Reviewer 1: This paper presents some case studies using the SIMA-MPAS GCM, which is a combination of the atmosphere and land component of the Community Earth System Model (CESM) and the nonhydrostatic MPAS dynamical core using the SIMA framework. The main simulations are performed at variable resolution with a 3km grid spacing covering the western US and a 60km grid spacing for the remaining globe. When compared to observations, SIMA-MPAS shows more realistic precipitation intensities, snowpack cover, and a smaller 2m temperature bias than simulations with the regional climate model WRF at a similar resolution (4km grid spacing).

The study provides confidence in the ability of SIMA-MPAS to produce realistic global climate simulations with variable grid spacing and the use of storm-resolving scales in regions of interest. This is a nice achievement considering the non-trivial coupling between the CAM physics package and the MPAS dynamical core (as described in the paper) and it encourages further research using this model framework. Therefore, I think the paper would be a useful contribution for the atmospheric modeling community.

We very much appreciate the reviewer's helpful input and the positive feedback. We have made substantial changes to the manuscript as enumerated below (blue colored). We have completed substantial revisions in the form of text revisions, figure modifications, additional simulations and analysis, to address each individual comment directly. We think the manuscript has been improved in many aspects.

However, there are some aspects of the paper that should be improved before publication.

Major points:

- A significant part of this work is the comparison of the model output with observational data, which is used as justification for the fitness of the model. However, the description of the different observational datasets is very brief (or non-existent). Without going through the references and performing their own literature research, many questions are unanswered to the audience. What are the respective resolutions of the datasets? How is the data obtained (i.e., what kind of product is it)? Are there any known biases (here, one gets the impression that they are the ground truth)? How do they compare to other observational products? Since the observations play such a significant role in this paper, I would expect a more detailed description of the respective datasets.

Thank you for the suggestion. We have added detailed descriptions of the observational datasets (one separate paragraph for each data product) in Section 2.2. As suggested by the second reviewer, we have also added the recently released Livneh precipitation data (Pierce et al., 2021) as another gridded observationally-based precipitation dataset to better account for extreme precipitation. The added text is also copied below:

"Detailed descriptions of the open-shared datasets used in this study are given below:

- CERES EBAF data products: we use gridded data from the Energy Balance And Filled (EBAF) product from the NASA Clouds in the Earth's Radiant Energy System (CERES), described by Loeb et al (2018). CERES provides high quality top of the atmosphere radiative fluxes and cloud radiative effects, as well as consistent ancillary products for Liquid Water Path (LWP) and cloud fraction. We start with monthly mean gridded products at 1˚ and make a 20 year climatology from 2000-2020.

- GHCN_CAMS Gridded 2m air land temperature: global analysis monthly data from NOAA PSL comes with resolution at 0.5 x 0.5°. It combines two large networks of station observations including the GHCN (Global Historical Climatology Network version 2) and the CAMS (Climate Anomaly Monitoring System), together with some unique interpolation methods (https://psl.noaa.gov; Fan and Van, 2008).

- PRISM observed data: the Parameter-elevation Regressions on Independent Slopes Model (PRISM) gridded observed data for daily precipitation and daily 2m air temperature is used at 4 km grid resolution (Daly et al., 2017; https://prism.oregonstate.edu/). Covering Continental U.S., PRISM takes the station observations from the Global Historical Climatology Network Daily (GHCND) data set (Menne et al., 2012) and applies a weighted regression scheme that accounts for multiple factors affecting the local climatology (Daly et al., 2017).

