Response to reviewers

Gallo, C., Eden, J. M., Dieppois, B., Drobyshev, I., Fulé, P. Z., San-Miguel-Ayanz, J., and Blackett, M.: Evaluation of CMIP6 model performances in simulating fire weather spatiotemporal variability on global and regional scales, Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2022-223, in review, 2022.

Please find the attached revised version of above paper. We thank the three reviewers for their very thorough and insightful reviews. We are pleased that the reviewers understood the study's motivation and recognise the work as a potentially valuable contribution to the literature in this field.

In responding to the comments of each Reviewer, we have made several changes to the manuscript. These were detailed in the 'Author Comment 1' (AC1) posted on 14/02/2023. A summary of the most important changes in response to the reviewers' comments is as follows:

- Reviewer 3 identified a potential issue with the way in which the fire weather indicators were calculated from CMIP6 fields. While Reviewer 3 accepts that our use of daily maxima and minima to approximate local noon conditions is justified, they suggested that we should use the same approach to approximate historical fire weather using input fields from ERA5. Therefore, in preparing our revision, we have taken the additional step of calculating fire weather indices using daily maximum temperature, daily minimum relative humidity, daily mean wind speed and total daily precipitation from ERA5, to compare with indicators taken directly from the GEFF-ERA5 fire danger reanalysis product. While the spatial patterns in the fire weather indicators derived from ERA5 and GEFF-ERA5 are broadly similar, we felt that there are sufficient differences to warrant using the ERA5-derived indicators as an evaluation reference. As a result, we have redone our analysis and updated Figures 2 to 9 (included at the end of this document), as well as amended the text where appropriate. We retain GEFF-ERA5 in Figures 2 and 3 for comparison, as this represents the most accurate and state-of-the-art representation of the fire weather indices, and is the dataset most likely to be used by others in the community.
- Reviewer 1 suggested that we acknowledge the "hot model" problem in CMIP6. In the revision, we ensure that this issue is discussed in sufficient detail in the Conclusions and Outlook section. We have also added relevant references.
- In response to all three reviewers, we have added a brief summary of multi-model mean performance in simulating the four individual meteorological components used to generate the fire weather indicators in supplementary material (namely, daily maximum temperature, daily minimum relative humidity, daily mean wind speed and daily total precipitation). We have produced an additional figure (Figure S1), which is placed in the Supplementary Material, and we have added text to the results and synthesis accordingly.

Below are the authors' detailed responses to Reviewers' comments.

Reviewer 1

https://doi.org/10.5194/gmd-2022-223-RC1

Synopsis

This paper evaluates the skill of 16 CMIP6 models in reproducing historical observed fire weather conditions as viewed through the CFDRS. The paper is well-written, well-structured and the methodology is robust. I think the paper has the potential to be published but I have a few comments below that need to be addressed before publication.

Response: We thank the reviewer for their efforts in reviewing this manuscript. We are pleased that the quality of the writing, structure and methodology was well-received. We have addressed each of the reviewer's comments in turn.

Major comments

Section 2.3 – Are authors aware of the hot model problem in CMIP6 ? See the paper below. This should be discussed and acknowledged.
https://www.nature.com/articles/d41586-022-01192-2

Response: We thank the reviewer for noting the importance of this issue, which is highly relevant and should indeed by properly discussed. In our revision, we have extended our *Conclusions and outlook* section to ensure that this issue is properly acknowledged, with reference to preceding work.

Change: The following changes have been made:

- [*Conclusions and outlook*, paragraph 3]: "We also note that CMIP6 models have been found to show a greater warming extent than CMIP5 (Coppola et al., 2020; Hausfather et al., 2022), with several models exhibiting far greater equilibrium climate sensitivity (Forster et al., 2020; Zelinka et al., 2020). It remains unclear to what extent some warming rates may be unrealistic, and how this might manifest in the calculated indicators."
- [References added]:
  - Coppola, E., Nogherotto, R., Ciarlo, J. M., Giorgi, F., van Meijgaard, E., Kadygrov, N., Iles, C., Corre, L., Sandstad, M., Somot, S., Nabat, P., Vautard, R., Levavasseur, G., Schwingskackl, C., Sillmann, J., Kjellström, E., Nikulin, G., Aalbers, E., Lenderink, G., Christensen, O. B., Boberg, F., Sørland, S. L., Demory, M., Bülow, K., Teichmann, C., Warrach-Sagi, K., and Wulfmeyer, V.: Assessment of the European climate projections as simulated by the large EURO-CORDEX regional and global climate model ensemble, J. Geophys. Res. Atmos., 126, e2019JD032356, https://doi.org/10.1029/2019JD032356, 2020.
  - Forster, P M., Maycock, A. C., McKenna, C. M., and Smith, C. J. Latest climate models confirm need for urgent mitigation, Nat. Clim. Chang., 10(1), 7–10, https://doi.org/10.1038/s41558-019-0660-0, 2020.
  - Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J., and Zelinka, M.: Climate simulations: recognize the 'hot model' problem, Nature, 605, 26-29, 2022.

- Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S.A. and Taylor, K.E.. Causes of higher climate sensitivity in CMIP6 models, Geophysical Research Letters, 47, e2019GL085782. https://doi.org/10.1029/2019GL08578210.1029/2019JD032356, (2020)

Line 166-170 – Defining a unique fire season for each GFED region is questionable given their spatial extent. The timing of the fire season has already been reported in a number of previous studies and has been shown to be much variable in space. The present coarse-scale analysis is thus likely to mix a number of different seasonalities within each GFED region. I would suggest to define a fire season locally (e.g. at the pixel scale) as done in most previous studies. Moreover, the model ranking (currently based on GFED region) would be much more relevant if authors would consider the spatial extent (number of pixels) where a model falls within a specific tercile. The current ranking is dependent on the size of each GFED region (a small region contributes as much as a large one).

Response: We thank the reviewer for this comment. We engaged in much discussion early in the process about the use of GFED areas for regionalisation. We were encouraged by the use of these areas in previous work (e.g., Mezuman et al. 2020; Grillakis et al. 2022; Liu et al 2022) and that they are potentially more meaningful for climate-fire analysis than, say, the IPCC regions. While we still consider the regionalisation to be a useful approach to presenting these results, we accept that the definition of region-wide fire seasons is rather coarse. In our revision, we highlight where similar approaches have been used before (2.4 Model evaluation) and make recommendations for local- and regional-scale studies to undertake their own evaluation taking into account a more precise fire season (Conclusions and outlook).

Change: The following changes have been made to the text.

- [Section 2.4, paragraph 2]: “Regional analysis is based on 14 GFED-defined fire regions originally presented by Giglio et al. (2006) and van der Werf et al. (2006), and widely used in subsequent work (e.g., Giglio et al., 2010; 2013; Andela et al., 2019; Mezuman et al., 2020; Grillakis et al., 2022; Liu et al., 2022).”
- [*Conclusions and outlook*, paragraph 3]: “A final point concerns the GFED fire regions taken as the basis for the regional-scale analysis: while they are a useful categorisation for the purpose of this evaluation, fire regimes vary substantially at the intra-regional scale. Potential alternative categorisations, in Europe for example, include the fire regimes defined by Galizia et al. (2021), while fire-prone areas may be better isolated using high-resolution land surface data (e.g., Normalized Difference Vegetation Index).”
- [Included additional reference]:

– Grillakis, M., Voulgarakis, A., Rovithakis, A., Seiradakis, K. D., Koutroulis, A., Field, R. D., Kasoar, M., Papadopoulos, A., and Lazaridis, M.: Climate drivers of global wildfire burned area, Environ. Res. Lett., 17, 045021, https://doi.org/10.1088/1748-9326/ac5fa1, 2022.

Minor comments

Line 65-68 - GCMs have also been used in attribution studies of FWI to quantify the current risk

Barbero R, Abatzoglou J T, Pimont F, Ruffault J and Curt T. 2020 Attributing increases in fire weather to anthropogenic climate change over France. Frontiers in Earth Science https://doi.org/10.3389/feart.2020.00104

Response: Thank you for drawing our attention to this relevant work. We have added its citation to the text.

Change: The following change has been made to the text.

- [Section *1 Introduction*, paragraph 4]: "GCMs have been used frequently to quantify the link between wildfire activity and weather conditions (Bedia et al., 2015; Williams and Abatzoglou, 2016) and, specifically, to simulate fire weather, both in the past and under future climate change scenarios (Moritz et al., 2012; Flannigan et al., 2013; Bedia et al., 2015; Littell et al., 2018; Abatzoglou et al., 2019); and also, in recent attribution studies to assess the influence of anthropogenic climate change on fire weather (Barbero et al., 2020; Liu et al., 2022)."
- [Reference added]:

- "Barbero, R., Abatzoglou, J. T., Pimont, F., Ruffault, J., and Curt, T.: Attributing increases in fire weather to anthropogenic climate change over France, Front. Earth Sci., 8, 104, https://doi.org/10.3389/feart.2020.00104, 2020."

Line 130-131 – I am not sure to understand how links between fire events and fire weather trends relate to the performance of ERA-5 in reproducing observed fire weather conditions Two independent variables can follow similar trends !

Response: Thank you for identifying this inconsistency. Fire weather (and hence fire danger) should not be expected to follow the same trend as the frequency of actual fire events. We have removed a sentence to avoid confusion.

Change: The following sentence has been removed.

- [Section 2.3, paragraph 1]: "Similarly, Vitolo et al. (2020) identified links between trends in fire weather indices and fire events."

Line 150-155 – As stated by authors, the record length of GFED is a bit short to draw such conclusions. What about using a land cover map to define what is burnable or not?

Response: Thank you for this suggestion. We consider the GFED product to be appropriate for defining fire-prone areas at the global and regional scale. Our 50km smoothing approach was taken to ensure that that no burnable areas were missed. This approach appears in a study published last year (Liu et al., 2022) which we duly cite. Defining fire-prone areas using NDVI (or similar) makes perfect sense smaller spatial scales, and we have added this recommendation to our *Conclusions and outlook* section.

