Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2022-223-RC3, 2023

Review of gmd-2022-223

Matthew Kasoar (Referee)

Referee comment on "Evaluation of CMIP6 model performances in simulating fire weather spatiotemporal variability on global and regional scales" by Carolina Gallo et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-223-RC3, 2023

The authors present an evaluation of historical fire weather performance by 16 CMIP6 models, comparing how well they reproduce ERA5 reanalysis estimates of the various components of the Canadian Forest Fire Weather Index System (the most widely-used set of fire weather danger metrics).

Such an evaluation is valuable and timely given the obvious application of using latest-generation climate models to study changing fire weather severity and frequency. The authors provide an indication of the current CMIP6 ensemble performance at capturing the mean, 90th percentile, and seasonal cycles of fire weather both globally and across various fire-prone regions, and also give an indication of which models (of the 16 studied) tend to provide the best (and worst) performances – a useful guide for many studies that often rely on single-model results. While it's a shame that some fairly prominent climate models (e.g. CESM2, HadGEM3/UKESM1, GISS-E2, EC-Earth3, NorESM, MIROC) are not included in the current analysis (due to the required variables not being available at the time the analysis was done), nonetheless it also provides a methodology that can be extended to validate FWI performance of further models given a small set of fairly standard output variables.

The manuscript is well written with a concise and very readable prose style, and is a well suited study for this journal. I have a number of mostly minor comments which I have listed below; mainly these are just requesting clarification on certain details. I also feel that the manuscript could be even more useful if the discussion touched a little more upon the drivers of model bias rather than being purely descriptive, as detailed in my second comment below. But the analysis seems sound (subject to my first comment below being addressed), and if the authors can provide the additional clarifications as detailed in the subsequent comments then I would certainly endorse it's publication in GMD.

Slightly more major comments:

- My main worry surrounding the methodology, is that the FWI indices for the CMIP6 models are calculated slightly differently from how they are derived in the GEFF-ERA5 reanalysis product. As the authors note, the FWI indices are ideally supposed to be calculated with local noon values of temperature, RH, wind speed, and accumulated precip, and these are what GEFF-ERA5 uses. However, local noon snapshots are not typically archived in CMIP6, and so the authors instead use daily maximum temperature, daily minimum RH, and daily mean wind speed and total precip, as proxies for local noon conditions. While these are necessary and reasonable approximations which previous studies have similarly used when working with climate model data, nonetheless it means it's not quite a like-for-like comparison when comparing the estimated CMIP6 FWI indices with the exact GEFF-ERA5 values. It's therefore important to verify first that this difference in calculation method doesn't make any difference to the resulting bias patterns. A priori, it seems plausible for instance that daily maximum temperatures would often be slightly higher than local noon values, which could result in the CMIP6 indices being positively biased on average simply due to the differing calculation methods. The standard ERA5 atmospheric reanalysis should have all the same daily max/min/mean variables that are used from the CMIP6 models, and therefore the authors should be able to calculate the FWI indices from scratch for ERA5 using the same method and approximations as for the CMIP6 models. This could then be differenced with the exact GEFF-ERA5 values to check what (if any) difference the calculation method makes. Provided the difference due to calculation method is negligible compared to the bias patterns then there's no problem with keeping the rest of the analysis as is, but this should be verified first.

- In Sections 4 and 5 (Synthesis/Discussion and Conclusions), it would add further value if the authors could comment a bit on what are the main meteorological drivers of good and bad model performance and inter-model spread. At the moment the paper principally illustrates the models' biases without any substantive discussion of what drives them. This is certainly still invaluable for end users of these models 'off the shelf' to study fire weather, while also providing a methodology to validate model FWI performance which can be extended to other models, and the authors are clear that this is the primary aim of the paper. But from a model development perspective, it would be very useful to also have some pointers towards what are the critical things that models need to get right to be able to simulate fire weather well.