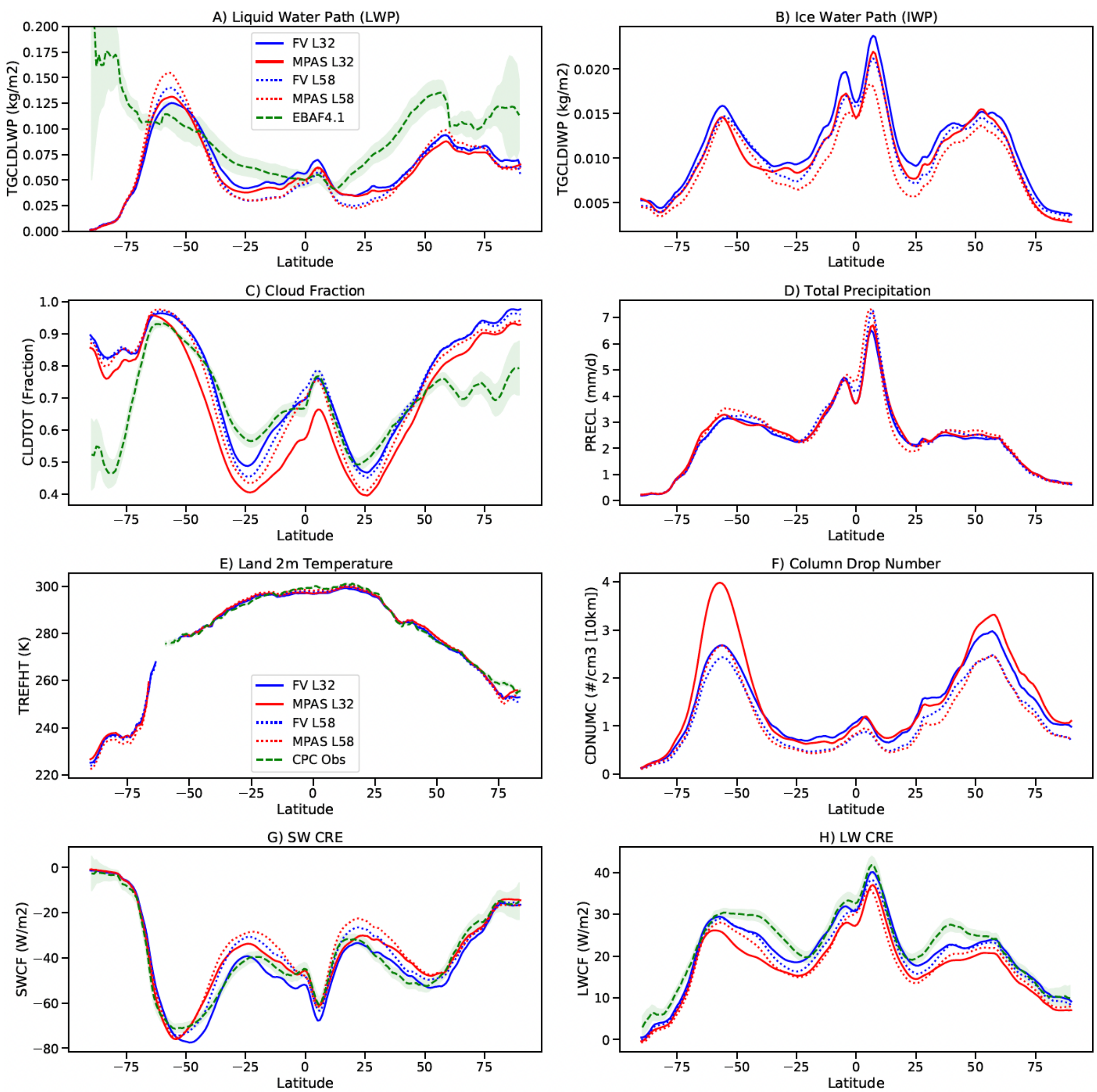
- Livneh gridded observationally-based precipitation dataset: in addition to PRISM data, to better account for extreme precipitation, a recently released Livneh precipitation data (Pierce et al., 2021; http://cirrus.ucsd.edu/~pierce/nonsplit_precip/) is also used for model evaluation. The data (~6km grid resolution) is shown to perform significantly better in reproducing extreme precipitation metrics (Pierce et al., 2021).

- Snow water equivalent (SWE) data over the CONUS: this is the observational data product we use for snowpack diagnostics. The data is available from National Snow and Ice Data Center (NSIDC) (at https://nsidc.org/data/nsidc-0719/versions/1). The product provides daily 4km SWE from 1981 to 2021, developed at the University of Arizona. The data assimilated in-situ snow measurements from the SNOTEL network and the COOP network with modeled, gridded temperature and precipitation data from PRISM (Zeng et al., 2018; Broxton et al., 2019).

