Comments by RC1, Anonymous Referee #1 in blue, our responses in black.

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.

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

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

The abstract was revised to read

This study compares two large hydrometeorological datasets, the Model Parameter Estimation Experiment (MOPEX), and the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS), with the aim of quantifying differences that might impact their mergers. This comparison focuses on 47 shared watersheds within the continental United States spanning daily, monthly, seasonal, and annual scales for the overlapping water years of 1981-2000. Results indicate significant differences between the datasets at daily timesteps, highlighting the challenge of high temporal resolution data reconciliation; however, compatibility markedly improves with temporal aggregation at monthly, seasonal, and annual scales. Systematic biases are evident, with MOPEX showing a warm bias for temperature and CAMELS displaying a wet bias for precipitation. Studies considered monthly or longer trends may use either or both datasets, as the biases are not significant for low-temporal resolution analyses. Studies focusing on high resolution hydrological characteristics, such as daily precipitation events, the frequency of wet and dry days per month or single-basin dynamics, may require a statistical bias correction to ensure accuracy. The variability between the datasets is comparable to the variability within each dataset and is neither a useful criterion for dataset selection nor a barrier to potential merger. In any case, model outputs should be calibrated against observational refence data to account for systematic errors. Statistical analyses demonstrate that both datasets are representative of climatic conditions, trends, and extreme events. Our findings validate the results of previous research employing either dataset. Furthermore, this study serves as a foundation for the merging and extension of MOPEX and CAMELS datasets without any alterations, providing a comprehensive, long-term dataset suitable for hydrological modelling and climate analyses while maintaining comparability across basin and temporal scales.

Beginning at line 56, after the sentence ending with "watershed-based datasets" we removed the remaining text (lines 56 to 59) and clarified the study to read as follows:

Our findings show that while MOPEX and CAMELS exhibit systematic biases, they can still be merged or reliably compared without requiring corrections beyond smaller time scales (i.e. a single day, month, or season). Statistical adjustments to daily data depend on study objectives, as no single method fits all needs. Raw data or direct model outputs typically require bias correction, and we intend for our results to help researchers determine necessary adjustments using appropriate methods, including equidistant quantile matching (EDCDFm) for temperature and quantile delta mapping (QDM) or PresRATe for precipitation (Lehner et al., 2023a; Pierce et al., 2015).

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

Thank you for your comment. We have clarified the statement as follows

This study provides researchers with detailed analyses regarding the uncertainties within the datasets and between them for a 20-year period through quantitative measurements of dispersion, distribution, central tendency, interval estimates, and statistical tests.

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.

Thank you for the suggestion. We have expanded the section and included two additional figures to aid in clarification. The following text was added immediately after Figure 2, beginning at line 140. The addition also includes two figures.

Terrestrial evapotranspiration is difficult to measure directly but can be evaluated using lysimeters or eddy covariance towers on local scales. Evapotranspiration can be estimated on a larger scale using satellite remote sensing or land surface models but these carry with them inherent biases due to varying algorithms, spatial resolutions, calibration, and input data (Long et al., 2014). Many studies have shown that derived ET products fail to reconcile the terrestrial water budget on multiple temporal scales (Carter et al., 2018). A water balance approach is commonly used on a catchment scale and with observed streamflow obtained from a measured outlet (Han et al., 2015). A water balance sets ET (mm) equal to the precipitation (mm) minus basin runoff (mm), with water storage and net groundwater flux assumed to be zero on an annual scale.

The MOPEX dataset does not contain daily ET. Studies that have made use of MOPEX data obtain ET via the water balance approach using the precipitation and observed runoff (Berghuijs et al., 2014; Coopersmith et al., 2012; Sawicz et al., 2014). As mentioned previously, CAMELS provides three different daily forcing datasets (Daymet, Maurer, NLDAS), which do not contain ET, in addition to three Sacramento Soil Moisture and Accounting Model (SAC-SMA) generated time series from each of the forcing datasets. In this study, daily ET values from the model output time series using Daymet forcing variables (CAMELS SAC-SMA) were compared to the water balance derived ET (CAMELS WB) using CAMELS catchment averaged Daymet precipitation and USGS runoff values to

evaluate any notable differences between methods and facilitate comparison of MOPEX and CAMELS. The CAMELS SAC-SMA model derived ET values are typically greater than the values derived from the CAMELS WB, which become more prominent at an annual scale, as plotted in Fig. 3. When the individual annual differences between CAMELS SAC-SMA model estimated ET and CAMELS WB estimated ET are averaged together, SAC-SMA model estimations are approximately 13 mm larger in arid regions (Fig. 3a), 36 mm larger in continental regions (Fig. 3b), and 50 mm larger in temperate regions (Fig. 3c).