Change: The following changes have been made.

- [Section 2.4, paragraph 1]: "Following the approach of Liu et al. (2022) in isolating burnable area, all grid points within a 50 km radius of a record of burned area are identified as 'fire-prone' in order, to account for the spatial randomness of fire activity and the relatively short record of the GFED4 data."
- [Section 5, paragraph 3]: "Potential alternative categorisations, in Europe for example, include the different fire regimes defined by Galizia et al. (2021), while fire-prone areas may be better isolated using high-resolution land surface data (e.g. Normalized Difference Vegetation Index)."
- [Reference added]:
  - Galizia, L.F., Curt, T., Barbero, R., and Rodrigues, M.: Understanding fire regimes in Europe, Int. J. Wildland Fire, 31, 56-66, https://doi.org/10.1071/WF21081, 2021.

Line 157 – why this specific time period? This needs clarification

Response: Thank you for requesting clarification. The period 1980-2014 is the longest extent for which the ERA5 reanalysis and CMIP6 simulations overlap. Of course, there is no temporal synchronicity between ERA5 and CMIP6 but it is important for the evaluation period to be consistent. We offer clarification in the text.

Change: The following change has been made

- [Section 2.4, paragraph 2]: "To understand the overall model representation of all CFWIS components (Fig. 1), historical simulations from each GCM are then compared to corresponding ERA5-calculated fields between 1980 and 2014, the maximum period for which ERA5 and CMIP6 data are concurrently available."

Line 158 – Are you referring to the 90th percentile of daily values?

Response: Indeed, that is correct. The text has been amended for clarification.

Change: The following sentence has been amended.

- [Section 2.4, paragraph 2]: "Additionally, to account for severe fire weather, performance is also quantified by representation of the 90th percentile, constructed for each month using daily CFWIS values across all years."

Figure 2 – the title indicates (mon_mean) while the caption reads annual mean. Please use the same terminology for consistency.

Response: Thank you for identifying this inconsistency. The Figure refers to annually-averaged monthly means and the caption has been amended accordingly.

Change: The caption has been changed.

- [Figure 2]: "Annually-averaged monthly means for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile,

white colour represents lower bias and darker blue/red higher negative/positive bias."

Figure 2 and 3 – I wonder if the (minor) differences between average and extreme statistics need to be reported. Please consider moving Figure 3 to the supplementary information and reduce the text accordingly.

Response: Thank you for the suggestion. Although in both figures CFWIS indicator biases present similar patterns, we have decided to keep both. We anticipate that some readers will be more interested in model performance at higher quantiles and, given that the two figures are not identical, we prefer to present both without asking the reader to visit supplementary material.

Line 197-198 – Given the (rather expected) similarities in the results, I suggest moving DSR results to the supplementary information.

Response: Thank you for this suggestion. Although we recognise the similarities between FWI and DSR, they are not exactly the same (e.g., differences in several regions like in South America and Sub-Saharan Africa, where DSR bias are lower). We, therefore, prefer to retain DSR in the results.

Line 217-219 – this belongs to the methods

Response: Thank you, text has been moved and amended.

Change: The following text has been amended.

- [Section 2.4, paragraph 2]: "To isolate CMIP6 performance during periods that are most conducive to fire activity, a fire season was established for each region based on available GFED4 burned area data. For each GFED-defined region, the fire season was defined by those months for which the total burned area is greater than 50% of the maximum burned area across all months, averaged for each month over the available 1996-2016 period."
- [Section 3.2, paragraph 1]: deleted "To isolate CMIP6 performance during periods that are most conducive to fire activity, a fire season was established for each region base on available GFED4 burned area data. For each GFED-defined region, the fire season was defined by those months for which the total burned area is greater than 50% of the maximum burned area across all months"

Figure 4 – please increase label size (x and y labels) as well as the colorbar. They are very hard to read.

Response: Thank you for this observation. We note that the figures in the original submission were rendered in raster format – the final versions will be vector format (e.g. .pdf). However, we have made increased the sizes of labels in Figures 2, 3 and 4 as suggested.

Change: The following changes have been made.

- [Figure 2 and Figure 3]: Panel labels and colourbar labels have been largened.
- [Figure 4]: x and y and colourbar labels have been largened.

Figures 5-6 – Would that make sense to add the MMM for comparison with each individual GCM?

Response: Thank you for this suggestion. Indeed, this is a useful reference point.

Change: The following changes have been made.
- [Figures 5 and 6]: An additional point showing the multi-model mean has been added to each panel of each figure.

Figure 5-6 – Please indicate the RMSE error on the figure for clarity

Response: Thank you for this suggestion. We have amended the figures accordingly.

Change: The following change has been made.
- [Figures 5 and 6]: An additional legend has been added to show that the black (dashed), grey (solid) and blue (dashed) lines represent intervals for correlation, RMSE and standard deviation respectively.

Figure 6 – please consider moving this figure to SI given the similarity with figure 5

Response: Thank you for this suggestion. After some consideration, we have decided to retain both Figure 5 and Figure 6 in the main article. As mentioned earlier, we anticipate that some readers will be more interested in model performance at higher quantiles and, given that the two figures are not identical, we prefer to present both without asking the reader to visit supplementary material.

Line 355-358 – Does it mean that a model performing well in reproducing historical variability is more credible in simulating future changes to fire weather conditions? I am just asking.

Response: Thank you for raising this point. As a first order evaluation, we believe our results show provide a basis upon which fire weather projections from different models might be interpreted. But additional analysis might be required to understand model capacity to simulate realistic trends. We have added a small reminder to our *Conclusions and outlook* section.

Change:
- [Section 5, paragraph 2]: "It is anticipated that the evaluation presented here, while based on solely historical spatiotemporal variability, will serve as an important resource for users of model-simulated fire weather, both during the CMIP6 era and beyond, in three different ways."

Line 402-405 – Yes, other regionalization such as the pyroregions presented in Galizia et al. (2021) over Europe would be more relevant in terms of fire activity and management.

Galizia, L. F. C., Curt, T., Barbero, R., Rodrigues, M. (2021). Understanding fire regimes in Europe, International Journal of Wildland Fire, 31, 56-66. https://doi.org/10.1071/WF21081

Response: Thank you for this suggestion. We have referenced this work accordingly.

Change: The following changes have been made.

- [Section 5, paragraph 3]: "Potential alternative categorisations, in Europe for example, include the different fire regimes defined by Galizia et al. (2021), while fire-prone areas may be better isolated using high-resolution land surface data (e.g. Normalized Difference Vegetation Index)."
- [Reference added]:
- Galizia, L. F., Curt, T., Barbero, R., and Rodrigues, M.: Understanding fire regimes in Europe, Int. J. Wildland Fire, 31, 56-66, https://doi.org/10.1071/WF21081, 2021.

A future interesting research question would be to examine what meteorological variables (temp, prcp, RH and wind) are responsible for the difference between observed and simulated CFDRS components. This could be discussed somewhere in the paper.

Response: Thank you for this suggestion. In response to this comment, and those of the other reviewers, we have added a brief summary of multi-model mean performance in simulating the four meteorological components used to generate the fire weather indicators. We are aware that previous work has given more attention to these (particularly to the evaluation of temperature and precipitation), but a summary is nevertheless helpful for the reader here and allows us to discuss some of our findings with respect to the representation of those component variables.

Change: The following changes have been made.

- [Figure S1]: An additional figure showing ERA5, CMIP6 multi-model mean, and bias in annual mean maximum temperature, precipitation, minimum relative humidity and wind speed.
- [Section 3.1, paragraph 4]: "The biases are driven by multi-model representation of the four meteorological components required as input for the CFWIS indicators: daily values for maximum temperature, mean wind speed, minimum relative humidity and total precipitation. The representation of these fields in ERA5 and CMIP6 is shown in Fig. S1 (see supplementary material). Biases are apparent in all four fields, most strikingly in the representation of relative humidity in the northern hemisphere (Fig. S1i). However, cooler maximum temperatures in boreal Eurasia (Fig. S1c) does not appear to impact on the representation of fire weather (Figs. 2 and 3; fourth column). Overestimation of precipitation in southern Africa (Fig. S1f) may be responsible for an underrepresentation of DC and DMC in particular (Figs. 2 and 3; fourth column)."
- [Section 4, paragraph 5]: "Our synthesis does not consider model representation of the meteorological components taken as input in deriving the CFWIS indicators. A first-order analysis of multi-model biases in these fields is given in Section 3.1 and Fig. S1 of the supplementary material, but more in-depth analysis of the relative contribution of biases in each field to the overall representation of fire weather is

beyond the scope of this study. Clearly, model development in fire weather representation of fire weather, especially in a changing world, should consider the reasons for model biases in key fire-prone regions. This includes the representation of temperature highs and relative humidity lows in large parts of the northern hemisphere."

- [Section 4, paragraph 5]: "Our synthesis does not consider model representation of the meteorological components taken as input in deriving the CFWIS indicators. A first-order analysis of multi-model biases in these fields is given in Section 3.1 and Fig. S1 of the supplementary material, but more in-depth analysis of the relative contribution of biases in each field to the overall representation of fire weather is beyond the scope of this study. Clearly, model development in fire weather representation of fire weather, especially in a changing world, should consider the reasons for model biases in key fire-prone regions. This includes the representation of temperature highs and relative humidity lows in large parts of the northern hemisphere."

Reviewer 2

https://doi.org/10.5194/gmd-2022-223-RC2

This work evaluates the performance of 16 CMIP6 models in reproducing historical fire weather indicators represented by the Canadian Fire Weather Index System by comparing the results to those produced by GEDD-ERA5. This paper is written in a concise manner, it is well structured and provides a robust and insightful analysis that allows a better understanding of the performance of CMIP6 models in capturing fire related weather. In my opinion, this work provides a useful contribution to the scientific community, aiding model selection for future related studies focused on the use of these for future climate driven fire projections. However, before this is considered for publication, there are several clarifications which should be provided by the authors.