- CONUS (Continental U.S.) II high resolution climate simulations: The WRF (Weather Research and Forecasting) nonhydrostatic model simulations we used for comparison are from Rasmussen et al. (2021) (accessible at https://rda.ucar.edu/datasets/ds612.5). Its horizontal grid resolution is 4 km with forcing from the mean of the CMIP5 model for both present (1996-2015) and future (2080-2099) mean climate, with hourly output. For the study region we focus on here (i.e. over the western US), the simulations provide a more realistic depiction of the mesoscale terrain features, critical to the successful simulation of mountainous precipitation (Rasmussen et al., 2021)."

- I find Fig. 3 confusing. For E) “Land 2m Temperature”, the legend says that EBAF 4.1 was used. However, in Sect. 2.2 you state that CERES EBAF was only used for cloud and radiation fluxes properties, whereas GHCN (which is mentioned nowhere except in Sect. 2.2) and/or PRISM was used for 2m temperature. So either the legend in the figure is wrong or the description in Sect. 2.2 is wrong. Please adjust and clarify. Also, if you have 2m temperature from PRISM and GHCN, why would you not include both into the analysis? Furthermore, I would suggest using a clearly different color for the observations to make it stand out more.

Thank you for catching this. We have revised the caption for Figure 3 clarifying that the observations for the land 2m air temperature is from GHCN_CAMS Gridded data. PRISM observation is not included here as it only covers the Continental U.S. region. For clarification, we have added the relevant data information in detail in Section 2.2. As suggested, we have updated Fig. 3 using a clearly different color (in green) for the observations (revised figure copied below).



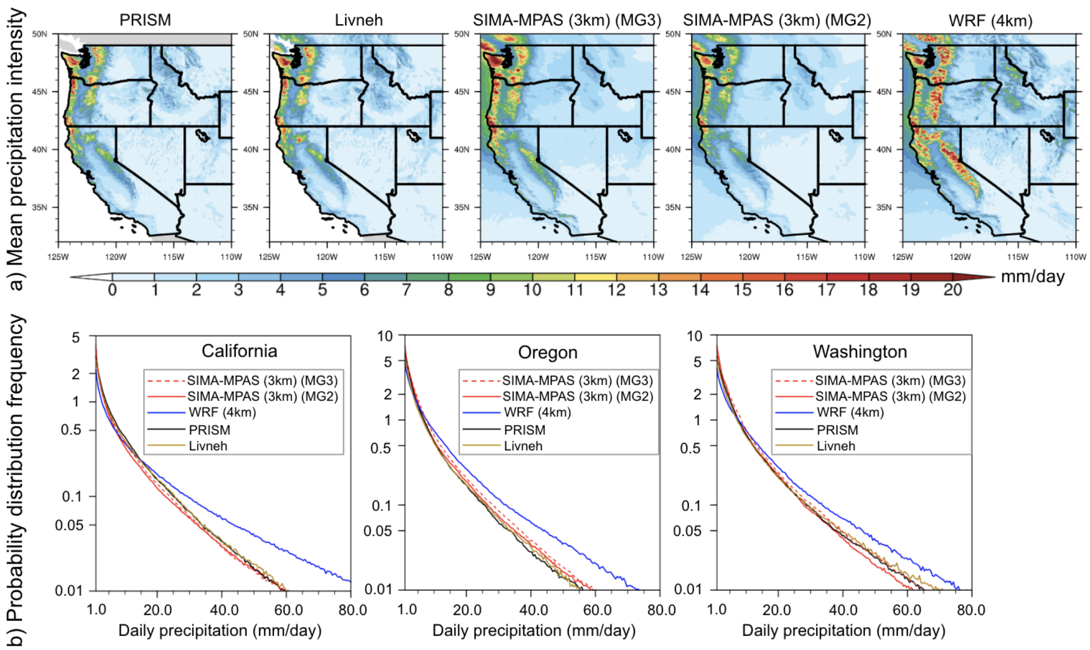
- I'm missing any information on the performed regridding for Fig. 5 and Fig. 6. Has the model and observational data been regridded for the analysis? If yes, to what grid and how? If no, how do you account for different grid spacings? This information is crucial for such an analysis (and the reproduction of the results!) and it should be clearly mentioned in the paper.

