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.