General response to all reviewers:

We are very thankful for the constructive and thorough review. You all clearly took meaningful time to help improve the presentation and robustness of this work, and I hope our attempt to incorporate your comments shows that we value it. There are a couple of broad changes to mention, but please see individual responses to each comment below. Our responses are in blue.

General changes:

1. Methodological details have been removed from the Results and Discussion section and moved into the Methods section (Section 2).
2. We have attempted to improve the writing of the Results and Discussion section as each reviewer brought up that this section was a bit messy.
3. Figures 5, 7, 9, and 10 have undergone more substantial changes to either clarify the interpretation of the figure or provide more quantitative information.
4. More quantitative information and statistical tests have been added for much of the analysis (e.g. Figure 7, 8, 9, 10, 12, 13).
5. Many supplemental figures have been added as we want to keep the main manuscript structured to focus on the most essential points for brevity.

Reviewer 2

In their study, Stuivenvolt-Allen et al. present an evaluation of a historical simulation with WRF (NAC6), as part of the North American CORDEX-CMIP suite simulations. The evaluation is performed comparing with several observation based data sets (GPM-IMERG, SNOTEL, among others), and with a convective-allowing simulation (C404). The authors find that the new simulation (NAC6, forced by ERA5) exhibits several improvements with respect to a previous analogous simulation (NAC5 forced by ERA-Interim), especially in the representation of the diurnal cycle of precipitation, precipitation extremes, and snow.

The manuscript is clear and well written. The analysis and findings are of interest to a broad community. However, some improvements could be made before publication, as suggested in the comments below. I suggest Major Revisions of this manuscript in its current state. I hope that the comments below would be helpful to the authors.

General response:

We are very thankful for your thorough and constructive review. Addressing your comments about how our simulation design through time-slices may affect some of the slower-varying land-surface variables is very helpful to show for us and for the reader. We also think that including more reference datasets made this work more robust. Please see our line-by-line responses below.

General comments:

A. About biases.

Regarding the comparison to ERA5-Land (Figs. 2, 3, 4), it would be interesting to also see a comparison to other observation based data sets, like those from CRU, CHIRPS, CHIRTS, or similar. The authors could also comment on why it is better to compare to ERA5-Land instead of those observation based gridded data sets. Whereas temperature in ERA5-Land benefits from direct data assimilation, precipitation is a product from the model (as the authors write in the text), with some biases.

In general, the authors make a description of the biases, with some references to other papers where specific processes/mechanisms have been identified to explain some of the biases. However, the authors can include some comments or hypothesis about part of the biases identified in their study (see specific comments below).

Thank you for this point! We recognize the value of comparing model performance against additional observationally based datasets. Accordingly, we have added comparisons of temperature and precipitation bias for the three WRF simulations (C404, NAC6, NAC5) against PRISM and Daymet V3, described below.

ERA5 and ERA5-Land remain central to our evaluation because ERA5 provides sub-daily, physically modeled fields at full North American coverage. However, we are limited in the number and fidelity of observational datasets with full North American coverage at 12 km or finer resolution spanning 1980 to near-present. We therefore also use the Parameter-elevation Regressions on Independent Slopes Model (PRISM), a weighted regression scheme that spatially interpolates sparse station observations over CONUS. PRISM was also used in the published evaluation of C404, allowing direct comparison with that work. We additionally include Daymet V3, another spatially interpolated product derived from sparse station observations with daily coverage. Both PRISM and Daymet V3 have been extensively validated against station observations (Thornton et al., 2021).

These results are provided in the supplement (Figures S6 through S9) to complement our existing ERA5-Land evaluation. The overall bias patterns are consistent across reference datasets, and our conclusions are unchanged.

B. About "Extreme events" (section 3.4)

Section 3.4 about extreme events is interesting: the authors compare a set of specific weather events (Figure 11 and 13), and some statistics for TCs (Figure 12). According to the authors, these comparisons are possible since all model simulations (C404, NAC5 and NAC6) use spectral nudging by reanalyses fields. However, for some of the comparisons it is important to bear in mind that two different reanalyses are used (ERA5 and ERA-Interim), with different intrinsic resolutions (equivalent grid-spacing close to 0.25° and 0.75°, respectively). In this sense, one could say, for example, that part of the improvement in the simulation of a TC (NAC6 vs. NAC5) could be associated with a higher resolution used to produce ERA5, and not mainly to the higher resolution of NAC6.