Response: We thank the reviewer for their assessment of our manuscript. We are pleased that the importance and practical use to the scientific community was recognised. We have addressed each of the reviewer's comments in turn.

Lines 125 – 129 – The work by Vitolo et al. (2020) describes GEFF-ERA5, the reanalysis dataset of FWI fire behaviour indices based on the ERA5 reanalysis, including the impact of resolution by comparing to GEFF-ERAI. Please consider including a reference to this work.

(Vitolo, C., Di Giuseppe, F., Barnard, C. et al. ERA5-based global meteorological wildfire danger maps. Sci Data 7, 216 (2020). https://doi.org/10.1038/s41597-020-0554-z)

Response: Thank you, we acknowledge the importance of the work done by Vitolo et al. (2020) in improving the previous dataset. To account for the impact of the resolution, we have added the reference also in this sentence.

Change: The following change has been made to the text.

- [Section 2.3, paragraph 2]: "In general, ERA5 provides a realistic and temporally coherent approximation of real-world weather states, with higher spatial and temporal resolutions and better estimates of meteorological variables compared to ERA-Interim (Dee et al., 2011; Hersbach et al., 2019), reducing biases and increasing correlation with observations (Graham et al., 2019; Gleixner et al., 2020; Tarek et al., 2020; Vitolo et al., 2020)".

Lines 150 – 155 – My understanding is that this paragraph is highlighting that this work will be focused on the fire-prone areas of the world, as defined in GFED4. It's not clear to me how the last sentence of this paragraph is relevant for this work, as no further mention to GFED (other than the regions) is made throughout this work.

Response: Thank you for this comment. GFED burned area data as a mask to isolate all 'fire-prone' areas. The 50km smoothing is used to ensure that that we (approximately) include any fire-prone areas where burned areas may not have been recorded in the relatively short GFED record. We have modified the text for clarification.

Change: The following text has been amended.

- [Section 2.4, paragraph 1]: “Following the approach of Liu et al. (2022) in isolating burnable area, all grid points within a 50 km radius of a record of burned area are identified as ‘fire-prone’ in order, to account for the spatial randomness of fire activity and the relatively short record of the GFED4 data.”

Section 3 – Throughout this section the authors analyse the differences in the fire indices between the different CMIP6 models and GEDD-ERA5. While this provides a useful insight and understanding on the FWI components bias, further expanding these to the fire weather components (e.g., temperature, wind, precipitation, etc…) would provide a better understanding of the drivers of said bias, strengthening the evaluation provided by this work. This is especially relevant at regional level and may help with model selection, as well as inform future model development.

Response: Thank you for this suggestion. In response to this comment, and those of the other reviewers, we have added a brief summary of multi-model mean performance in simulated the four meteorological components used to generate the fire weather indicators. We are aware that previous work has given more attention to these (particularly to the evaluation of temperature and precipitation), but as summary is nevertheless helpful for the reader here and allows us to discuss some of our findings with respect to the representation of those component variables.

Change: The following changes have been made.

- [Figure S1]: An additional figure showing ERA5, CMIP6 multi-model mean, and bias in annual mean maximum temperature, precipitation, minimum relative humidity and wind speed.
- [Section 3.1, paragraph 4]: “The biases are driven by multi-model representation of the four meteorological components required as input for the CFWIS indicators: daily values for maximum temperature, mean wind speed, minimum relative humidity and total precipitation. The representation of these fields in ERA5 and CMIP6 is shown in Fig. S1 (see supplementary material). Biases are apparent in all four fields, most strikingly in the representation of relative humidity in the northern hemisphere (Fig. S1i). However, cooler maximum temperatures in boreal Eurasia (Fig. S1c) do not appear to impact on the representation of fire weather (Figs. 2 and 3; fourth column). Overestimation of precipitation in southern Africa (Fig. S1f) may be responsible for an underrepresentation of DC and DMC in particular (Figs. 2 and 3; fourth column).”
- [Section 4, paragraph 5]: “Our synthesis does not consider model representation of the meteorological components taken as input in deriving the CFWIS indicators. A first-order analysis of multi-model biases in these fields is given in Section 3.1 and Fig. S1 of the supplementary material, but more in-depth analysis of the relative contribution of biases in each field to the overall representation of fire weather is beyond the scope of this study. Clearly, model development in fire weather representation of fire weather, especially in a changing world, should consider the reasons for model biases in key fire-prone regions. This includes the representation of temperature highs and relative humidity lows in large parts of the northern hemisphere.”

Figure 2 and 3 – The caption refers to these figures as Annual means, however subtitles on each top tile refer to mon_mean. This should be reviewed and made consistent.

Response: Thank you. This indeed refers to annual mean of monthly means. Have made changes to clarify.

Change: The following change has been made:

- [Figure 2 and Figure 3]: Column titles changed.
- [Figure 2 caption]: "Annually-averaged monthly means for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias."
- [Figure 3]: "Annually-averaged values of monthly climatology of the 90th percentile for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias."

Figure 4 – Figure label are not legible, please consider increasing the font.

Response: Thank you for this observation. We note that the figures in the original submission were rendered in raster format – the final versions will be vector format (e.g. .pdf). However, we have made increased the sizes of labels in Figure 4.

Change: The following changes have been made.

- [Figure 4]: x and y and colourbar labels have been enlarged. The updated figure has been appended to the end of this document.

Lines 344 – 346 – Although it is stated that it is difficult to identify systematic reasons for inter-model differences, having an analysis of the meteorological driver (e.g., temperature and precipitation) between models may help better understand the inter-model bias.

Response: Thank you for raising this issue again. As stated above, we have briefly summarised the multi-model biases in the meteorological components and made recommendations for further analysis in future study.

Change:

- [Section 4, paragraph 5]: "Our synthesis does not consider model representation of the meteorological components taken as input in deriving the CFWIS indicators. A first-order analysis of multi-model biases in these fields is given in Section 3.1 and Fig. S1 of the supplementary material, but more in-depth analysis of the relative contribution of biases in each field to the overall representation of fire weather is

beyond the scope of this study. Clearly, model development in fire weather representation of fire weather, especially in a changing world, should consider the reasons for model biases in key fire-prone regions. This includes the representation of temperature highs and relative humidity lows in large parts of the northern hemisphere."

Lines 346 – 350 – It is mentioned that there is little evidence on the impact of spatial resolution, but it is mentioned that MPI-ESM1-2-HR consistently performs better than MPI-ESM-ESM1-2-LR and MPI-ESM-1-2-HAM. Comparing the impact of resolution using different models may not provide the robust framework to draw conclusions, as the effect of different resolution may be impacted by different model formulation (e.g., different dynamics, physics, inputs, etc…). Furthermore, Vitolo et al. (2020) shows the benefits of resolution in FWI between GEFF-ERAI GEFF-ERA5.

Response: We thank the reviewer for raising this important point. We have revised the text in our *Synthesis and discussion* section to add clarification.

Change: The following text was amended.

- [Section 4, paragraph 3]: "It is also notable that the CanESM5 model has the lowest resolution (2.8° x 2.8°) but outperforms many higher resolution models in several regions, particularly Boreal North America (BONA) and Central America (CEAM). However, this observation aside, there is little evidence for a model's original spatial resolution as an important factor in its performance. Comparison of different models does not provide an ideal framework to draw conclusions as the impact of resolution is likely to be driven by internal model physics and dynamics."

Line 390 – resolution should be considered a caveat, especially following the mention of the fire regimes vary substantially at the intra-regional scale in lines 402 – 405.

Response: Thank you again for identifying the caveat related to resolution. This paragraph in our *Conclusions and outlook* section has been amended.

Change: The following text was amended.

- [Section 4, paragraph 3]: "Potential alternative categorisations, in Europe for example, include the different fire regimes defined by Galizia et al. (2021), while fire-prone areas may be better isolated using high-resolution land surface data (e.g. Normalized Difference Vegetation Index)."

Reviewer 3

https://doi.org/10.5194/gmd-2022-223-RC3

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.

Response: We're grateful to the reviewer for their robust and helpful review. We are pleased that the potential value and importance of the results presented. We have addressed each of the reviewer's comments in turn.

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.

Response: We're grateful to the reviewer for raising this point. We were mindful of this potential discrepancy when preparing our original submission (Section 2.3, paragraph 2). The suggestion to calculate the CFWIS indicators from ERA5 fields using the same approach applied to CMIP6 models is a sensible suggestion. In preparing our revision, we did exactly this, using ERA5 daily maximum temperature, daily minimum relative humidity, daily mean wind speed and total daily precipitation as proxies for local noon conditions. The results are presented in revised Figures 2 and 3; in each figure, an additional column has been added to show mean and p90 values derived directly from ERA5 input fields. While the spatial patterns in the fire weather indicators derived from ERA5 and GEFF-ERA5 are broadly similar, we felt that there are sufficient differences to warrant using the ERA5-derived indicators as an evaluation reference. As a result, we have redone our analysis and updated Figures 2 to 9. We retain GEFF-ERA5 in Figures 2 and 3 for comparison, as this represents the most accurate and state-of-the-art representation of the fire weather indices, and is the dataset most likely to be used by others in the community.

Change: The following changes have been made.

- [Figures 2 and Figure 3]: An additional column has been added to each. The updated figures are appended to the end of this document.
- [Figure 2 caption]: "Annually-averaged monthly means for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias."
- [Figure 3 caption]: "Annually-averaged values of monthly climatology of the 90th percentile for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias."

- [Figures 4-9]: These figures have been updated with the results now made with reference to ERA5. The overall results are very similar to before; the text has been updated to reflect any changes.
- [Section 2.3, paragraph 1]: "In our case, as the CFWIS indicators generated from CMIP6 rely on daily values for the four meteorological components as proxies for noon conditions, and to ensure a fair comparison, we apply generate CFWIS indicators for ERA5 using the same input components. We make a comparison between ERA5 and GEFF-ERA5 to illustrate the consistency between the two sources of CFWIS information."

