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 1

Thank you Dr. Solman for your constructive review of this work. In particular, the suggested inclusion of more reanalysis-based reference data is useful context for the reader, and we think the manuscript has benefitted a lot from the other suggestions you've made. Please see line-by-line comments below and thank you for your time! Your questions have been surrounded in quotations, while ours are unquoted.

"The manuscript assesses the improvements in the most updated WRF evaluation simulation for the CORDEX-NA domain compared with a previous generation simulation. It also compares the simulated results against a convection-permitting simulation available for North America. The analysis is valuable and needed, as it represents the basis for identifying strengths and limitations of the most updated WRF simulation for the NA-CORDEX domain. The manuscript is very well written and conveys a clear message. However, in some passages the narrative seems to be somewhat disorganized and disconnected. This is particularly evident when analyzing snow features and extreme precipitation. I would recommend separating the case study analysis from the climatological assessment given the different approach followed."

There is a number of general and detailed comments, listed below, that should be considered to improve the overall readability and quality of the manuscript, before publication.

General comments

- The overall objectives of the manuscript are too vague. It is mentioned that the authors pursue assessing improvements from the corresponding previous-generation simulation. On what? What phenomena or climate feature are being assessed? I recommend clarifying on what features the analysis is focused.

Thank you for this comment, the objectives have been clarified by directly calling out which phenomena we focused on. Please see lines 45-50 in the revised manuscript.

- I recommend restructuring the analysis and differentiate the climatological analysis from the case studies analysis. The content starting in page 8 line 240 should be moved to a new section dedicated to case studies.

We agree that this section needed some better organization. We have created subsections under the Section 4 “Evaluation of extremes”, titled “Extreme precipitation”, “Tropical cyclones”, and “Case study of a mesoscale convective system”. After moving many of the methodological content from these sections into the actual Methods section, we hope that this all flows better.

- The discussion on the representation of snow seems to be written in a very different style as compared with previous sections. Please harmonize the writing style if possible.

This section has been revised, and an additional Figure has been moved from the supplement into the main manuscript to better quantify the snow biases. Please see the new section 3.3 and the new Figure 10 which is now properly referenced in the paper.

- Revise the discussion around Figure 10. There is a Figure 10 in the manuscript but it does not contain what is discussed in the paragraph referring to it. Moreover, one figure seems to be lacking (snow features).

Thank you for calling this out, we had a Figure reference error in the original manuscript. This has been corrected!

- The analysis focused on on case studies (TC and MCS) should include some reference to compare with. ERA5 may be too coarse but at least may provide some overall idea of the features evaluated. Particularly for the MCS, since the discussion is around the synoptic environment features, the comparison against ERA5 should be presented.

Thank you for noticing this, we think including the comparisons with ERA5 is useful for demonstrating the benefits and fidelity of these simulations. We have included ERA5 and ERA-Interim visual comparisons for these events in the supplement (Figures S13 and Figure S15) and the text now includes more quantitative comparisons. Note lines 399-409 and 447-455.

Detailed comments

- Page 2 lines 35-45: Are the physical parameterizations mentioned the actual selection used for the NA-CORDEX WRF simulation being assessed? This is not clear and needs to be clarified. As it is written, it seems that the only improvements in the WRF model are related with the physics schemes mentioned in the paragraph (which is certainly not the case). I recommend to clarify that the schemes highlighted are those used in the new simulation.

This has been clarified.

· Page 2 line 44: Who are you referring to as “our contribution” to NA-CORDEX? Who are “our”? Please clarify.

The author list indicates who is contributing these simulations. We have elected to keep this sentence as is.

· Page 3 line 57: Add CORDEX domain after North America.

Done.

· Page 3 lines 70-73: It would be recommendable to consider that the sensitivity tests for identifying the optimal physics selection may not end up in the same choice if the tests were done with different combinations of physical parameterizations, given the strong interdependencies among physical processes and how these interdependences are treated by different combinations of parameterizations.

We agree that our methods resulted in some untested combinations of physical parameterizations that could have performed differently or even better than our final configuration. We have inserted a statement to acknowledge that this testing was imperfect, but efficient and reproducible. Please see lines 80-94.

· Page 3 lines 75-79: The simulated period should be clearly indicated in this paragraph.

Added.