Furthermore, the representation of features related with some "Extreme events" (e.g. the TC features in Figs. 11 and 12) could depend on the resolution of the ESMs or GCMs used for the planned NA-CORDEX-CMIP6 simulation suite (coarser than ERA5). In this case, it could happen that the simulated TCs with the NAC6 configuration could be of the same type as the NAC5 structure and statistics in Figs 11 and 12. In this context, the analysis in section 3.4 might not be representative of what would be expected in the NA-CORDEX-CMIP6 simulation suite.

I suggest the authors take these points into consideration in the analysis of section 3.4, and in the conclusions (e.g. lines 299-302).

This is a great point, and we agree that it should be stated and elaborated upon in our work. Here and in the supplement, we have provided an additional figure which shows the differences in Hurricane Ivan as represented by ERA-Interim and ERA5 (Figure S13). We have also included an ERA-Interim and ERA5-based analysis of the mesoscale convective system, which is provided in Figure S15.

ERA-Interim has a nominal horizontal resolution of approximately 0.7° (T255), slightly higher than the 50 km NAC5. Likely due to data assimilation methods, ERA-Interim displays a more symmetrical storm. However, the storm completely lacks a discernible eyewall and has very weak intensity metrics (minimum surface pressure, maximum precipitation rate and wind speed). Quantitative comparisons of these reanalysis-based results with the WRF simulations are now included in the manuscript. Please see lines 399-409 and 447-455 for examples.

C. About other capabilities and skills: interannual variability.

While sections 3.1-3.3 are devoted to the mean state, and mean seasonal and diurnal patterns, section 3.4 is devoted to weather systems. To complement these analyses, it would be interesting to see the mean states and/or their biases associated with interannual variations in the boundary conditions, e.g. associated with ENSO phases. The common time period 1980-2010, and the use of reanalysis data as forcing (along with spectral nudging) would allow to explore the statistics of temperature and precipitation during El Niño vs. La Niña. Or comparing composites derived from particularly cold vs. particularly warm SSTs in nearby regions. This kind of analysis could provide statistics for the mean states or the mean biases associated with different climate regimes, which could be relevant for the interpretation of the planned NA-CORDEX-CMIP6 simulation suite. To avoid going much further beyond the scope of the present study, a relatively simple analysis of composites (El Niño vs. La Niña) would suffice.

This is a good suggestion for a free-running simulation, but the spectral nudging of temperature, wind, and geopotential height above the boundary layer largely constrains the large-scale circulation and thermodynamic patterns associated with ENSO to match the forcing reanalysis by design. An ENSO composite would therefore mostly reproduce the ENSO signal already present in ERA5/ERA-Interim rather than reveal something about the regional model's own response to interannual forcing, limiting its diagnostic value for this evaluation run. Given this and the number of supplementary figures already added in this revision, we prefer not to add this analysis here but agree it would be a valuable direction if we were providing free running simulations.

Specific Comments:

L52. Please provide one or two examples of the similarity among simulations: model domain? time period?

This has been clarified, please see the new sentence starting on line 55 in the revised manuscript.

L58. This seems like a good place to refer the reader to Figure 1.

Done.

L64. Table 1: please provide some details about the physics in Noah-MP, e.g. dynamic vegetation, groundwater scheme, depth of roots, etc.

We have added a description of the NoahMP configuration:

"Within Noah-MP, we retain the default compiled configuration: dynamic vegetation is disabled with leaf area index prescribed from look-up tables, stomatal resistance follows the Ball-Berry scheme, runoff and groundwater are represented through free drainage, and four vertical levels are output."

Root depth is a value from the lookup table (MPTABLE.TBL) for USGS vegetation categories in NoahMP and this has not been modified.