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.

Response: We concur that analysis of the meteorological components has considerable merit. We appreciate that the reviewer recognises that evaluation of fire weather is the primary importance of the paper, but we agree that analysis of the meteorological components would be useful to the reader. In response to this comment, and those of the other reviewers, we have added a brief summary of multi-model mean performance in simulated the four meteorological components used to generate the fire weather indicators. We are aware that previous work has given more attention to these (particularly to the evaluation of temperature and precipitation), but nevertheless it is helpful for the reader here and allows us to discuss some of our findings with respect to the representation of those component variables.

Change: The following changes have been made.

- [Figure S1]: An additional figure showing ERA5, CMIP6 multi-model mean, and bias in annual mean maximum temperature, precipitation, minimum relative humidity and wind speed.
- [Section 3.1, paragraph 4]: "The biases are driven by multi-model representation of the four meteorological components required as input for the CFWIS indicators: daily values for maximum temperature, mean wind speed, minimum relative humidity and total precipitation. The representation of these fields in ERA5 and CMIP6 is shown in Fig. S1 (see supplementary material). Biases are apparent in all four fields, most strikingly in the representation of relative humidity in the northern hemisphere (Fig. S1i). However, cooler maximum temperatures in boreal Eurasia (Fig. S1c) do not appear to impact on the representation of fire weather (Figs. 2 and 3; fourth column). Overestimation of precipitation in southern Africa (Fig. S1f) may be

responsible for an underrepresentation of DC and DMC in particular (Figs. 2 and 3; fourth column)."

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?

Response: In addition to the figure and text added in response to the previous comment, we have recognised that our synthesis does not consider the meteorological input fields. We make reference to this and the implications for model development in our response.

Change: The following additions have been made.

- [Section 4, paragraph 5]: "Our synthesis does not consider model representation of the meteorological components taken as input in deriving the CFWIS indicators. A first-order analysis of multi-model biases in these fields is given in Section 3.1 and Fig. S1 of the supplementary material, but more in-depth analysis of the relative contribution of biases in each field to the overall representation of fire weather is beyond the scope of this study. Clearly, model development in fire weather representation of fire weather, especially in a changing world, should consider the reasons for model biases in key fire-prone regions. This includes the representation of temperature highs and relative humidity lows in large parts of the northern hemisphere."

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.

Response: Thank you for these comments. In addition to the text added in response to the two previous comments, we have made further additions.

Change: The following additions have been made.

- [Section 5; paragraph 3]: "To truly understand the sources of error and biases for a given index, an in-depth analysis of relative contribution of the meteorological fields used to construct it is required. Such an analysis is not trivial and should be an important focus for future study."
- [References added]: "Grillakis, M., Voulgarakis, A., Rovithakis, A., Seiradakis, K. D., Koutroulis, A., Field, R. D., Kasoar, M., Papadopoulos, A., and Lazaridis, M.: Climate drivers of global wildfire burned area, Environ. Res. Lett., 17, 045021, https://doi.org/10.1088/1748-9326/ac5fa1, 2022."

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.).

Response: Thank you for this. This is true, the data of hundreds of millions of hectares make reference to global burned area in general, not specific to forests

Change: The following change has been made to the text.

- [Section 1, paragraph 1]: "Wildfires burn hundreds of millions of hectares each year around the world (Giglio et al., 2013; Yang et al., 2014; van Lierop et al., 2015; van Wees et al., 2021)."

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.

Response: Thank you for this remark. We agree and have changed it in the text.

Change: The following change has been made to the text.

- [Section 2.4, paragraph 2]: "To understand the overall model representation of all CFWIS components (Fig. 1), historical simulations from each GCM are then compared to corresponding ERA5-calculated fields between 1980 and 2014, the maximum period for which ERA5 and CMIP6 data are concurrently available".

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?

Response: Thank you for requesting clarification on this. The monthly climatologies are based on daily values of each CFWIS indicator. We calculated monthly means and monthly 90th percentiles from all daily values across all years. We have amended the text to provide clarification.

Change: The following change has been made to the text.

- [Section 2.4, paragraph 2]: "Model performance is then quantified through the ability of GCMs to simulate monthly mean climatologies of daily values of each CFWIS indicator with ERA5 used as a reference. Additionally, to account for severe fire weather, performance is also quantified by representation of the 90th percentile, constructed for each month using daily CFWIS values across all years."

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).

Response: Yes, it refers to spatial variance.

Change: The following change has been made to the text.

- [Section 2.4, paragraph 2]: "Multiple model performance metrics are used, including (i) spatial correlation to assess the representation of spatial variability; (ii) root mean squared error (RMSE) to assess the representation of mean states and the extent of model bias; (iii) ratio of observed standard deviation to assess the representation of spatial variance."

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.

Response: Thank you for requesting clarification on this. For each GFED region, burned area was averaged across all months, with the highest monthly total noted. In addition to this month, we added the following.

Change: The following change has been made to the text.

- [Section 2.4, paragraph 2]: “To isolate CMIP6 performance during periods that are most conducive to fire activity, a fire season was established for each region based on available GFED4 burned area data. For each GFED-defined region, the fire season was defined by those months for which the total burned area is greater than 50% of the maximum burned area across all months, averaged for each month over the available 1996-2016 period.”

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.

Response: It is intended for CMIP6 models, clarification has been added in the text.

Change: The following change has been made to the text.

- [Section 3.1, paragraph 1]: “For all CFWIS components, global patterns of the CMIP6 multi-model mean are generally similar for both the annually-averaged monthly mean (Fig. 2; centre-right column) and 90th percentile statistics of daily values (Fig. 3; centre-right column)”

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?)

Response: Figures refer to annual monthly means and monthly climatology of the 90th percentile of daily values, changed captions of the figures and amended text for more clarity].

Change: The following changes have been made to the text.

- [Section 3.1, paragraph 1]: “For all CFWIS components, global patterns of the CMIP6 multi-model mean are generally similar for both the annually-averaged monthly

mean (Fig. 2; centre-right column) and 90th percentile statistics of daily values (Fig. 3; centre-right column)"

- [Figure 2]: "Annually-averaged monthly means for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias"
- [Figure 3]: "A Annually-averaged values of monthly climatology of the 90th percentile for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias."

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.

Response: Thank you for this observation. We note that the figures in the original submission were rendered in raster format – the final versions will be vector format (e.g. .pdf). However, our revised Figures 2, 3, 4 and 9 have increased labels sizes.

Change: The following changes have been made.

- [Figure 2 and Figure 3]: Panel labels and colourbar labels have been enlarged. The updated figures have been appended to the end of this document.
- [Figure 4]: x and y and colourbar labels have been enlarged. The updated figure has been appended to the end of this document.
- [Figure 9]: Axes tick labels have been enlarged.

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.

Response: Thank you for spotting this, caption was incomplete.

Change: The following change has been made to the caption.

- [Figure 4]: "Bias in monthly means and 90th percentiles in seven CFWIS components simulated by the CMIP6 multi-model mean with respect to ERA5 across 14 GFED fire regions"

L256: "monthly burned area for each region". Presumably this is from GFED4; it could be helpful to specify this in the caption.

Response: Thank you, we have specified it in the caption for better clarity.

Change: The following change has been made to the caption.

- [Figure 4]: "Bar plots show the average monthly-burned area for each GFED region, represented as a fraction of the monthly maximum."]

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?

Response: Thank you, indeed, this is not properly worded. What is intended is that models draw similar patterns for those 3 indices, but no similarity is found among models themselves.

Change: The following change has been made to the text.

- [Section 3.3, paragraph 2]: "At the global scale, the representations of DMC, DC and BUI present similar patterns, with greater inter-model variability and thus greater uncertainty than the other indices, for both monthly mean (Fig. 5b, c, e) and 90th percentile annual values (Fig. 6b, c, e)"

L270: "one indicator to the other" -> "one indicator to another" (because there's more than one other indicator)

Response: Thank you, this has been corrected.

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.

Response: Thank you for noticing this inconsistency. We have amended the text for clarification.

Change: The following change has been made

- [Section 3.3, paragraph 3]: "GFDL-CM4 is an example of a model that performs well for all CFWIS components (Fig. 5)"

Figure 8: Could a row also be added for the global rankings?

Response: Thank you for this suggestion. We have a row for "WORLD" accordingly.

Change: The following change was made

- [Figure 8]: Additional row "WORLD" has been added showing respective ranking across all regions collectively.

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.

Response: Thank you for considering the implications of allowing all three metrics to determine the rankings. Indeed, the metrics are not strictly independent. But the same is true of the fire indicators themselves. Previous attempts to generate similar rankings (e.g. Vautard et al., 2021, https://doi.org/10.1029/2019JD032344) have also used co-dependent variables (e.g., maximum and minimum temperature). We appreciate that caution should be exercised when developing such a ranking, but we feel that our approach is not undermined but this issue.

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).

Response: Indeed, at this point these are spatial metrics. The sentence has been changed for clarification.

Change: The following change has been made.

- [*Synthesis and discussion*, paragraph 1]: "All 16 models were ranked according to two different measurements: (1) the count of the number of times for which each model falls into the upper tercile in terms of all three skill metrics (i.e., correlation, normalised RMSE and the ratio of standard deviation) for the seasonal mean and 90th percentile in each of the seven CFWIS components and across each of the 14 GFED fire regions (Fig. 9a); and (2) the count of the number of times in which a model falls into the lower tercile, indicating which models exhibit poorer performance more frequently (Fig. 9b)."

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.

Response: Thank you, we agree that comprehensive might not be the right word to use in this case, and that we should more accurately refer to a "subset" of CMIP6 models.

Change: The following change has been made to the text in our *Conclusions and outlook* section.