· Page 3 lines 75-79: The simulated period should be clearly indicated in this paragraph.

Done.

· Figure 1 is not referenced in the text.

Good catch, thank you! This has been fixed and the figure is now referenced when we describe the domain, the land use categories, and the lake model.

· Page 5 – lines 142-143: Should “from July 1 through August 31” be “from June 1 through Augus 31”, to be consistent with Fig. 5 legend?

Correct, this has been fixed.

· Page 6 – line 157: NAM region has not been defined. Please clarify.

We have added the lat/lon bounds for the NAM region and the northern and southern sub-domains used in this study to the text. They are also present in the figure caption.

· Page 6 – line 163: Are you referring to Figure 5d for IMERGE? If this is the case, replace Fig. 5d-g by Fig. 5d and refer to Fig. 5e-g when referring to the WRF simulations.

We have adjusted the reference for this statement on line 270.

· Page 6 – lines 163-164: I don’t see IMERGE rainfall rate being systematically smaller than WRF in Figure 5d. Can you identify where this feature is apparent?

Thank you for noticing this inconsistency. We had an incorrect reference to Figure 5d that has been corrected. Note the smaller magnitudes in the diurnal precipitation magnitude that are most visible over the Southeast United States and the Northern Sierra Madre mountain range in Mexico, and the provided reference (Hosseini-Moghari and Tang, 2022) systematically evaluates IMERG and comes to this conclusion.

- Page 7 - lines 173-178: Revise. It seems that some parentheses are lacking.

Fixed.

- Figure 7: Revise figure legend and panel titles.

Done.

- Page 8 - lines 201-205: This paragraph seems to be disconnected from the narrative. I recommend inverting the order of this and the following paragraph to ease readiness.

We have reorganized this section to provide a more direct description of the results with some discussion included in the paragraph of this subsection. Please see the new section 3.3.2.

- Page 8 – lines 209-210: The information on how Figure 9 is organized is already in the figure legend. It is not necessary to repeat it in the text. Please remove this line.

Done.

- Page 8 - lines 214-215: Why is the seasonality of the snow cycle not possible to be evaluated against the University of Arizona SWE data? Please clarify.

Seasonal results were provided per basin as a supplemental figure. We recognize that the corresponding supplemental figure was incorrectly referenced so that it was not clear, and have since moved this figure into the main manuscript for better viewing. Please see Figure 10 which includes seasonality of SWE compared to SNOTEL sites for a more robust, station-based comparison.

- Page 8: there is no analysis based on Figure 10, but what is the expected result. Figure 10 only displays the results from the simulations with no observation. Please provide an analysis of that figure, otherwise, it should be removed.

Addressed below.

- Page 8 – line 216: Please revise the reference to the Figure. Figure ??a is indicated.

Addressed below.

- Page 8 –lines 216-223: Revise this paragraph. No figure is displayed related with what is discussed here. There is a reference to Figure 10b/10c-e but Figure 10 only shows the histogram of monthly precipitation. Check carefully the consistency between the text and the figures before submitting the manuscript, otherwise it is impossible to provide a comment on the analysis.

I apologize for this figure reference problem. The supplemental figure looking at the seasonality of SWE compared to the SNOTEL station data is not properly referenced in the manuscript.

- Page 8: last paragraph: Figure 10 is referenced but referring to a different content. Please revise the list of figures included in the manuscript.

Same as above.

- Page 8 – Page 9 - lines 225-239: The analysis of extreme precipitation lacks a more in-depth discussion. It is written in a very disorganized way. Please re write.

We have changed this figure in response to another reviewer comment and have also rewritten the section. Please see the new Figure 11 and the associated text. This now includes a more nuanced view at extreme precipitation rates at 1-hourly, 6-hourly, and daily time scales. We have also provided percentile estimates of precipitation to better quantify the difference between the WRF simulations (Figure 11d-f)

- Page 9 – lines 236-237: Where is the observed histogram displayed to state that the C404 dataset closely matches station-based observational estimates? The observational-based results should be included in this figure. Otherwise, make it clear that you are just referring to referenced literature.

Please see Figure 8 in Rasmussen et al. (2023) for the comparison of C404 precipitation extremes with station based data. We have rewritten this sentence to be clearer, and have also added a recent reference from Schumacher et al. (2026) which further evaluates extreme precipitation in C404.