L73-77. From the text, I assume no restart files form WRF are used, but a new simulation is started for each time slice: is this correct? Please provide a justification as to why no restart files are used, what would be the consequences for slow evolution fields (e.g. soil moisture, snow cover, and those associated with vegetation dynamics), and would be the potential differences in this sense with respecto to C404 (continuous simulation?).

Thank you for this comment — we think this is a more important problem to address and highlight in the manuscript and for our data users.

The real-world constraints of our computing allocation and project timeline made it infeasible to run the entire simulation in serial. Running the entire (relatively short) 43-year evaluation simulation would have taken an estimated 107 days of wall-clock time (disregarding any time spent waiting in the queue); it would have been even more unworkable for the 150-year GCM-driven simulations we are currently producing. Therefore, we ran the simulation in parallel in decadal chunks with 2.4 years of spin-up, giving us a significant speed-up. We have clarified this motivation in the text (line 94).

These long integration times result in part from the fact that adding more processors for a WRF simulation does not provide a proportional increase in speed. Processes like input/output (I/O) do not scale well.

This is also part of the reason we elected to use spectral nudging, as it ensures some synoptic consistency between the time-sliced simulations. We have added a supplemental figure to get a sense of the impact of these decadal "discontinuities." Figure RX displays two snapshots of geopotential height (contours) and specific humidity (shading) at 500 hPa for 2009-12-31_18:00:00 and 2010-01-01_00:00:00 — the time periods straddling the discontinuous time step. The synoptic evolution is very consistent and there are no clear discontinuities in the middle atmosphere's circulation and moisture (even though moisture is not nudged).

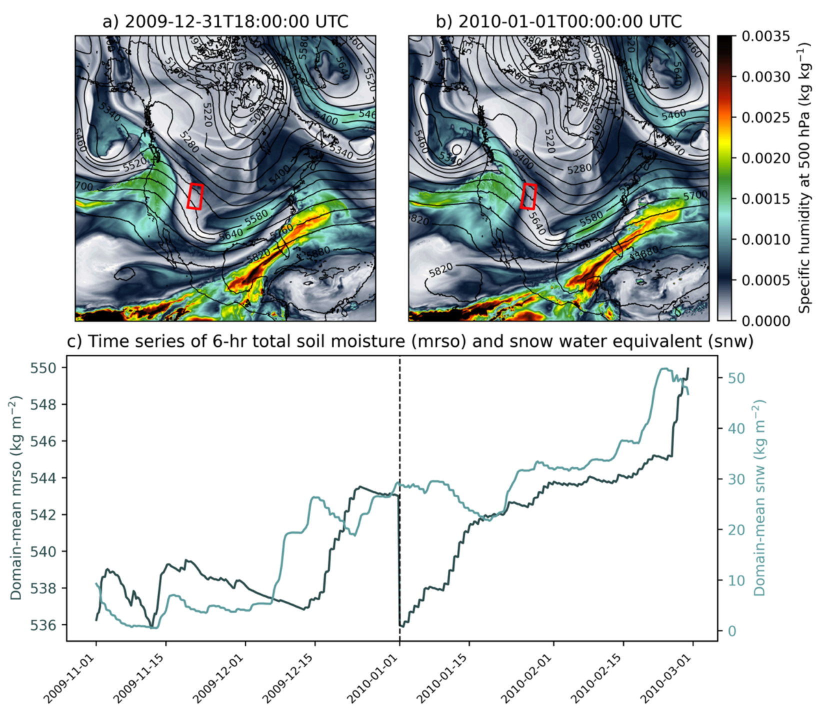

Figure R1 and S1. Snapshots of geopotential height and specific humidity at 500 hPa for (a) 2009-12-31 at 18:00 UTC and (b) 2010-01-01 00:00 UTC. (c) Time series of 6-hour total soil moisture (dark blue) and snow water equivalent (light blue) from 2009-11-30 through 2010-02-28.

