Comments by RC2, Anonymous Referee #2 in blue, our responses in black.

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.

Thank you for your insightful feedback on our manuscript. We have carefully considered each of your suggestions and have made the following revisions accordingly. Specific changes:

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.

Thank you for your comment. We have made the necessary updates to the abstract and text, indicating that the bias is systematic for clarification.

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.

Thank you for your suggestions. The final paragraph in the discussion section was edited. The three sentences beginning on line 725 with "MOPEX estimated daily mean" were replaced with:

A key distinction between the MOPEX and CAMELS datasets lies in their methodological foundations. While both rely on ground-based weather stations as their primary source of meteorological input, they differ in how these observations were processed and spatially interpolated. Daily observations from a network of weather stations are made available by NOAA's National Centers for Environmental Information (NCEI). These data undergo quality assurance checks and processing prior to dissemination. However, uncertainties inherent to station observations remain due to limitations in instrumentation, despite adherence to established standards and calibration protocols, such as those outlined by the National Weather Service (https://www.weather.gov/coop/standards).

MOPEX used observed, gauge-based inputs of precipitation and temperature from Cooperative Observer Program (COOP) and Snowpack Telemetry Network (SNOTEL) weather stations to estimate mean areal values at the catchment scale (Schaake et al., 2006). For precipitation, MOPEX employed the Mean Areal Precipitation (MAP) methodology developed by the National Weather Service River Forecast Systems (NWSRFS), which combined an inverse distance weighting algorithm with monthly climatological means from PRISM to enhance spatial representativeness (Daly et al., 1993). In contrast, CAMELS derived its meteorological forcing data, specifically precipitation and temperature, from Daymet version 2, a gridded dataset that interpolates and extrapolates surface observations from Global Historical Climatology Network (GHCN) stations, including those from COOP (Thornton et al., 2014). Daymet version 2 employs a Gaussian convolution kernel interpolation method to produce spatially and temporally consistent values across the CONUS. These gridded values were then spatially averaged to the catchment scale within CAMELS (Newman et al., 2015).

Evapotranspiration values in CAMELS are estimated via the SAC-SMA hydrological model and as our results highlight, these model-based ET values can sometimes produce implausible behavior – such as overestimation or ET demand exceeding available precipitation. As MOPEX does not provide ET data, it can be calculated as a water balance residual. This approach benefits from empirical grounding, particularly in well-instrumented basins, but may suffer in regions with poor data quality or significant anthropogenic influences that are not explicitly accounted for in the water balance in addition to the uncertainties present in precipitation and runoff estimates (Carter et al., 2018).

MOPEX provides a longer historical record, which is valuable for evaluating long-term trends and hydroclimate variability. CAMELS is particularly well-suited for regional-scale hydrological analyses and climate sensitivity studies, especially in areas where gauge coverage is minimal or where spatial variability in meteorology is high (Addor et al., 2017; Newman et al., 2015). For spatially extensive or gridded analyses where consistency and meteorological realism are priorities, CAMELS offers advantages. Thus, when using daily data, the choice between CAMELS and MOPEX depends on the application.

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.

We have added the citations which are indicated below along with a brief description of COOP and SNOTEL, that is incorporated into the previous response above.

Operated by National Resources Conservation Service, US Department of Agriculture. Snowpack Telemetry Network (https://www.wcc.nrcs.usda.gov/factpub/sntlfct1.html).

Wuertz, David; Lawrimore, Jay; and Korzeniewski, Bryant (2018). Cooperative Observer Program (COOP) Hourly Precipitation Data (HPD). NOAA National Centers for Environmental Information. doi:10.25921/p7j8-2170.

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.

We have added the following to the caption of Figure 2

The blue colors in southern Florida represent regions within the tropical climate group, which is not represented in this study.

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

References were added for all applicable statistical tests.

Aho, K.A. (2013). Foundational and applied statistics for biologists using R. Chapman & Hall/CRC Press, 596 p.

Fligner, M.A. and Killeen, T.J. (1976). Distribution-free two-sample tests for scale. Journal of the American Statistical Association, 71(353), 210-213. https://doi.org/10.1080/01621459.1976.10481517

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J. (2020). Statistical methods in water resources. USGS Techniques and Methods 4 – A3.

Welch, B.L. (1951). On the comparison of several mean values: an alternative approach. Biometrika, 38(3-4), 330-336. https://doi.org/10.1093/biomet/38.3-4.330