Thank you for pointing this out. Yes, we have regridded the model data to the same grid resolution as the PRISM observation (i.e. 4 km). For the regridd method and procedure, we first regridded CAM-MPAS data from unstructured grids to regular rectilinear lat/lon grids at 0.03 degree with ESMF software functions, and then regridded to the same grid spacings as the PRISM using the bilinear interpolation with relevant CDO command.

We have now added this information to Section 2.2: "For the regridd method and procedure, first CAM-MPAS data is regridded from unstructured grids to regular rectilinear lat/lon grids at 0.03 degree, and then the rectilinear data is regridded to the same grid spacings as the PRISM using the bilinear interpolation." and mentioned that "The SIMA-MPAS model data is regridded to the same resolution as the PRISM grid spacings (i.e. 4 km)." when it applies.

- SIMAS-MPAS (3km) with MG2 microphysics seems to perform very well in spatial representation of precipitation (Fig. 4) and daily precipitation frequency (Fig. 5) when compared to observations over 5 seasons. However, when only looking at one season for the comparison with MG3, the MG2 version underestimates heavy precipitation frequency. So there seems to be quite a bit of variability and this does not exactly provide confidence in the robustness of the results, especially for Fig. 6. Is one season really enough to conclude that MG3 performs better? Maybe MG3 would overestimate heavy precipitation frequency over all 5 seasons? You also state in the paper that this issue requires more investigation. Therefore, I'm not sure whether it's wise to include these results that prominently in the paper. Maybe these results would be better suited for the appendix.

Thank you for the suggestion. We agree that the variability for using one season in assessing MG3 performance. To better understand this, we have added two new simulations as we can (as each simulation is computation intensive) for another two wet seasons using MG3 microphysics. We have updated the results section 3.2.2, combining the three seasons in total for MG3. Overall, the precipitation statistics are well represented in SIMA-MAPS compared to observations both with MG2 and MG3. We do recognize that MG2 tends to underestimate heavy precipitation frequency in certain regions compared to observations, while MG3 shows a closer match in those cases with more intense precipitation produced. Changes are made to the manuscript. The updated Fig. 6 is copied below.



(updated) Figure 6: MG2 vs. MG3 microphysics used in SIMA-MPAS for the wet-season (Nov-March) precipitation over western US (1999-2002).

- Is the SST and ice sheet for Set A constant? Or are the forcings from different years or only from year 2000? How is the model initialized? From the paper it is not entirely clear to me how the mean climatology is obtained. Also, the last sentence at Section 3.1 does not make much sense to me (lines 224-225). Please clarify.

The SST and ice data for Set A are prescribed at the same yearly climatology (i.e. 12 months) with mean from the time period 1995-2005. We have clarified the referred sentence in Section 3.1 to “Simulations results are averaged over the five years output under the present day climatology (with SST and ice forcings from the mean of time period 1996-2005)”.

Minor points / typos:

- Lines 14-16 in the abstract read like SIMA is the atmospheric component of CESM, whereas from reading the introduction and the website, SIMA is just a framework that allows for the coupling of different components. Please clarify or reformulate.