As you highlighted, the bigger impact of this discontinuity can be seen in the slower varying soil moisture fields. Figure R1c shows an area average spanning the Colorado Rockies to produce a time series of 6-hourly total soil-moisture (mrso) and snow water equivalent (snw) across the temporal discontinuity. The snw fields have no noticeable step function change. The total soil moisture fields however do experience a step-function reduction at 2010-01-01 of 7 kg/m^2. While this looks substantial in the time series of 6-hourly data, it is a comparatively small perturbation in the context of larger monthly to seasonal changes over the entire simulation. When looking at the daily time series from 1980-2023, these discontinuities in soil moisture are barely visible (Figure R2 and Figure S2). Nonetheless, this is an important effect to mention in the manuscript so that users are aware of this slight discontinuity (This is now discussed in Lines 100-112).

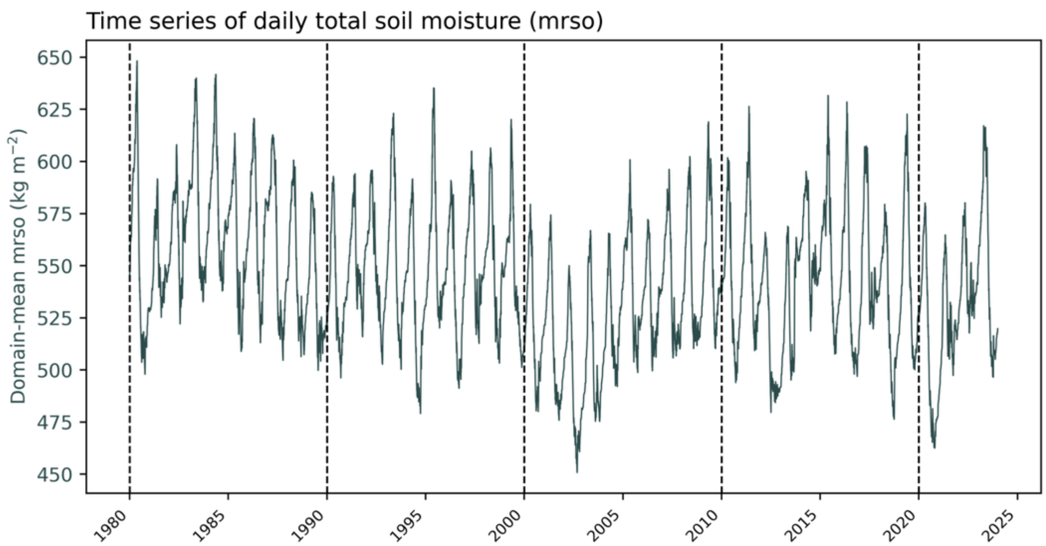

Figure R2 and Figure S2. Time series of 6-hour total soil moisture over the Colorado Rockies from 1980-01-01 through 2023-12-31.

L103-109 The C404 warm biases over complex terrain (e.g. western U.S.) could be more directly associated with lower valleys in the 4 km representation of topography, in comparison with the 9 km representation of ERA5-Land. The same reasoning would not directly apply to the comparison between NAC6 and ERA5-land. Could you please expand on these difference?

Thank you for raising this. Our understanding of the mechanism differs slightly from a relative-resolution comparison to ERA5-Land. Cold-air pooling and nocturnal drainage flows typically develop at scales well below 4 km, so even C404's convection grid does not fully resolve the physical process that would otherwise keep cold air trapped in subgrid valleys. This limitation is not unique to the comparison against ERA5-Land: although ERA5-Land's 9 km grid also cannot resolve these smaller cold pools, its data assimilation may still provide a more realistic representative temperature for the region than either WRF simulation captures.

More directly, C404 and NAC6 share the same warm bias pattern in this region despite their different horizontal resolutions (4 km vs. 12 km), while NAC5, run with the older Noah land surface model, instead shows a cold bias here. Since C404 and NAC6 are the two simulations that use Noah-MP, and NAC5 does not, this hints at Noah-MP, rather than grid spacing, as the more likely source of the warm signal in the Mountain West. We have revised the manuscript to state this more precisely (Lines 204-213).

L109-110. Do you have any hypothesis to explain the cold biases throughout central and eastern North America?

