry air pressure. This is not sensible. If the calculation is done at fixed dry air pressure (more sensible since this is what would actually happen with a fixed mass of dry air and g), the curve for q will look similar to that for e. I recommend deleting the figure entirely and dropping all claims or innuendo in the paper about the inaccuracy of saturation algorithms—you are beating a dead horse, these small errors are already documented in the literature, and there is no way that errors of no more than 2% that don’t begin to appear until temperatures are 30C higher than any on Earth today are of any significance.