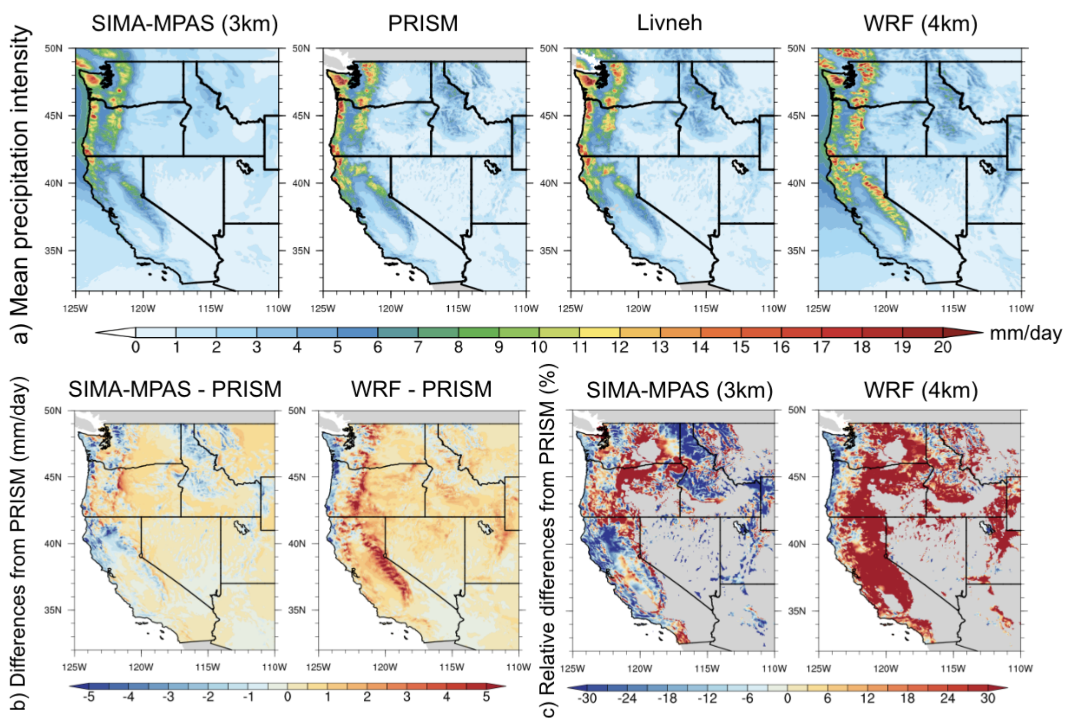
Thank you for catching this. We have rephrased the referred sentence to “This study uses a state-of-art storm-resolving GCM with a non-hydrostatic dynamical core - the Model for Prediction Across Scales (MPAS), incorporated in the atmospheric component (Community Atmosphere Model, CAM) of the open-source Community Earth System Model (CESM), within the System for Integrated Modeling of the Atmosphere (SIMA) framework.”

- Section 3.4 & Fig. 9: I would at least expect a sentence about the use of gravity wave drag parameterization for the different simulations (I assume the 60km uses one, whereas it's not really necessary for 3km), as this will likely have an effect on the strength of the jet.

Thank you for the suggestions. We note in the discussion of the simulations in section 2.2 under topography how the gravity wave drag scheme operates: The orographic gravity wave drag scheme in SIMA-MPAS (used in CESM2-CAM6) uses a 'sub-grid' orography to force the scheme. The 'sub-grid' orography is calculated for each grid cell from a standard high resolution (1km) Digital Elevation Model and used to drive the scheme. Thus the sub-grid orography forcing is small at 3km, and is larger at 60km. The theory being that more of the drag and waves are resolved at higher resolution. So the overall drag should be somewhat similar with scale, but partitioned differently between resolved and unresolved. We have added those explanations to the main text.

- I would reverse the color bar for Fig. 4 b) and c). In Fig. 4 a), red means more precipitation and blue means less. For the differences it is the other way around. I believe it would make the plots easier to read to reverse it for b) and c) (as you have done it in Fig. 8).

Thank you for the suggestion. We have updated Fig. 4 with reversed colorbar in b) and c). The revised Fig. 4 is copied below.



- The terms "non-hydrostatic" and "nonhydrostatic" are both used in this paper.

For consistency, we have replaced "non-hydrostatic" with "nonhydrostatic" throughout the text.

- The term SST is used without definition.

Thank you. We have added this information.

- Lines 122, 143, 149, 181, ….: “We would” instead of “We’d”

Thank you. We have corrected those wordings.

- Line 421: “vertical wind patterns” sounds like you have analyzed vertical winds. Maybe use “cross sections of zonal and meridional winds” or something similar.

Thank you for pointing out this. We have rephrased this accordingly.