- [Section 5, paragraph 1]: “We presented a detailed evaluation of the performance of a subset of CMIP6 models in simulating spatiotemporal variability in fire weather across all parts of the world currently vulnerable to wildfire”.

L365: “for the period 1979-2014” – Earlier in the text (L157) the analysis period was given as 1980-2014; which range is correct?

Response: Thank you, the correct range is 1980-2014.

Change: The following change has been made to the text.

- [Section 5, paragraph 1]: “A set of fire weather indicators, defined by the CFWIS, were generated for 16 different CMIP6 models and compared with corresponding fields from the ERA5 fire danger reanalysis for the period 1980-2014”

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?

Response: Thank you for this comment, this will be part of our next analysis. We have added a sentence in the *Conclusions and outlook* section.

Change: The following change has been made to the text.

- [Section 5, paragraph 2]: “Future analysis will explore how the multi-model mean bias could be potentially reduced using a weighted mean or a MMM with those models showing better performance, and see how it is reflected in the projections for different Shared Socioeconomic Pathways (SSP) scenarios.”

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…

Response: Thank you, in the same line as previous comment, this is something that will be part of our next analysis. We have added a sentence in the conclusions.

Change: The following change has been made to the text.

- [*Conclusions and outlook*, paragraph 2]: “Future analysis will explore how the multi-model mean bias could be potentially reduced using a weighted mean or a MMM with those models showing better performance, and see how it is reflected in the projections for different Shared Socioeconomic Pathways (SSP) scenarios.”

New and updated figures

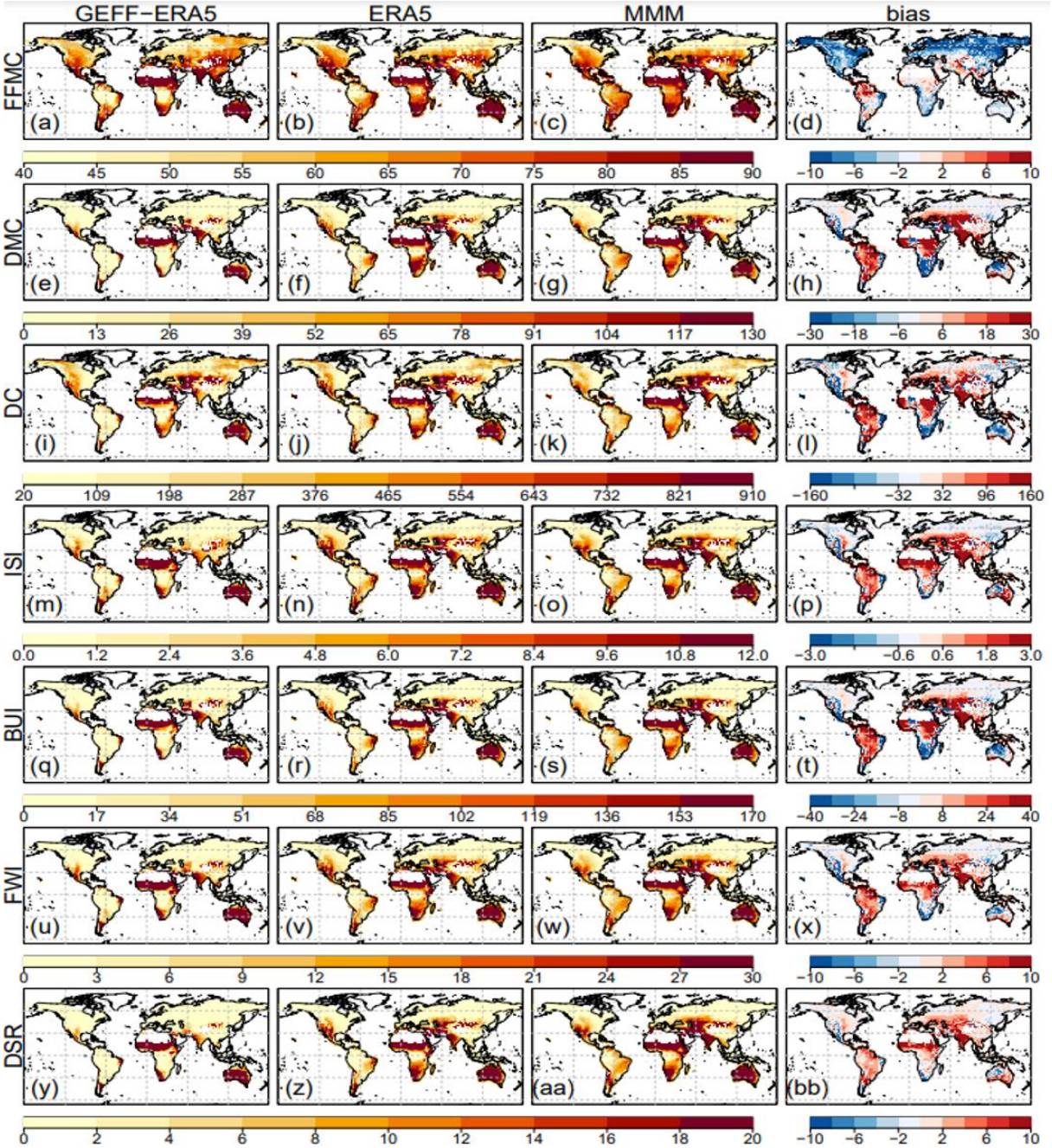

Figure 2: Annually-averaged monthly means for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t), FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias.

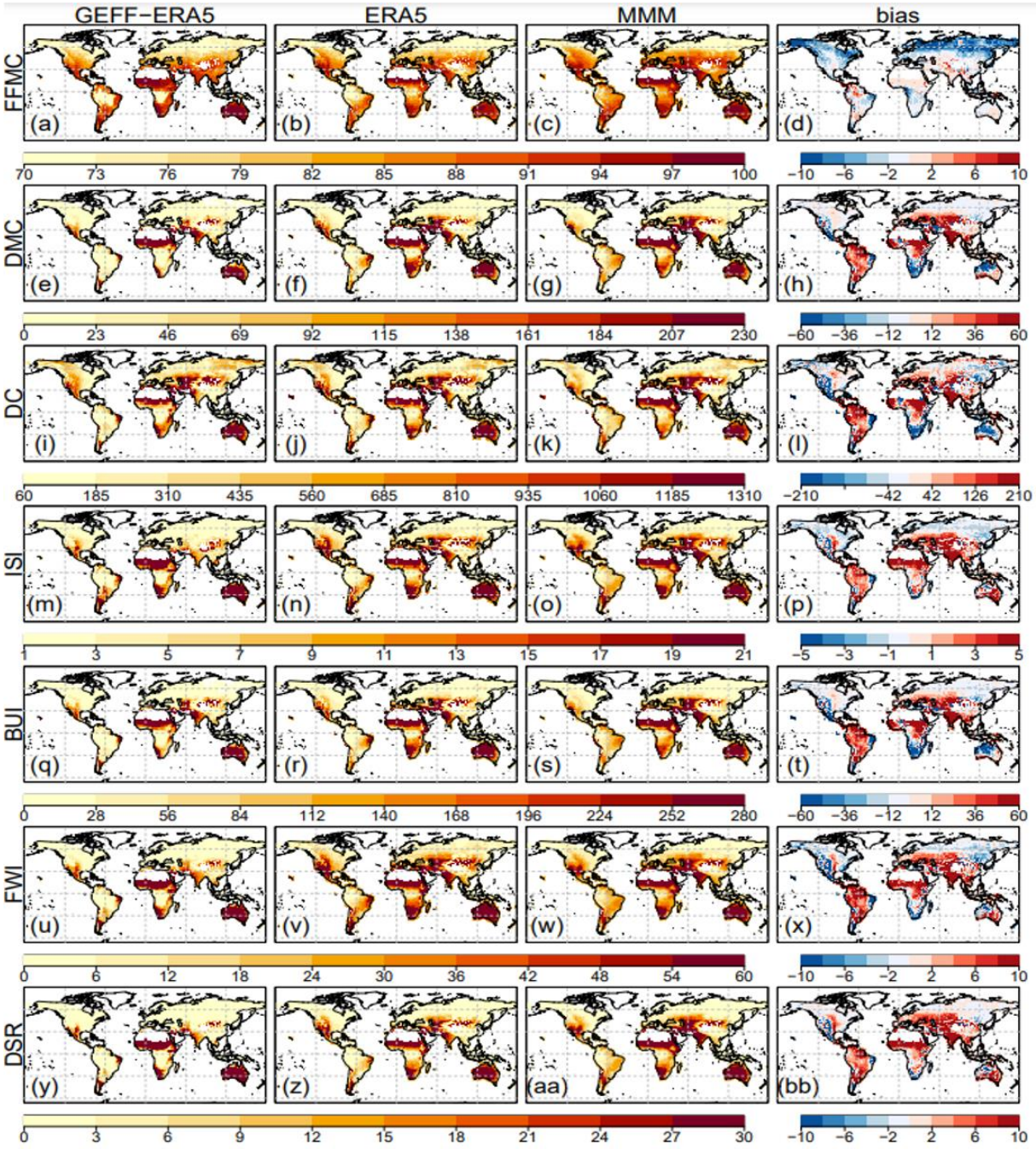

Figure 3: Annually-averaged values of monthly climatology of the 90th percentile for GEFF-ERA5 (first column), ERA5 (second column) and the CMIP6 multi-model mean (third column), and bias in the CMIP6 multi-model mean with respect to ERA5 (fourth column) for FFMC (a-d), DMC (e-h), DC (i-l), ISI (m-p), BUI (q-t),FWI (u-x) and DSR (y-bb). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias.