"Rasmussen et al. (2023) showed that the C404 dataset simulates extreme precipitation rates that closely match station-based observational estimates, rendering this simulation a quality benchmark for our precipitation comparisons (Schumacher et al. 2026).

- Page 9 – lines 247-255: To what extent reanalysis data (ERA5) can be used to compare against simulations? The discussion focus on highlighting that the C404 results are comparable with NAC6 results, but are these realistic? Consider including ERA5 reanalysis results just to have a reference to compare with.

We have included the ERA5 and ERA-Interim representation of the TC and the MCS case study. Please see the supplemental Figures S13 and S15 as well as the associated text and quantitative comparisons of the reanalysis-based case studies and their WRF simulation counterparts (Lines 399-409 and 447-455 are good examples).

- Figure S4 is not referenced in the text.

Fixed!

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).

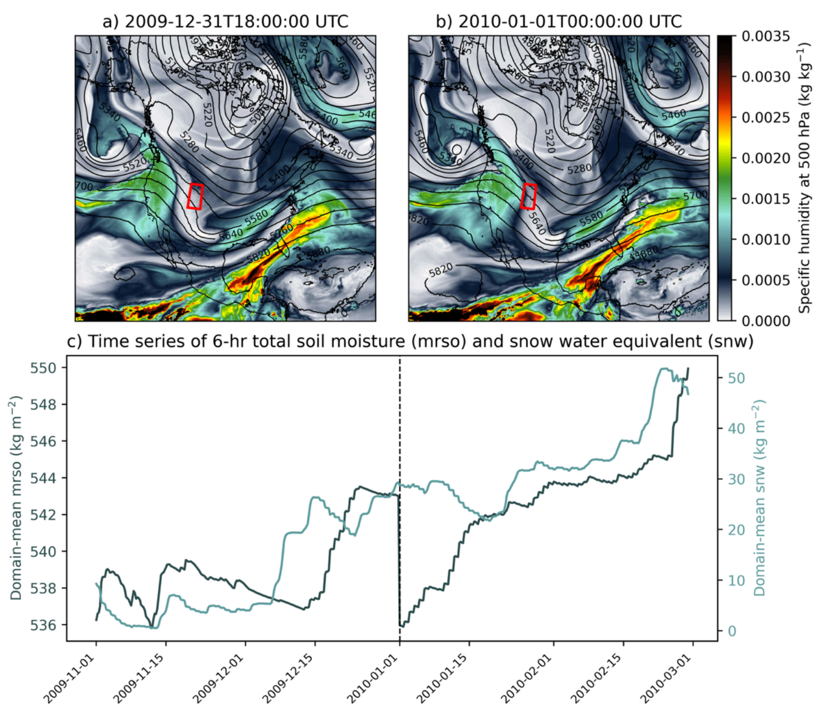

Figure R1 and S1. Snapshots of geopotential height and specific humidity at 500 hPa for (a) 2009-12-31 at 18:00 UTC and (b) 2010-01-01 00:00 UTC. (c) Time series of 6-hour total soil moisture (dark blue) and snow water equivalent (light blue) from 2009-11-30 through 2010-02-28.

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).