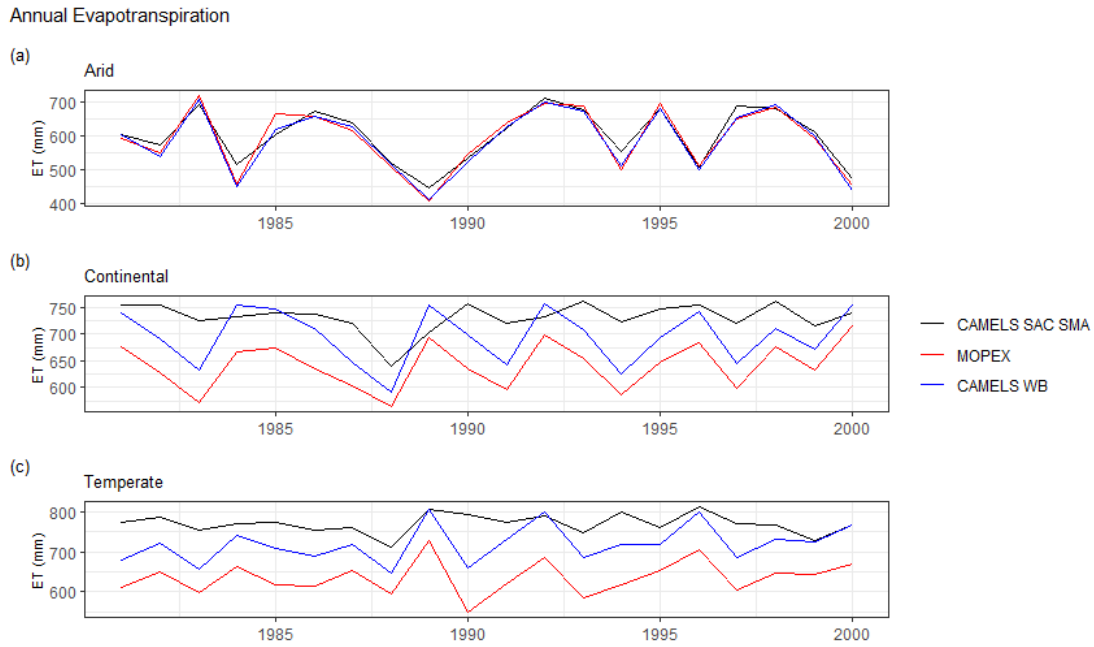

Figure 3. Total annual evapotranspiration for a) arid, b) continental, and c) temperate regions. The annual values are the overall mean of all basin totals in a region. The model output ET (CAMELS SAC-SMA), water balance derived MOPEX ET (MOPEX), and water balance derived CAMELS ET (CAMELS WB) are shown in each plot.

Higher ET values lead to reduced runoff. As shown in Fig. 4, estimated ET values from CAMELS SAC-SMA model were subtracted from the provided CAMELS (Daymet) precipitation data to calculate estimated runoff (SAC_RUN), which was then compared to observed runoff (OBS_RUN). Incorporating ET values from the model output time series as an input variable to a hydrologic

model may result in slightly lower discharge estimates, primarily reflecting the influence of ET values rather than actual runoff conditions.