  There is some brief discussion of structural differences between models, and the authors note for instance that model resolution doesn't seem to correlate with performance. But the models don't simulate FWI directly; they simulate meteorology, and it feels like it should be possible to say something about which meteorological factors (and/or regions) are the ones that model developers should be concentrating on to try and improve the representation of fire weather. A thorough exploration of this is no doubt beyond the scope of the current paper, as it could be a whole separate

analysis in itself. But it would be great if the authors could comment a bit on some of the broader patterns. For instance, from Figure 4 we see that the MMM consistently does badly in certain tropical regions like NHSA, SEAS, and EQAS. Is this because all the models consistently struggle to represent a certain driving variable in these regions, e.g. maybe they all tend to underestimate tropical precipitation? Or are the models all bad for different reasons in these locations?

In terms of inter-model spread, I also found it very intriguing that (L376-377) "strong model performance for one indicator does not necessarily mean strong performance for another", and some of the fire weather indices (FFMC, ISI, FWI, and DSR) are more consistently well-simulated than others, even though those other indices are largely calculated from the same meteorological variables as the ones that are well-simulated. I assume this can only be because the relative influence of the various meteorological variables is different for different indices. But therefore, it again seems like it should be possible to say something broadly about which meteorological variables are more responsible for driving inter-model spread in performance; e.g. which are the driving variables that are relatively more important in those indices which tend to be poorly simulated, which can explain why those variables are often less well simulated than others? N.B. it's not quite the same thing, but as a tangential example the authors could look at this paper: Grillakis et al. ERL 2022, https://doi.org/10.1088/1748-9326/ac5fa1. In Figure 4 of that paper, we ranked which FWI input components were the most important for driving burnt area in each of the different GFED regions (RH and temperature tended to be the most important, but it varied by region which one was dominant, and occasionally it was something else like wind speed that mattered most). This was looking at the drivers of burnt area, but it should be possible to say something similar about which input variables are most important for influencing the different CFFWIS indices, and therefore make some general statement abouts which meteorological variables tend to be more/less consistently well-simulated, and therefore result in certain indices to be more/less consistently well-simulated.

Minor comments:

- L28: "*Wildfires burn hundreds of millions of hectares of forest each year around the world (Giglio et al., 2013…)*". Hundreds of millions of hectares is actually the total burnt area of all land cover types (~ 350 Mha in Giglio et al., although this may be an underestimate). The vast majority of this is savannah fires; the amount of forest burnt is only a small fraction (~ 5%) of the total (c.f. Figure 4 in Giglio et al.).
- 2.4: More precise details of the data processing and metrics used are needed here to fully understand the comparison being made. E.g.:
  - L156-157: "*simulations from each GCM are then compared to corresponding GEFF-ERA5 fields between 1980 and 2014"*. I assume that by 'simulations' the authors mean the "historical" experiment from CMIP6, but please specify this. Similarly I assume the analysis period goes from 1980 to 2014 because the 'historical' experiment in CMIP6 only goes up to 2014, but again for the benefit of readers who aren't familiar with the details of the CMIP6 suite of scenarios, it would be useful to clarify this.

- L158: "*monthly mean and 90th percentile statistics"*. I'm assuming that the monthly mean analysis is done for a monthly climatology, i.e. where the daily FWI indices are averaged for each month across all years between 1980-2014, rather than being compared year-by-year. However please clarify this. Similarly, please clarify what the 90th percentile is the 90th percentile of. g. is it the 90th percentile of all the individual monthly means? Is it the 90th percentile of the daily FWI values for each month, which are then averaged into a monthly climatology? Is it the 90th percentile of all the daily FWI values across the whole year? Or something else?
- L163: "*ratio of observed standard deviation to assess the representation of variance"*. Is this the spatial variance (i.e. s.d. between different gridbox values), or temporal variance (i.e. s.d. of the year-to-year timeseries)? I'm assuming it's spatial, given that it's later plotted on a Taylor diagram against the spatial correlation and RMSE, but please clarify. Assuming that is it a spatial variance, it may also be worth clarifying that these three metrics are not entirely independent measures of performance (which is of course why they can be plotted together on a Taylor diagram in 2 dimensions) – if you know any two of these metrics then it uniquely determines the third. (This is relevant for interpreting Section 4, where model skill is ranked based on how often a model scores well for all three metrics together).
- L169-170: "*those months for which the total burned area is greater than 50% of the maximum burned area across all months"*. Is this the maximum burned across all 12 months of a monthly climatology (i.e. averaged for each month over 1996-2016), or is it the maximum month from any point in the raw 252-month time series? If the latter, this strikes me as a potentially restrictive definition of fire season, since it could be very sensitive to one extreme year where there was much higher burned area than usual.