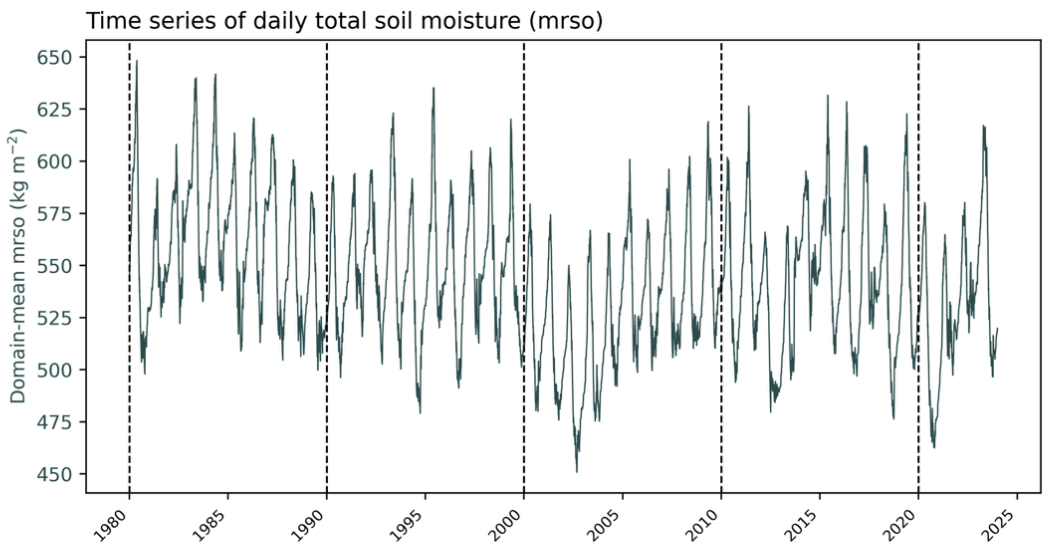

Figure R2 and Figure S2. Time series of 6-hour total soil moisture over the Colorado Rockies from 1980-01-01 through 2023-12-31.

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.

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 3

General response:

Thank you for your thorough and constructive review! The push for more quantitative rigor throughout the manuscript, in both figures and text, has substantially improved the paper and helped clarify the extent to which the updates from NAC5 to NAC6 actually matter. We address each comment in detail below.

Review of: The North American CORDEX-CMIP6 WRF evaluation run: comparing historical simulations from 25 km to convection-permitting scales.

Jacob Stuivenvolt-Allen et al., 2026. Preprint egusphere-2026-1638

General comments

The manuscript presents a comparative analysis of an ensemble of the NA-CORDEX-CMIP6 regional model, NA-CORDEX-CMIP5, and CONUS404. It provides valuable intercomparison information among the three datasets, as well as comparisons with other sources of information for the representation of precipitation, temperature, snow water equivalent, moisture flux, extratropical cyclones, and mesoscale convective systems, among other variables. The information presented is valuable for both the North American and the broader scientific community. However, several aspects should be clarified before publication, particularly the following key points:

1. The overall experimental design is somewhat confusing, as methodological details are introduced progressively alongside the presentation of the results. I therefore suggest a complete restructuring of the Methods section so that all methodological aspects are presented before moving directly to the Results section, avoiding the need to explain methodological procedures while presenting the results.

Thank you for this comment. We had organized the manuscript this way to try and eliminate the need for jumping back and forth from methods to results to conclusions. We recognize this might be a bit out of the ordinary for a manuscript, so we have extended the methods section and reduced methodological content from the results section when possible. In response to yours and other reviewer comments, we have re-written the results describing extreme precipitation and snow as well to hopefully improve the flow of the content.

2. Although this is partly a matter of scientific writing style, presenting the Results together with the Discussion makes the manuscript difficult to follow. In several instances, a topic is introduced for the first time within the Results section to establish the current state of the art, followed by the presentation of the results, and then by their discussion, making the manuscript difficult to read in a linear fashion. I strongly recommend separating the Discussion (interpretation and comparison with previous literature) from the Results (quantitative findings).

We appreciate this suggestion and recognize that separating Results from Discussion is more conventional. However, given the wide range of topics covered in this paper, we believe many readers will focus only on the sections most relevant to their interests (for example, someone evaluating tropical cyclones in the NAC6 data may skip content related to snow). In our internal review, we found that separating Results and Discussion led to more back and forth between sections for these targeted readers. The other two reviewers also commented favorably on the current writing style. For these reasons, we have elected to largely retain this structure, while improving clarity and flow of the writing, particularly in the Methods and Results and Discussion sections.

3. In general, the manuscript should be more quantitative in describing the results. Although the maps present numerical information, the text provides very few numerical values and instead relies heavily on qualitative adjectives to describe model performance or differences.

We have gone through the manuscript and added more quantitative data to both the figures and the text. Please see a few key examples on Lines 265, 292, 300, 308, 315, 328, etc in the revised manuscript. Quantitative information has also been added to figures when sensible.

4. Several figures are truncated over North America, particularly outside the CONUS region.

The domain of the figures has been adjusted to match the domain of the whole NAC6 simulation which follows the CORDEX specifications. All the figures have been updated.

5. Although the dataset is presented as having potential for ensemble applications, no analysis exploiting this capability is shown, suggesting that only the ensemble mean is presented.