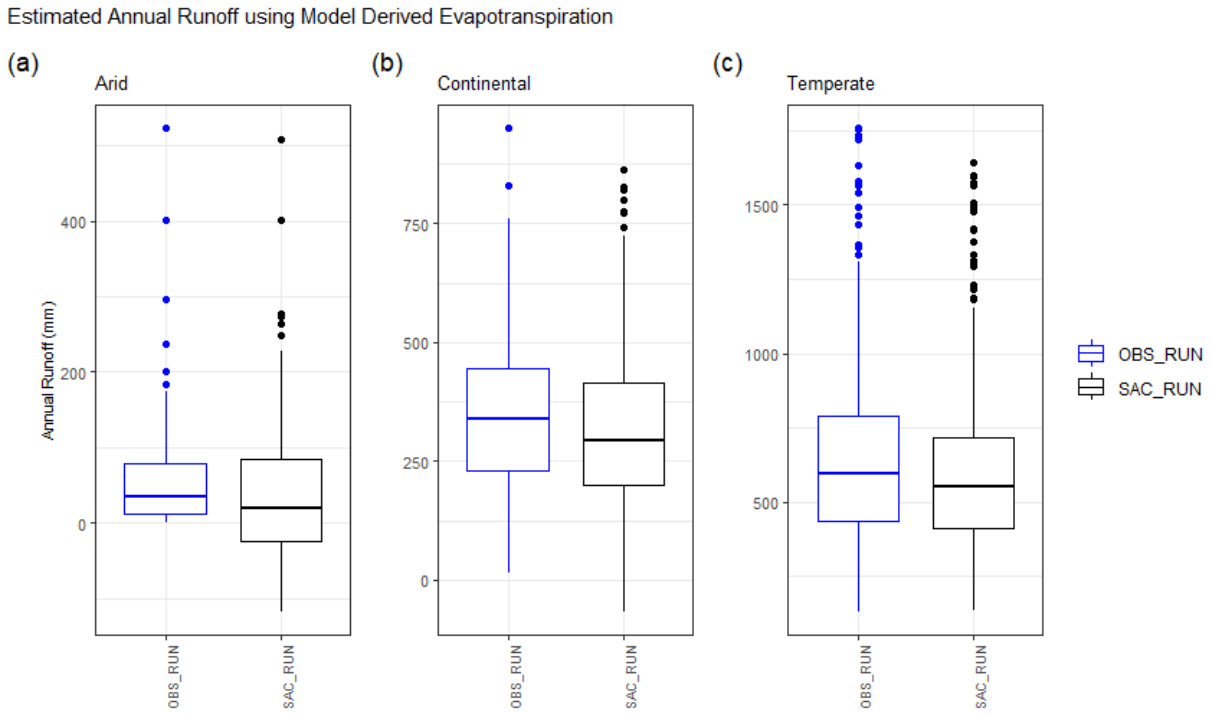

Figure 4. Total annual runoff for a) arid, b) continental, and c) temperate regions. The boxplots represent the annual totals for all basins in a region, with measured observed runoff (OBS_RUN) and water balance calculated runoff (PRCP minus ET) using CAMELS SAC-SMA model output ET (SAC_RUN).

Figure 3 (page 7) was renumbered to Figure 5 and the subsequent figures in the manuscript and captions were updated to reflect the insertion of the new figures.

The text resumes with the original line 140 "A Budyko diagram …" and ends with "4.88% in continental regions from line 153. The text was slightly modified to read as follows

A Budyko diagram, plotting evaporative versus aridity indices, clarifies the predominant hydrologic processes versus climate type (Fig. 5) for the common basins. The CAMELS SAC-SMA evapotranspiration (solid symbols, Fig. 5) exhibits large discrepancies from CAMELS-WB (open symbols, Fig. 5) and MOPEX ET values for most catchments (arrows, Fig. 5). Furthermore, several CAMELS SAC-SMA gauges plotted above the water limit (i.e. to extreme values in the Budyko context) and were 10 to 12 % larger than the water-balance-calculated evapotranspiration indices. The higher SAC-SMA model-derived ET for CAMELS could reflect additional non-precipitation sources of water to the catchment, but that was not evaluated in this study. The largest discrepancies between model-derived ET/P and water balance derived ET/P for CAMELS for each climate region are 46 %, 12 %, and 60 % for arid, continental, and temperate regions respectively.

Average discrepancies for both CAMELS evapotranspiration values are largest in arid regions, 12 %, followed by average discrepancies of 11 % in temperate regions, and 5 % in continental regions.

The next portion of the sentence in line 153 "Studies that have made use of MOPEX data obtain evapotranspiration via the water balance approach" was deleted and relocated to the inserted text presented on page 3 of this response document.