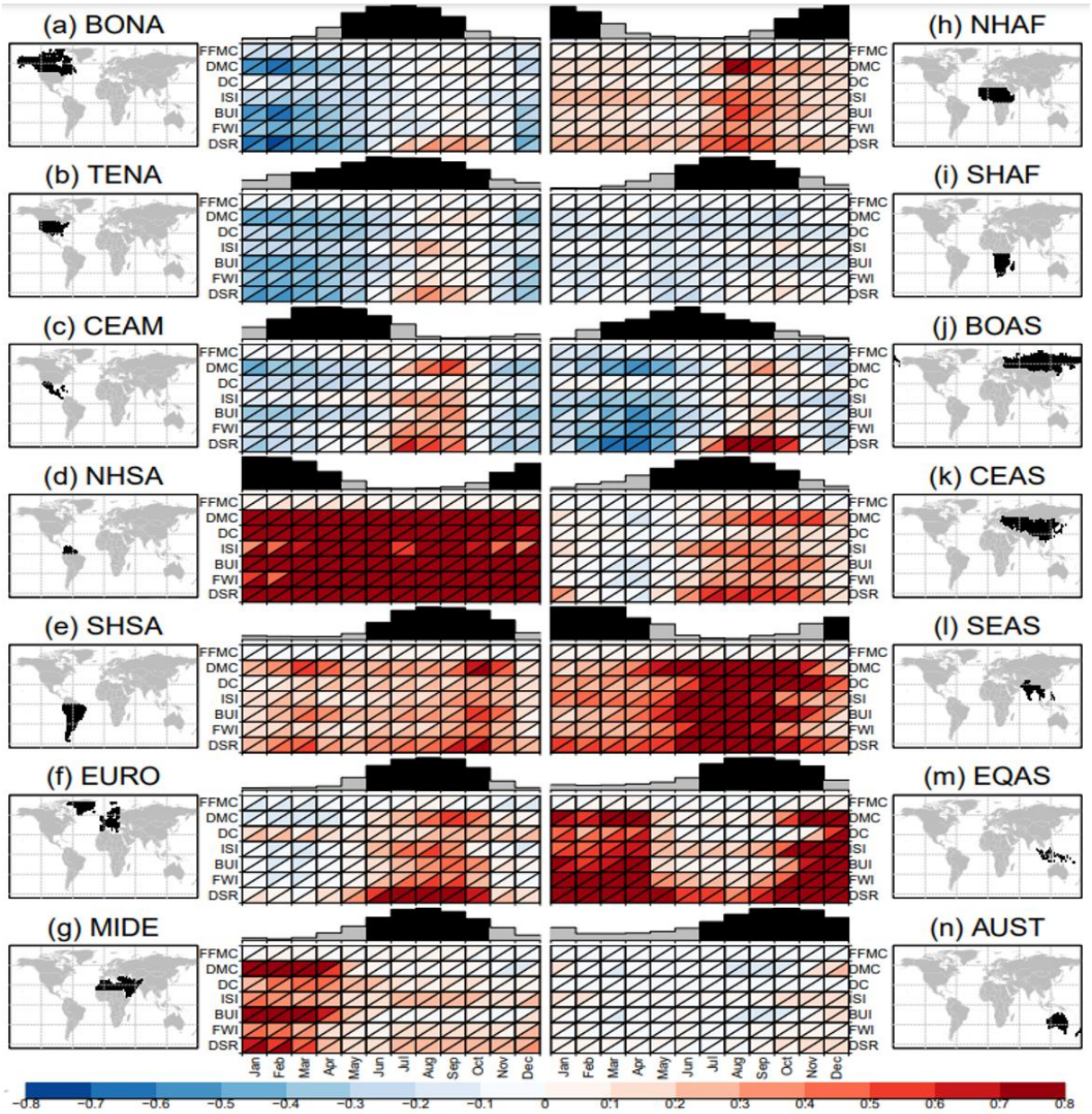

Figure 4: Bias in monthly means and 90th percentiles in seven CFWIS components simulated by the CMIP6 multi-model mean with respect to ERA5 across 14 GFED fire regions: (a) Boreal North America (BONA); (b) Temperate North America (TENA); (c) Central America (CEAM); (d) Northern Hemisphere South America (NHSA); (e) Southern Hemisphere South America (SHSA); (f) Europe (EURO); (g) Middle East (MIDE); (h) Northern Hemisphere Africa (NHAF); (i) Southern Hemisphere Africa (SHAF); (j) Boreal Asia (BOAS); (k) Central Asia (CEAS); (l) Southeast Asia (SEAS); (m) Equatorial Asia (EQAS); (n) Australia and New Zealand (AUST). Results show overall model performance, with blue shading indicating underestimation and red shading overestimation. The lower right triangle represents the monthly mean and the upper left triangle the monthly 90thpercentile. Bar plots show the average monthly-burned area for each GFED region, represented as a fraction of the monthly maximum. Black bars highlight months that constitute the 'fire season', defined as those months for which the average burned area is greater than 50% of the monthly maximum.

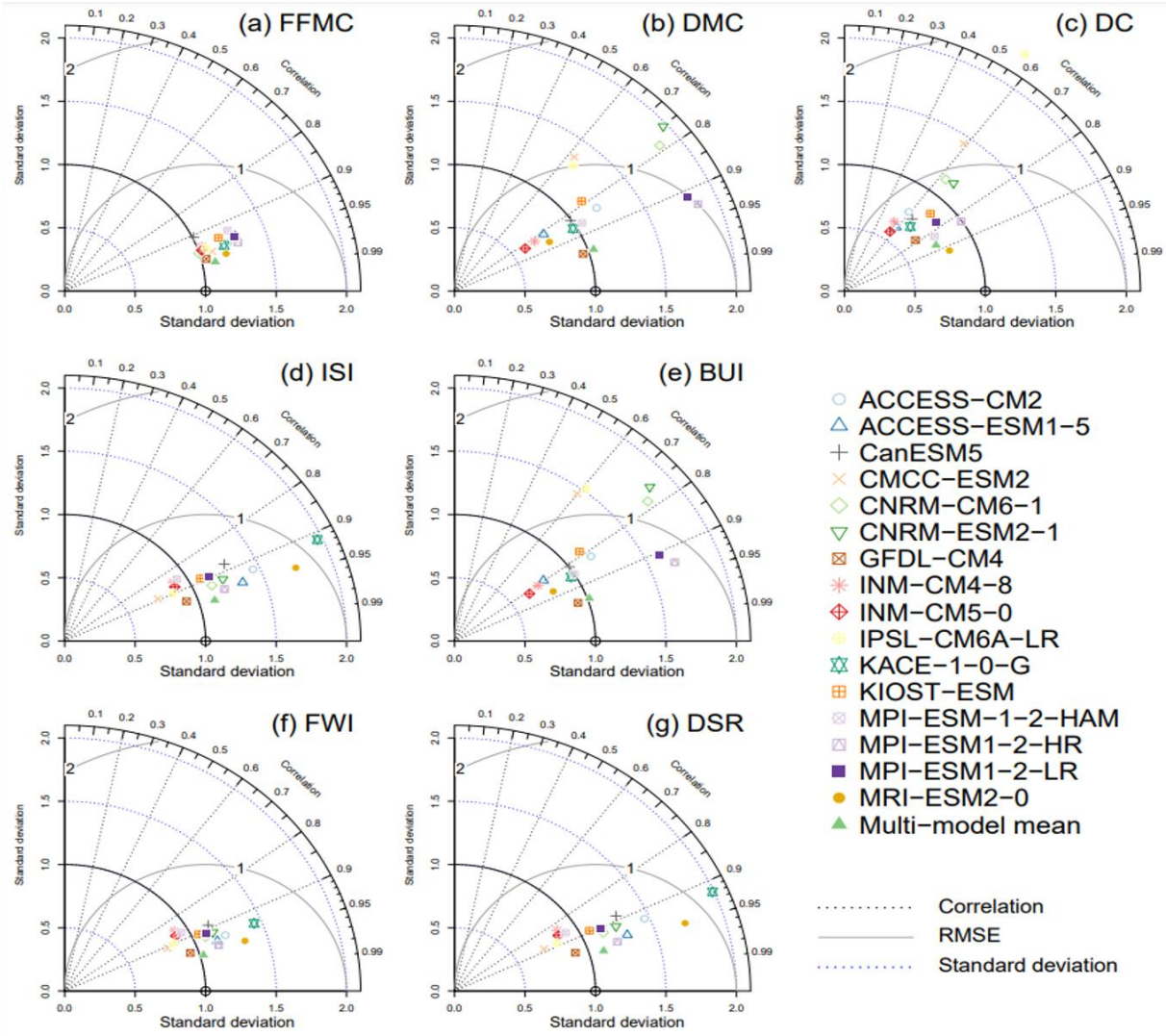

Figure 5: Taylor diagrams showing the capacity of 16 CMIP6 models to simulate annual means in the seven CFWIS indices. The correlation coefficient is plotted in relation to the polar axis, the normalised RMSE in relation to the internal circular axis, and the normalised standard deviation in relation to the horizontal axis. ERA5 is represented by an empty dot on the horizontal axis.