None of the results in this manuscript are an ensemble mean. These results are a comparison of our “evaluation runs”. These are the reanalysis driven runs – and as stated in the discussion, the runs driven by general circulation models are nearing completion at time of writing.

6. I strongly recommend performing statistical significance tests to evaluate whether the similarities and/or differences found for each analyzed variable are statistically significant.

We have included statistical significance testing for Figure 13. Additionally, we have added more statistics such as percentiles, to quantify the difference between simulated results more clearly.

Specific comments are provided below.

Abstract

Please note that it is not appropriate to refer to "biases" when the comparison is made against other models that are themselves subject to uncertainty.

Thank you for raising this important point. We agree that ERA5-Land, while benefiting from data assimilation, is not a true observational dataset and carries its own uncertainties relative to station and satellite-based estimates. We have replaced "bias" with "difference" throughout the abstract, figures, and figure captions where comparisons are made against ERA5-Land. We have retained "bias" in select instances in the text for conciseness (e.g., "cold bias" rather than "negative temperature difference relative to ERA5-Land"), and have added a clarifying statement in the Methods section explaining this terminology choice and acknowledging the limitations of using a reanalysis-based product as a reference. We have also included more evaluations of these differences in our simulation against other observationally derived datasets (PRISM and Daymet) to provide some context on the role of reference dataset in influencing magnitudes and patterns of climatological differences in these WRF simulations.

L18. Correct the parenthesis: **CORDEX)(Giorgi** should be **CORDEX; Giorgi et al.**

Done.

At the end of the Introduction the authors state:

"Although these simulations employ different physics configurations and spatial resolutions, they are otherwise very similar, and allow us to assess the quality of the new NAC6 simulations with minimal uncertainty due to structural differences between modeling systems"

Please justify what is meant by "minimal uncertainty." It would be useful to specify what the structural differences actually are and how this uncertainty can be quantified.

Thank you for calling out this vague statement. We have rewritten it as follows to be more clear:

“Because all three simulations share the same WRF dynamical core, domain-nesting approach, and spectral nudging methodology, the remaining differences are largely isolated to physics parameterizations, resolution, and driving reanalysis, allowing us to assess the quality of the new NAC6 simulation with reduced confounding from structural differences between modeling systems."

Methods

L63. Please provide references for the following statement:

"Spectral nudging of temperature, winds, and geopotential height is applied above the boundary layer for scales >1000 km to maintain synoptic consistency while resolving mesoscale processes"

We have rewritten this sentence to clarify our methodology and intent, while also including the relevant references. Please see lines 67-70.

Since this section describes the methodology, please indicate which metrics and which components of temperature and precipitation were analyzed.

The methods section now includes a more clear description of our temperature and precipitation evaluations. Please see Section 2.4 and 2.4.1.

L71.

"The resulting test simulation with the smallest annual bias in temperature and precipitation was retained, and the next physics scheme was evaluated in subsequent tests".

From L65 to L70, it is not clear whether these correspond to the results of (Bukovsky and Karoly, 2009) or to results obtained in the present study. If the latter, the tested physics schemes and the resulting errors should be reported.

These results correspond to sensitivity testing conducted for this study, not Bukovsky et al. (2009) and we have revised the text to clarify this (lines 80-93). We have also added a description of the schemes tested for each parameterization category (microphysics, radiation, PBL/surface layer) alongside the existing cumulus scheme example. Scheme selection was based on visual comparison of annual and seasonal bias patterns against ERA5 for temperature and precipitation, rather than formal error statistics; the underlying model output from these 1977-78 test runs is no longer archived, so we are unable to retroactively compute quantitative metrics. Given that this screening was qualitative rather than quantitative, we have not included a full results table, as we do not want to overstate the rigor of the exercise.

Please provide references for the paragraph beginning at L75.

We have added a reference for the spectral nudging statement, but the rest of the content is unique to this work.

L85. Please indicate the data sources used for the prescribed variables and clarify whether these vary over time.

Done.

L88–91. A comparison with ERA5 could also be included to evaluate performance at the continental scale.

Correct, our NAC6 simulation is directly compared with ERA5 and ERA5-Land for this purpose.

L99–102. This material belongs in the Methods section. Also note that it is not appropriate to refer to "biases"; "differences" would be more appropriate.