The text then continues with "Differences between …" from line 160.

The remainder of the original sentence on line 155 beginning with "further research …" and ending with "snowmelt is negligible." on line 158 was moved to immediately after the original Figure 3 (now figure 5), concluding Section 3.1

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.

Thank you for your comment. We agree that this could be clarified, and we have made changes to assist the reader with comparisons. After the end of the sentence on line 251 ("both within and between them."), we modified the remaining text for Section 4 to read:

By considering multiple statistics, we can evaluate the representativeness of each dataset and identify any systematic differences that may need further investigation. If both datasets exhibit similar means and variability within a climate region, it suggests that their distributions are comparable. Differences in variance and skewness, on the other hand, highlight potential biases between the datasets. Though there are consistent biases, they are minimal for aggregations beyond a daily time step, making them suitable for combined application in climate studies and hydrologic modelling at monthly, seasonal, or annual aggregations. Although efforts were made to distinguish results for internal analyses within datasets and intercomparisons between datasets, the results are often presented together to provide a clearer understanding of how each dataset behaves independently, while also enabling direct cross-dataset evaluation. Consequently, some overlap does occur.

We have relocated the original tables 4 and 5 to Section 4.2 because they coincide with the other results for between datasets. Original tables 4 and 5 (renumbered to 7 and 8) are now located at line 506, immediately after the caption for original figure 17. All tables have been renumbered accordingly in the captions and manuscript text. The text preceding the relocated tables, beginning at line 506 reads:

Overall statistics for precipitation are shown in Table 7 and were calculated over all shared basins within a climate region. Temperature statistics are shown in Table 8. The mean values and corresponding confidence intervals are based on the averages derived from bootstrapping results, shown in Figs. 7-9 for monthly, seasonal, and annual precipitation values and Figs. 10-12 for monthly, seasonal, and annual temperature values. The tables highlight the commensurate central tendencies, variabilities, and dispersion values within the datasets and provide insight into the comparisons between the datasets.

The manuscript resumes after the inclusion of the relocated tables (original tables 4 and 5 now tables 7 and 8) with original text beginning on line 506 "To assess the magnitude…"

We have included two new tables, now Tables 4 and 5 that show the internal variability of the datasets with the minimum and maximum values for each statistic. These tables show the range of median, mean, variance, standard deviation, and skewness values for the basins within each climate region. We have edited the text in Section 4.1 (removing original lines 254 through 260) so that it now begins on line 254 with the following:

The internal uncertainty and variability of the MOPEX and CAMELS datasets were assessed using median, mean, variance, standard deviation, skewness, coefficient of variation, and confidence intervals for each climate region at daily, monthly, seasonal, and annual scales. Precipitation statistics shown in Table 4 were determined for each individual basin, with minimum and maximum values representing all basins within each climate region (number of observations, Table 3). Due to the large differences between temporal precipitation totals, the ranges for each statistic were normalized by finding the difference between the maximum and minimum values and then dividing the difference by the mean of the maximum and minimum values. The normalized ranges were then used to assess the variability of each statistic within the dataset across the different temporal aggregations and are summarized in Table 4. When calculating the normalized range, a minimum value of zero or close to zero (i.e. median or skew) will conflate the range, making it appear larger than it truly is. For this reason, normalized daily median ranges and normalized skew values are ignored.

Table 4. Minimum (min) and maximum (max) median, mean, variance, standard deviation, and skewness for MOPEX (M) and CAMELS (C) total daily, monthly, seasonal, and annual precipitation (PRCP) totals in mm. Values are based on all basins within a climate region. The normalized ranges (NR) are based on the maximum and minimum values (maximum – minimum / mean (maximum + minimum)).