We have stated a hypothesis for the eastern cold bias, citing Zhang et al. (2025), who identify snow thermal conductivity parameterization in Noah-MP as the likely driver (lines 230 and 297-301) of the revised manuscript). The second instance expands on the mechanism: increasing thermal conductivity permits more atmosphere-snowpack heat exchange, raising winter temperatures by up to ~4 K in the Northeast. We also discuss the potential for this bias to be related to surface to canopy heat fluxes on Lines 290-301, as the bias seems to be strongly delineated by vegetation category (Figure S11).

L131. Could this be associated with too much cloudiness at night?

Persistent nightly cloud cover could cause part of this, but we did not evaluate cloud fraction, longwave/shortwave cloud radiative effects, or OLR in this study, so we cannot support or refute a cloudiness-driven mechanism with the current analysis. We note this as a direction for future work rather than speculate without supporting evidence.

L159-160. Any hypothesis as to why NAC6 exhibits a better picture of nocturnal precipitation, compared to NAC5?

We cannot fully disentangle two potential explanations. First, part of this may be attributable to the driving reanalysis data as discussed earlier. ERA5 has substantially higher vertical and horizontal resolution and it assimilates a lot more observations. Potentially, the better-resolved thermodynamic structure in ERA5 could improve nocturnal convective initiation and propagation. It is hard to disentangle these effects from the changes in physics schemes (particularly the microphysics and cumulus schemes). Finally, the Noah-MP land-atmosphere coupling could strongly impact local moisture entrainment and precipitation propagation.

L171. From figures 6c and 6d it is not too obvious that NAC6 exhibits a better eastward propagation than NAC5. Perhaps a version of these figures, in terms of fractional precipitation (w.r.t. to total daily precipitation) would help to better see the eastward propagation in NAC6.

Thank you for this comment. We think the format of the previous figure somewhat overstated the extent to which the eastward propagation is improved in NAC6. Please see the new Figure 5, which shows the peak diurnal precipitation hour in JJA represented as a difference from the IMERG climatology. NAC6 misses a lot of the nocturnal precipitation peaks near the Great Lakes and the Eastern half of the continent, but the timing of daily precipitation in NAC6 does not homogeneously occur in the afternoon like it does in NAC5 (Figure S10). The results and discussion for this figure have been revised accordingly on lines 275-283)

L181. Fig. 7a: I guess this refers to ERA5-Land data. Caption of the figure also needs corrections.

Fixed, thank you for noticing this.

L216. Fix the number of the figure: "Figure ??a" --> "Figure S4a" ?

Done.

L221. Change to "He et al. (2021)"

Done.

L212-220. The authors talk about comparison with SNOTEL data, and to figures 10c-e. However, in the current version of the PDF of the main text, Figure 10 only shows one graph ("CONUS maximum precipitation histogram"), with no reference to SNOTEL data. It seems that the authors are referring to Figure S4.

Correct! Apologies for this missed reference, it has been fixed. Please see the new Figure 10 and associated text.

L226. Instead of "all datasets use spectral nudging", I would say "all simulations use spectral nudging".

Fixed.

L247. Instead of "solar flux", I guess you are talking about "shortwave radiation". It would be better to visualize OLR, which is related with deeper clouds and precipitation.

We agree, but we do not have the needed temporal frequency of OLR from each of the simulations as output. We have added a line indicating this limitation – please see line 389-391.

L240-255. The results in figure 11 are interesting. Take into account, that forcing for NAC5 (ERA-Interim) is coarser than for C404 and NAC6 (ERA5), which probably has an effect in the resulting structures (e.g. in the gradients in the forcing fields) and features via the spectral nudging (e.g. compare the TC in ERA-Interim vs. ERA5).

Great point – we have added a supplemental figure and some discussion on this point. Please see our response to the general comment above.

L261-265. Results from figure 12 are very interesting. Are the differences between means (medians) statistically significant?

Yes – please see the new Figure 12 and the associated text describing it.

L275. Instead of "all datasets use spectral nudging", I would say "all simulations use spectral nudging".

Fixed.