The methodological content has been moved and the point about terminology is responded to above.

L105. The authors could consider applying an elevation correction based on a standard atmosphere to evaluate this effect.

We are actively working on bias adjustment for the evaluation runs and for our simulations driven by CMIP6 models (ongoing work) and consider this to be a separate and substantial task for inclusion into other manuscript(s). The purpose of this work is to highlight the evaluation simulation and benchmark it against its most similar products.

Section 3.1

P116. Please provide a more quantitative description of the results.

We have gone through and added more quantitative descriptions from the climatological differences.

Section 3.2

Additional regions of North America should also be evaluated, given that the dataset is presented for a domain larger than CONUS.

Our methods in section 3.2 present the diurnal evaluation for all North America in Figure S10. In the main text, we then focus on three well-documented, dynamically supported diurnal precipitation mechanisms (Southern Rockies, Great Plains propagation, and Southeast US) to provide a more in-depth discussion given the number of processes evaluated across the paper. We recognize the interest in diving deeper into each region and note that Figure S10 already provides panel-by-panel comparisons across the full NA domain for readers interested in other regions.

IMERG should be described in the Methods section.

Good catch, this is now described in the Methods section.

L180. This information should also be moved to the Methods section.

Done.

L185. Please explore the effect of an elevation correction.

It’s not clear to us how an elevation correction would resolve the cold bias near the Hudson Bay. We are actively working on bias adjustment for the evaluation runs and for our simulations driven by CMIP6 models (ongoing work) and consider this to be a separate and substantial task for inclusion into other manuscript(s).

Section 3.3

Although daily differences are computed, it is also important to calculate accumulated and percentage climatological differences. For example, a difference of 0.5 mm day^{-1} does not necessarily correspond to 182 mm year^{-1}, since precipitation does not occur every day.

Of course — we had assumed that this was implicitly understood, but we have provided seasonal accumulations and their biases to make this explicit. Please see the updated Figure 8 and associated text which quantifies the July through September difference in simulated precipitation for the WRF simulations when compared to IMERG.

L195. Please cite the reference supporting the statement that C404 is closer to the observations.

Rasmussen et al. (2023) is cited for this sentence. Please see their Figure 8.

L205. Please be more quantitative.

We have made section 3.3 much more quantitative.

The comparison with SNOTEL and the evaluation metrics should be described in the Methods section.

Done – this methodological text has been moved.

L216. Please correct the citation reference error.

Please quantify expressions such as "Too short", "too early", "Notable elongation", etc.

Done, we have quantified the difference in peak SWE across each dataset with respect to SNOTEL and these metrics are reported in Figure 10 and in the text.

Section 3.4

L236–239. The narrative is somewhat fragmented.

Section 3.4 has been largely rewritten and reorganized, please see the updated section.

L259. Add reference to TC category as defined by the US National Hurricane Center.

A reference to Ivan as a Category 3 TC was already in place, but we have added that the TC also reached Category 5 status over the Gulf of Mexico.

I recommend separating this section into two subsections, one devoted to tropical cyclones (TC) and the other to mesoscale convective systems (MCS).

Done.

Equations 2 and 3 should be presented in the Methods section, with all variables and terms properly defined.

Done.

Conclusions

It is not clear which dataset the following sentence refers to. The paragraph begins discussing NAC6, but the phrase "no uniformly superior dataset" creates ambiguity.

"However, for precipitation, the comparison is more nuanced: NAC6 exhibits a less severe dry bias in parts of the western US, and relative skill varies by region and season with no uniformly superior dataset".

This sentence has been rewritten and clarified. Please see lines 457-461.

Figures

Overall, it is difficult to follow the narrative of the manuscript because the figures are located far from the corresponding text. Some of the comments below may already be addressed in the intervening paragraphs; nevertheless, I include the following observations. In general, I recommend formatting the figures to occupy the full page width, as considerable page space is currently unused.

The figures have been embedded within the manuscript and given the full page width.

Figures 1, 2, 3, 4, 5, and 7. The NAC6 domain is cropped at both the northern and southern boundaries.

All the figures have been updated to match the domain of our simulations.

Figure 1. The purpose of panel (c) is not clear. If it is used later in the manuscript, this should be explicitly stated in the figure caption.