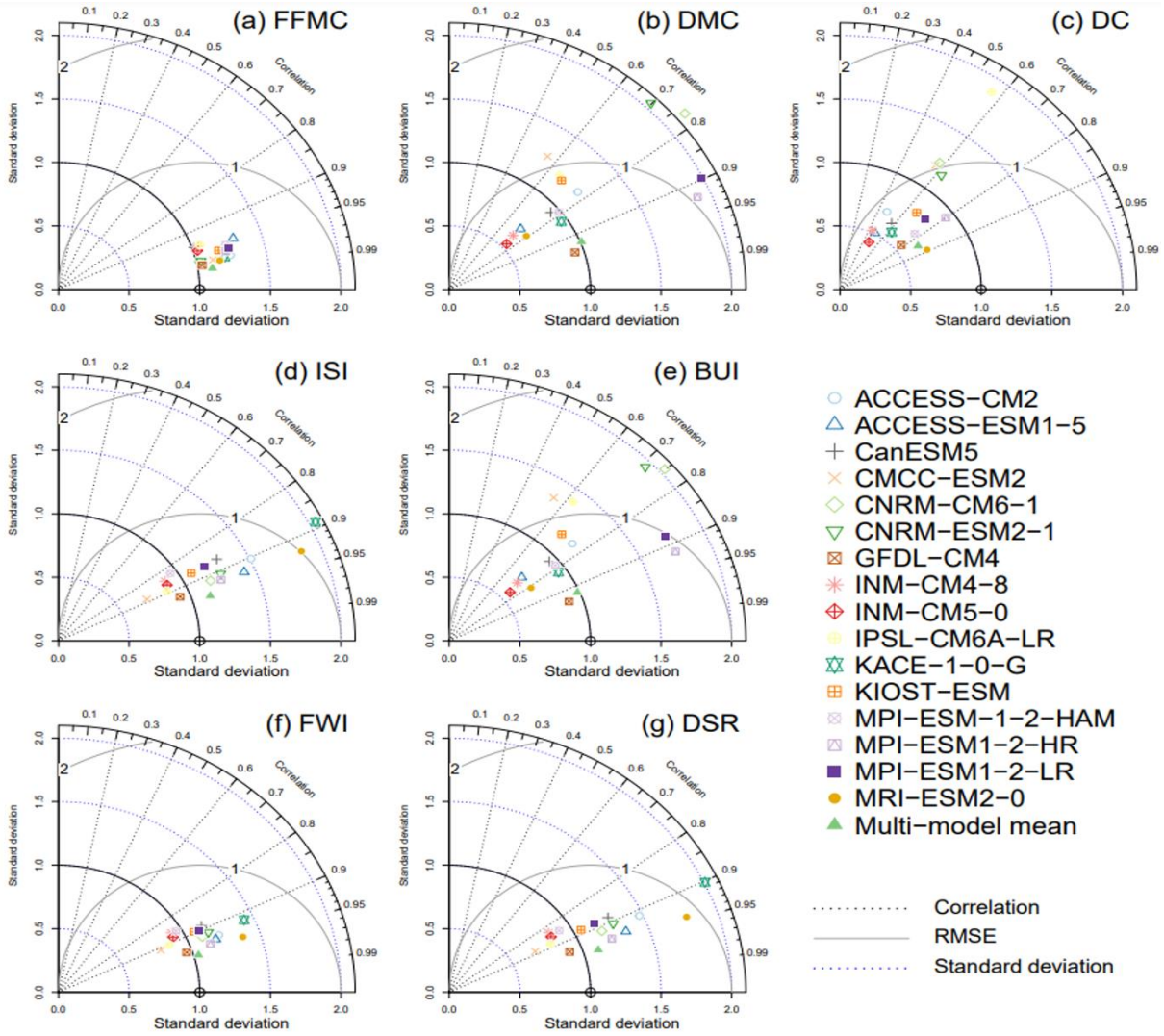

Figure 6: Taylor diagrams showing the capacity of16 CMIP6 models to simulate annual 90th percentile in the seven CFWIS indices. The correlation coefficient is plotted in relation to the polar axis, the normalised RMSE in relation to the internal circular axis, and the normalised standard deviation in relation to the horizontal axis. ERA5 is represented by an empty dot on the horizontal axis.

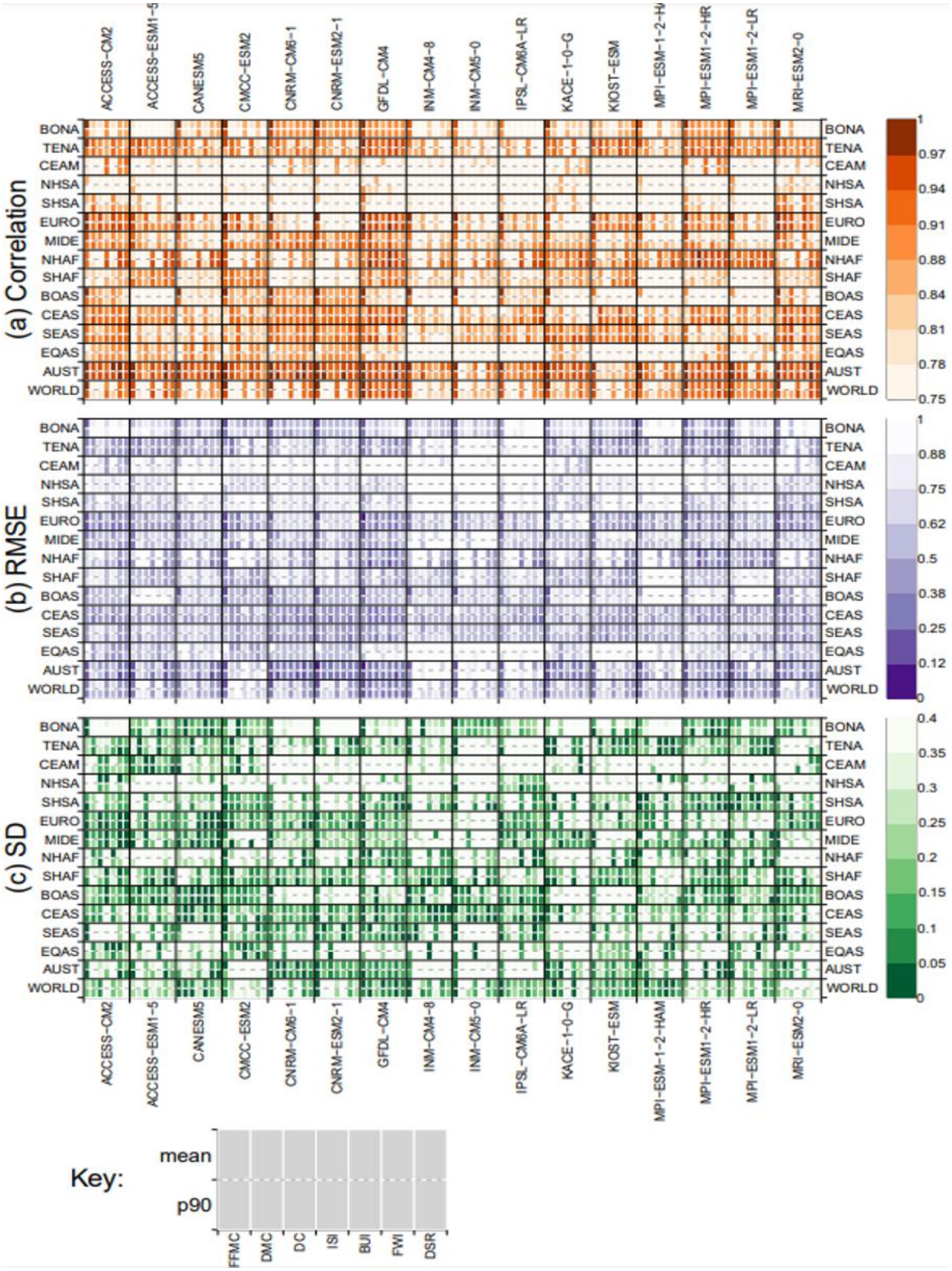

Figure 7: Individual CMIP6 model (a) correlation, (b) RMSE and (c) absolute log of the ratio of standard deviation with respect to ERA5 for the fire season mean and 90th percentile across each of the seven CFWIS indices and each of the 14 GFED fire regions. Darker colours show higher spatial correlations, and lighter colours lower. The fire season for each region is defined as those months for which the average burned area is greater than 50% of the monthly maximum (see Fig. 4).

