Line numbers refer to the updated revision pdf with tracking changes enabled, not the original submission. The content of the added text, tables, and figures were detailed in our individual responses to the referee comments. We did not include them again in this response for brevity's sake.

We sincerely want to thank the reviewers and editor for their time and comments. Your suggestions have greatly contributed to the improvement of our manuscript.

Comments by RC1, RC2 in blue, our responses in black

Anonymous Referee #1 Comments

Overall, this is an interesting study comparing two commonly utilized catchment data sources. The analysis for the continental US (CONUS) appears to demonstrate statistically significant differences in the aggregate regarding temperature and precipitation. The differences shown are important, however greater attention to explaining the differences, and their significance would significantly improve the manuscript. The use of machine learning is not clearly articulated in the work and its significance is not yet clear. Greater attention should be paid to discussing the impacts of these differences on future modeling efforts as well.

The manuscript would also benefit from a clear statement of the goals of the research, i.e. is the goal to show that the two data sets are equivalent and therefore can be merged? Or is to identify where the two data sets differ and to explain why they are different, with the goal of adjusting one, or the other to allow merging? See line 45 for the first time this is made clear in the text. I would suggest clearly stating this in the abstract as well

- Revised the abstract to include clear statement of the goals.
- Edited and added text beginning at line 70 to address the results of the study along with the goals.

Line 103. How will this study address "uncertainties within the data sets? This is an unclear statement

- Line 129 clarifies the statement regarding the methodology to address uncertainties.

Lines 140-150. This is a confusing paragraph for those not intimately familiar with either data set. You state there are large discrepancies between the CAMEL SAC model ET and CAMEL-WB. Why is this important when comparing CAMELS to MOPEX, the goal of this work? Please expand this section and make it clear why these differences in ET with CAMELS is important to the goal of this work.

- Added text beginning at line 168 to line 184 to expand the section and clarify the differences in evapotranspiration in the CAMELS dataset, the modeled ET and the water balance derived ET.

- Added Figure 3 at line 184 to depict the differences in annual ET for CAMELS model and water balance derived along with MOPEX water balance ET for comparison.
- Added text at line 190 to 196 to explain the new figures 3 and 4.
- Added Figure 4 at line 196 to demonstrate the differences between the modeled ET when used to calculate the runoff (Precipitation – evapotranspiration = runoff) compared to the observed, measured runoff from the USGS.
- Text beginning at line 205 was edited to clarify the depictions in Figure 5.
- Text on lines 216 – 221 was moved to line 239 – 242.

Tables 4 and 5: These tables need far more explaining. The text indicates that they are internal variability of the two data sets, yet in each case, only a single mean is presented. The text is unclear as the tables do not provide the reader with any form of comparison here. The text indicates "within" the data sets, but the tables appear to provide "between" the data sets. Please expand section 4.1 to be clearer here.

- Added text at line 360 – 363 and at lines 365 – 368 to clarify the comparisons.
- Added Table 4 at line 389 to include the minimum, maximum and ranges for the median, mean, variance, standard deviation, and skewness for precipitation for both datasets. This table shows the internal variability of the datasets.
- Added text at lines 370 – 380 and lines 394 – 409 to explain results in Table 4.
- Added text at line 411 – 425 to explain results in Table 5.
- Added Table 5 at line 445 to include minimum, maximum, and ranges for median, mean, variance, standard deviation, and skewness for temperature for both datasets. This table shows the internal variability of the datasets.
- Removed sentence at line 452.
- Added text at lines 457 – 461 to explain Table 4 and its relationship with the coefficient of variation results for precipitation.
- Added line 469 to clarify the confidence interval estimates.
- More descriptive text added at lines 564 and 568.
- Text at 694 – 698 was added to explain Tables 7 and 8 (previously Tables 4 and 5), which were relocated to line 699 after the addition of new Tables 4 and 5.

Line 272. Does the fact that averaging over greater temporal scales reduce the dispersion a major finding here? it would seem like this would be an expected result?.

- Now line 452-453, was removed.

Line 323. It's not surprising that the variation in arid region precipitation is greater but what does " remain the most consistent" in the text mean? Consistent between data sets? Please be specific.

- Modified lines 508-509 for clarity.

Line 375: Some discussion of why these differences exist would be valuable here. . A bit of speculation will be helpful and appropriate.

- Added to discussion and machine learning, indicated below.

Line 630, Section 4.4 It is not fully apparent why machine learning validation was undertaking for this work and how it helps in the analysis. Please justify its use in more clarity.

- Added significantly to Section 3.3 to include further machine learning validation. Outlines the procedures.
- Text added at lines 300 – 337 to explain the machine learning models and simulated runoff using linear regression, random forest, gradient boosting, and support vector regression.
- Lines 339 – 342 removed.
- Lines 346 – 351 added and edited to explain binary classification.
- Modified Section 4.4 begins at line 836. Added new text, Tables 12, 13, 14 and 15 to present results from the 4 machine learning models and the predicted compared to observed runoff results.
- Added Figure 27 at line 892 which depicts the predicted vs observed runoff values and the 1:1 reference line for each of the four machine learning models for CAMELS and MOPEX.
- Numbered points in Conclusion beginning at line 1035 were renumbered and revised to include more defined results of the study.

Anonymous Referee #2 Comments

This manuscript presents a detailed comparison between two widely used streamflow and meteorological datasets for the continental United States, MOPEX and CAMELS, investigating their consistency and discrepancies from daily to annual scales. The study is based on a carefully designed statistical analysis and is relevant to the hydrological modeling and large-sample hydrology communities. The work is rigorous, and the results are clearly communicated and well discussed. I have a few remarks and suggestions for improvement that the authors might find useful.

In the abstract and elsewhere, the term ‘bias’ is used to describe the differences between MOPEX and CAMELS. Since bias is typically defined with respect to a reference or ground truth, it would be helpful to clarify that this refers to relative bias (i.e., systematic differences between datasets), rather than absolute error. While this becomes clearer within the manuscript, the abstract might mislead readers into thinking that MOPEX is definitively too warm or CAMELS too wet.

- Modified abstract incorporates these comments.

The manuscript could benefit from a more in-depth discussion of which dataset may be more reliable under certain conditions. Lines 685–687 touch upon this subject but could be expanded. For instance, CAMELS uses Daymet meteorological forcing, which could be potentially considered more reliable for regional hydrological analyses. However, its evapotranspiration values are derived from the SAC-SMA hydrologic model and, as the authors show, can exhibit implausible behavior. These trade-offs, i.e., between modern gridded meteorological inputs and model-based ET estimates, deserve a more explicit discussion to help guide dataset selection for different hydrological applications.

- Removed lines 994-997 and added lines 997-1025 to discussion section to expand.

Line 725: Please provide a citation for the NCDC COOP and SNOTEL datasets used in MOPEX. Additionally, a brief explanation of the nature of these data sources, including their observational basis and common sources of uncertainty, would help readers better understand the reliability and limitations of the meteorological data used in these databases.

- Citations added and discussion incorporated into lines 997-1025.

Figure 2: Could the authors clarify the meaning of the blue color in the map? It's not evident from the caption or figure description.

- Added explanation to figure caption.

Section 3.2.2: Please include references for all the statistical tests used (e.g., Fligner-Killeen test, Welch's t-test).

- Added references

Additional points

- Renumbered all figures/tables and captions to reflect the updated numbering.
- Fixed minor typos and punctuation.