Fixed! We intended to give the reader a closer view of how 12 km resolves the spatial extent of specific land use categories (urban in particular), lakes, and mountains. Some new text describes this on line 75.

Figures 2, 3, and 4. Please use "differences" instead of "biases," since ERA5-Land is also a dataset subject to uncertainty.

Done.

Figure 5. Consider using IMERG as the first column, while the remaining columns show the differences between each model and IMERG. This would facilitate interpretation of the results, while the individual model fields could be moved to a supplementary figure.

This suggested edit has been incorporated.

Figure 6. The analysis region should be indicated in Figure 1, and the caption of Figure 6 should explicitly refer to this spatial domain defined in Figure 1.

We might be misunderstanding this comment, but the analysis region is displayed in Figure 5d and the caption for Figure 5 where it is displayed.

Figure 7. Similar comment as for the previous figure. Consider showing model differences while displaying only the ERA5 magnitude.

This suggested edit has been incorporated.

Figure 8. The analysis region should be indicated in Figure 1, and the caption of Figure 8 should explicitly refer to this spatial domain defined in Figure 1.

We have elected to just keep the description of the lat/lon box to keep Figure 1 uncluttered.

Figure 9. Consider adding a panel showing the reference SWE climatology.

We have added a supplemental figure with the SWE climatology. Please see Figure S12a which shows the SWE climatology from the University of Arizona data product. As Figure 9 has undergone some changes, namely focusing on SWE biases in UofA SWE gridcells that climatologically have at least 50 kg/m of SWE in JFMA, Figure S12b shows the resulting gridcells that are considered in our analysis. The purpose of this is elucidated in Section 3.3.2, from Lines 308 through 313.

Figure 10. Does this refer to **monthly maximum hourly precipitation**? Given the units, it would also be interesting to repeat the analysis for 6-hour, 12-hour, and 24-hour maxima, for example by adding new rows to this figure.

Thank you for this question, we think the description of these methods was too vague. Please see the updated Methods describing the extreme precipitation analysis, Figure 11 along with its descriptive text, which evaluates 1-hour, 6-hour, and daily precipitation extremes. We choose to include 6-hourly precipitation as this is a very common output/aggregation frequency for precipitation and can be better compared with other RCM or GCM based results.

We have also expanded this analysis to include percentile estimates of each dataset for the hourly, 6-hourly, and daily timescales.

Figure 11. Consider using satellite products as the reference, allowing all model differences to be calculated relative to a single external dataset instead of requiring visual comparison among the models.

For panels (a-c), we note that downward surface shortwave radiation is not a variable directly retrieved by satellite products, so a quantitative difference plot against an observational dataset is not feasible here. We instead use C404 as a benchmark, since a convection-permitting, reanalysis-driven simulation represents about the best achievable comparison at 12 km resolution. For additional context, we have added figures showing the ERA5- and ERA-Interim-driven simulations directly (new Figures S13 and S15).

For panels (d-f), we agree an external reference improves the comparison. We have added IBTrACS best-track minimum pressure and maximum wind speed as a reference in Figure 12d and 12f. Precipitation is not part of the IBTrACS best-track record, so we retain C404 as the reference for panel (e).

Figure 12. Same comment as for Figures 6 and 8. Consider performing statistical tests between samples to assess the statistical significance of differences between the datasets.

Done, this has been added. Please see the new Figure 12 and its associated discussion.

Figure 13. This is a very interesting analysis that deserves a more in-depth discussion in the manuscript. A similar analysis could also be performed for additional pressure levels and for IVT as complementary diagnostics.

We appreciate this positive feedback and have expanded the discussion of this case study in the manuscript. Please see Section 3.4.3. We chose 850 hPa moisture flux, moisture flux convergence, and temperature advection because these fields characterize the low-level moisture and thermal environment most directly relevant to convective initiation and organization for this case. IVT is a valuable diagnostic for atmospheric river and broader moisture-transport studies, but is less standard for characterizing MCS/derecho environments, which are more directly tied to low-level jets and boundary-layer moisture convergence. For this high-level case study, we believe the existing diagnostics sufficiently describe the storm environment without additional pressure levels or IVT. Additionally, we have added representations of this event in ERA5 and ERA-Interim as part of other reviewer comments, which help provide more context of the storm and its representation in reanalysis and the WRF simulations.