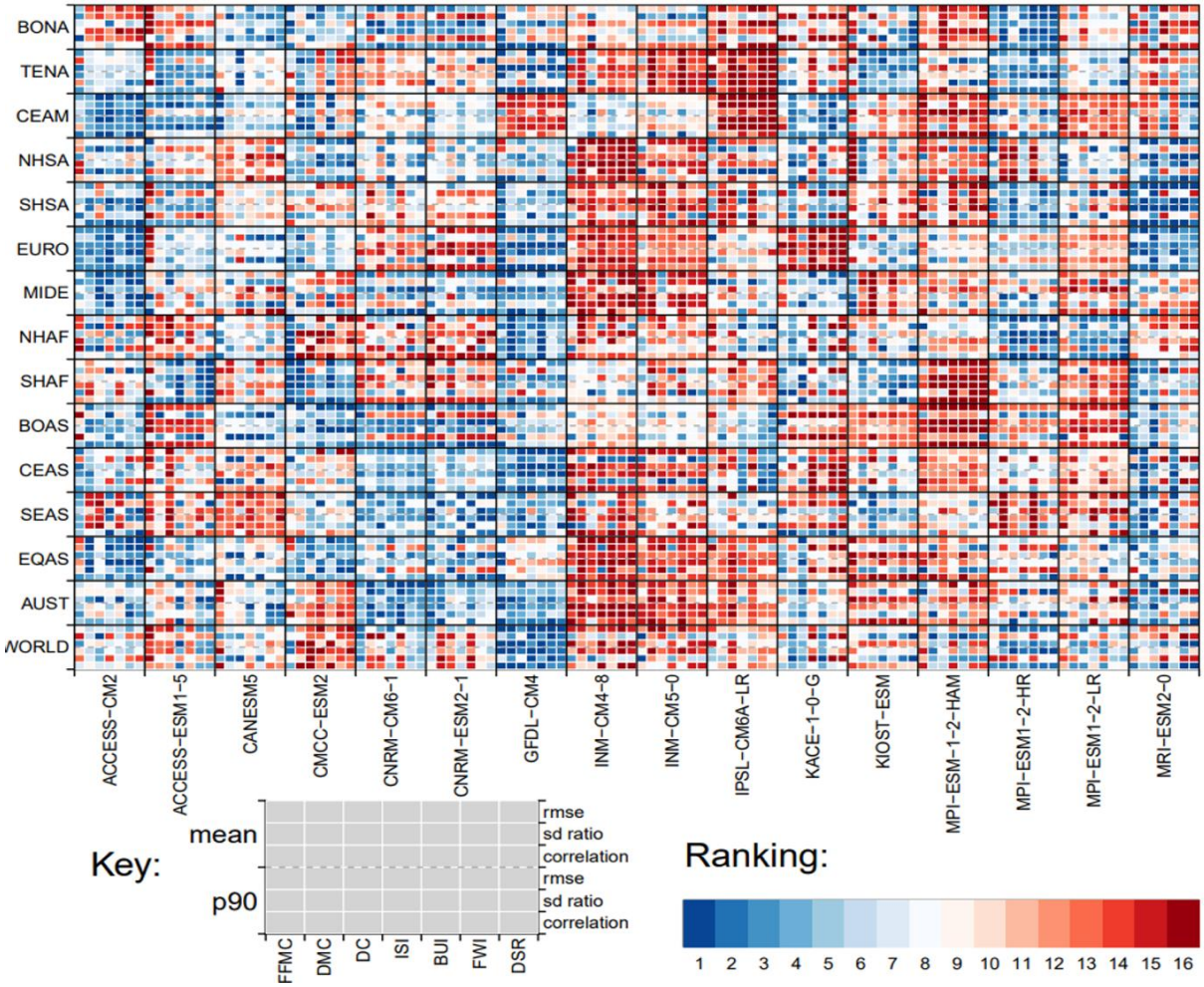

Figure 8: CMIP6 inter-model ranking for 14 GFED regions, 7 CFWIS components and 3 x 2 skill metrics (correlation, RMSE, and ratio of standard deviation for the mean and 90th percentile). For a given region and CFWIS component, models are ranked from 1 (the strongest) to 16 (the weakest) accordingly to a given skill metric. Blue (red) shading is thus indicative of strong (weak) model performance.

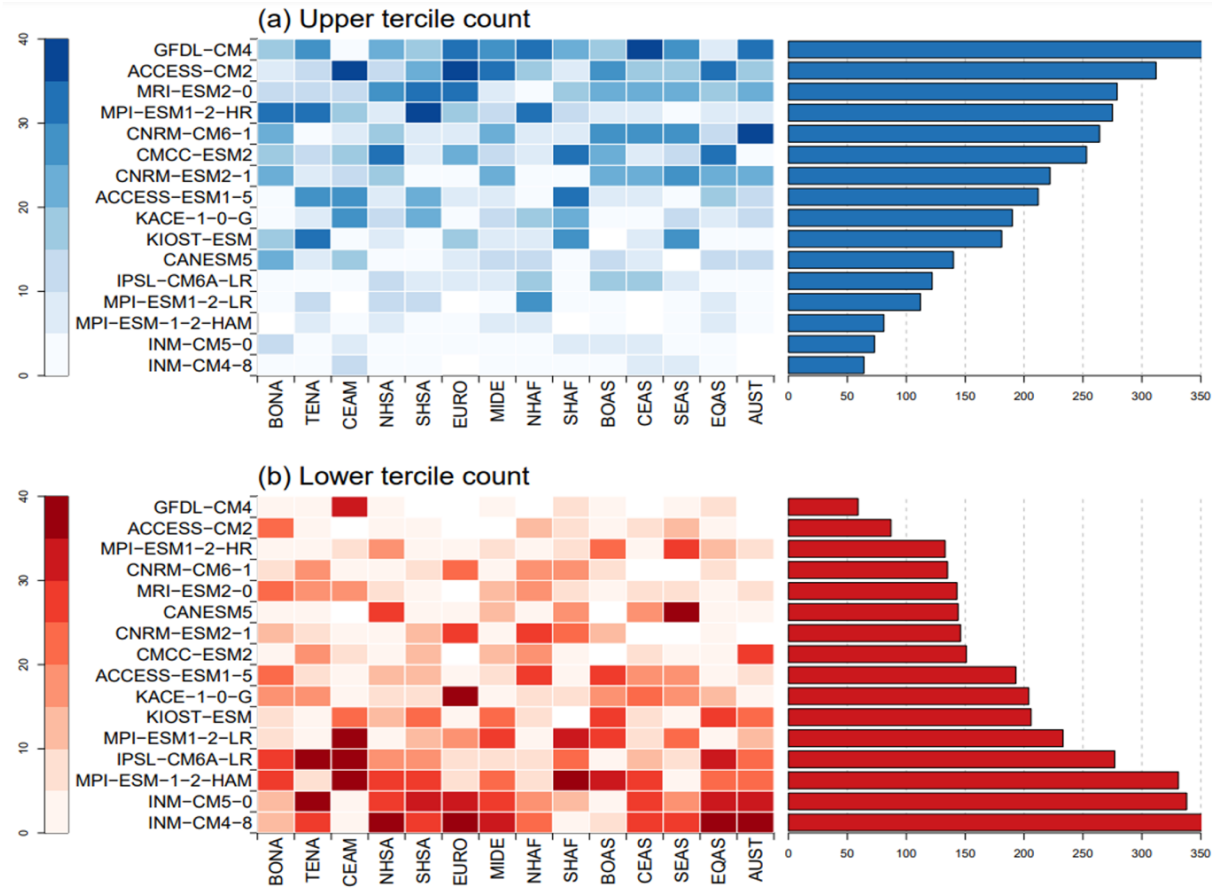

Figure 9: (a) Counts of the number of times that each CMIP6 model is ranked in the upper tercile (top 5) across all 7 CFWIS components and 3 x 2 skill metrics (correlation, RMSE, and ratio of standard deviation for the mean and 90th percentile). The grid (left) shows the breakdown of total counts for each of the 14 GFED regions. The bars (right) indicate the total count across all regions. (b) As (a) but for the lower tercile (bottom 5).

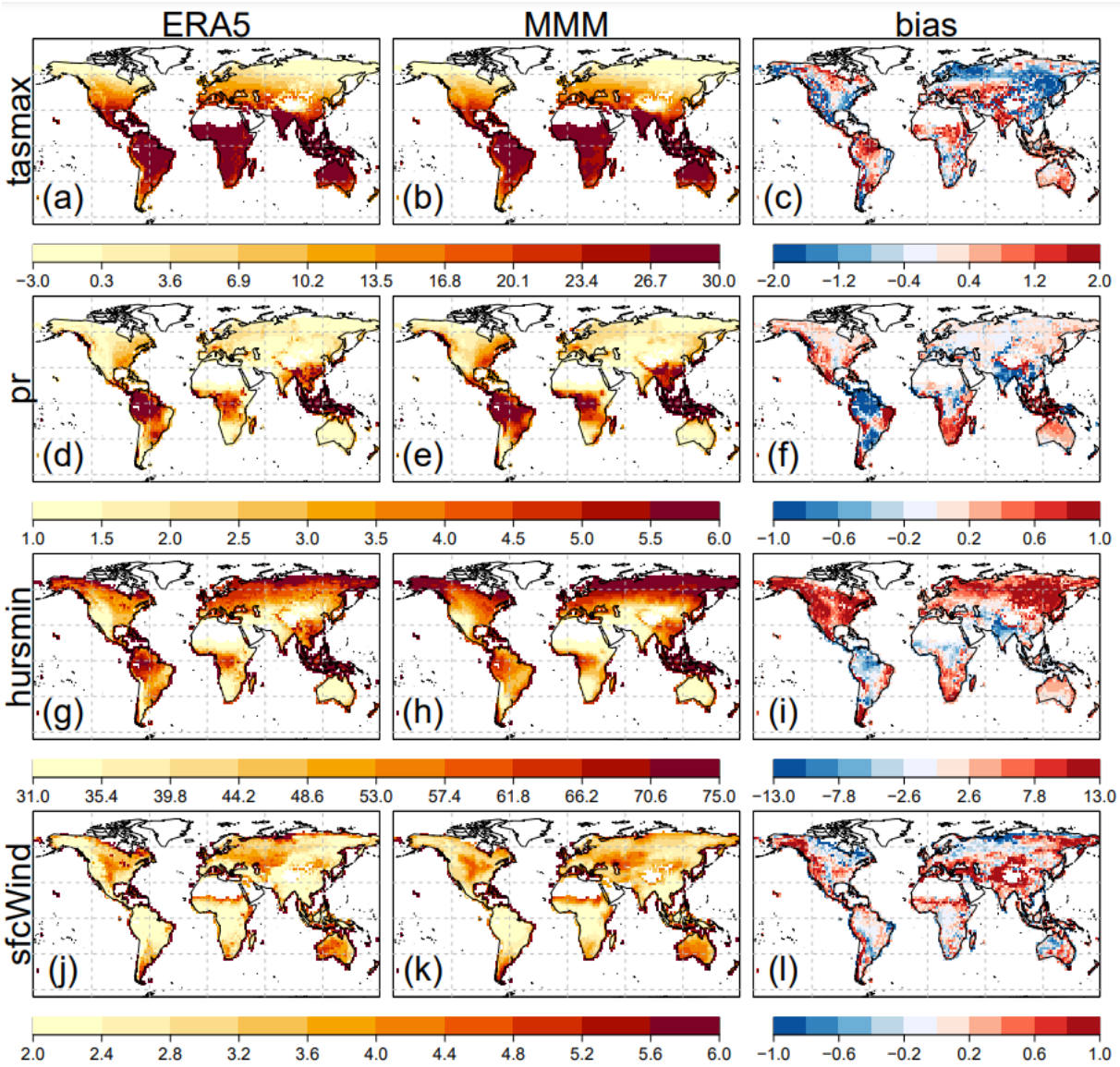

Figure S1: Annually-averaged monthly means for ERA5 (left column) and the CMIP6 multi-model mean (centre column), and bias in the CMIP6 multi-model mean with respect to ERA5 (right column) for maximum temperature (a-c), precipitation (d-f), minimum humidity (g-i), and wind speed (j-l). Lighter yellow colour represents lower danger and darker brown represents higher danger. Meanwhile, white colour represents lower bias and darker blue/red higher negative/positive bias.