- L174-175: "*For all CFWIS components, global patterns are generally similar for both the annually-averaged monthly mean (Fig. 2; centre column) and 90th percentile statistics (Fig. 3; centre column)"*. Is this talking about the CMIP6 models, or is it still talking describing the GEFF-ERA5 patterns? The previous sentence only talks about GEFF-ERA5, but the centre column of Figs 2 and 3 relate to the CMIP6 models.
- Figures 2 and 3: The caption says 'Annual means', but the labels at the top of each column say 'mon_mean' and 'mon_p90' respectively, which is a little confusing. (Especially for Figure 3, c.f. my previous comment that it's confusing what the 90th percentile is of – e.g. is it the 90th percentile of monthly mean values, or is it a monthly climatology of the 90th percentile of daily FWI values?)
- Figures 2, 3, 4, 9: Axes label text is much too small to read. Figure 4 is the worst offender; I printed it out in A4 and it's impossible to read any of the colourbar, row, or column labels. Colourbar labels on Figs 2, 3, and 9 also need to be bigger.
- L249: Title line of Figure 4 caption describes it only as "Bias in monthly means…." however the figure shows the bias in 90th percentile as well (with equal prominence), which should therefore also be reflected in the title description.
- L256: "*monthly burned area for each region"*. Presumably this is from GFED4; it could be helpful to specify this in the caption.
- L265-267: "*At the global scale, the representation of DMC, DC and BUI is similar among models, which all present similar patterns, with greater inter-model variability and thus greater uncertainty, for both monthly mean (Fig. 5b, c, e) and 90th percentile annual values (Fig. 6b, c, e)"*. As currently worded, this is quite a confusing sentence – it initially says that all models show similar patterns of DMC, DC and BUI, but then says there's large inter-model variability, which seems contradictory? Also unclear: what is the 'greater inter-model variability' greater than?
- L270: "one indicator to the other" -> "one indicator to another" (because there's more than one other indicator)
- L270-271: "*model performance varies greatly from one indicator to the other. For instance, the GFDL-CM4 model performs well for all CFWIS components"*. Another slightly confusing wording, as 'GFDL-CM4 performs well for all CFWIS components'

appears to be a counterexample to the preceding statement that model performance varies greatly from one indicator to the other, rather than an instance of it.

- Figure 8: Could a row also be added for the global rankings?
- Section 4: "*models were ranked according to… the count of the number of times for which each model falls into the upper tercile in terms of all three spatiotemporal skill metrics (i.e., correlation, normalised RMSE and the ratio of standard deviation)"*. Is there a danger of double counting by ranking the models in this way? Since (as far as I understand), these three metrics are not independent – any model which performs well by two metrics will automatically perform well on the third (I think?), since they are related by the Taylor diagram.
- L324: "*all three spatiotemporal skill metrics"* – what is the temporal element of these skill metrics? As far as I've understood, all three are purely spatial metrics calculated across the gridbox values of time-averaged FWI indices (though I may well have misunderstood; if so perhaps Section 2.4 could be clarified to give more detail on how the metrics are defined).
- L362: "*a comprehensive evaluation of CMIP6 performance"* – while an excellent evaluation, I'm not certain it can be described as 'comprehensive... of CMIP6' when only 16 out of ~50 CMIP6 models are included.
- L365: "*for the period 1979-2014"* – Earlier in the text (L157) the analysis period was given as 1980-2014; which range is correct?
- L381-384: "*the large differences in model performances highlight the importance of a comprehensive model selection. This could significantly affect the conclusion provided in previous assessments... using a multi-model mean"* – it would be interesting to check how much the MMM bias improves by in an ensemble where only the best performing models are included. Or do the errors in different models cancel each other such that the MMM performance is actually similar either way?
- On a related note, and just as a quick aside, good (bad) performance at simulating historical FWI isn't necessarily a guarantee that models will project future changes in FWI well (badly). This is probably beyond the scope of the current paper, but if the authors have any plans to extend this work, it could be interesting to take the best performing models and see whether or not they project the same changes in FWI for a given future SSP scenario, or whether they diverge in their future projections…