	PRCP	Median			Mean			Variance			Standard Deviation			Skewness	
	mm	Min	Max	NR	Min	Max	NR	Min	Max	NR	Min	Max	NR	Min	Max
ARID CAMELS	Day	0.00	0.00	-	1.30	2.50	0.63	13.93	60.91	1.26	3.73	7.80	0.71	4.01	5.84
	Month	11.50	63.25	1.38	39.55	76.01	0.63	1421.16	4110.82	0.97	37.70	64.12	0.52	1.23	2.23
	Season	91.10	209.51	0.79	118.66	228.03	0.63	6608.73	16794.05	0.87	81.29	129.59	0.46	0.45	1.37
	Year	423.07	920.40	0.74	474.63	912.11	0.63	13628.52	64193.54	1.30	116.74	253.36	0.74	-0.46	0.87
ARID MOPEX	Day	0.00	0.15	-	1.27	2.31	0.58	13.36	54.01	1.21	3.66	7.35	0.67	5.40	7.81
	Month	9.13	55.67	1.44	38.55	70.33	0.58	1329.37	3569.40	0.91	36.46	59.74	0.48	1.00	2.34
	Season	69.89	202.28	0.97	115.65	211.00	0.58	6289.11	18484.56	0.98	79.30	135.96	0.53	0.37	1.64
	Year	381.36	878.77	0.79	462.60	844.02	0.58	11933.15	54745.76	1.28	109.24	233.98	0.73	-0.41	0.83
CONT CAMELS	Day	0.00	0.57	-	2.25	3.19	0.35	11.44	64.58	1.40	3.38	8.04	0.82	2.26	4.70
	Month	58.74	88.78	0.41	68.51	97.06	0.34	1529.66	4577.49	1.00	39.11	67.66	0.53	0.71	1.99
	Season	203.99	281.15	0.32	205.52	291.19	0.34	5124.76	20848.61	1.21	71.59	144.39	0.67	-0.01	0.95
	Year	796.50	1197.26	0.40	822.10	1164.76	0.34	11734.23	63049.45	1.37	108.32	251.10	0.79	-0.5	1.14
CONT MOPEX	Day	0.03	0.87	-	2.08	3.16	0.41	9.68	62.68	1.46	3.11	7.92	0.87	2.9	5.33
	Month	60.05	86.34	0.36	63.35	96.11	0.41	966.07	4001.93	1.22	31.08	63.26	0.68	0.83	1.62
	Season	209.10	277.72	0.28	190.04	288.34	0.41	3920.88	18210.45	1.29	62.62	134.95	0.73	0.03	1.00
	Year	765.73	1143.82	0.40	760.18	1153.35	0.41	11497.26	53327.77	1.29	107.23	230.92	0.73	-0.41	1.09
TEMP CAMELS	Day	0.00	0.70	-	2.90	5.71	0.65	26.61	135.48	1.34	5.16	11.64	0.77	2.09	4.80
	Month	79.12	165.40	0.71	88.14	173.68	0.65	1872.09	6241.27	1.08	43.27	79.00	0.58	0.39	1.24
	Season	251.52	508.24	0.68	264.42	521.04	0.65	6496.89	28687.61	1.26	80.60	169.37	0.71	0.18	1.12
	Year	1014.72	2181.86	0.73	1057.68	2084.18	0.65	23220.68	128322.97	1.39	152.38	358.22	0.81	-0.66	0.91
TEMP MOPEX	Day	0.04	1.09	-	2.59	5.28	0.68	27.73	144.50	1.36	5.27	12.02	0.78	2.88	5.34
	Month	71.48	151.99	0.72	78.98	160.71	0.68	1458.72	5823.13	1.20	38.19	76.31	0.67	0.48	1.29
	Season	224.14	445.36	0.66	236.93	482.12	0.68	4861.81	26089.63	1.37	69.73	161.52	0.79	0.05	1.12
	Year	923.61	2004.61	0.74	947.70	1928.48	0.68	18274.22	136058.9	1.53	135.18	368.86	0.93	-0.63	0.85

The range in median values decreases in all regions when moving from a monthly aggregation to seasonal (CAMELS arid 1.38 to 0.79, CAMELS continental 0.41 to 0.32, CAMELS temperate 0.71 to 0.68, MOPEX arid 1.44 to 0.97, MOPEX continental 0.36 to 0.28, MOPEX temperate 0.72 to 0.66). Median ranges continue to contract in arid regions at an annual scale (CAMELS 0.74, MOPEX 0.79). Continental and temperate regions show a slight expansion in median ranges between seasonal and annual aggregations. The range in mean values is uniform in each region over all temporal scales (CAMELS arid 0.63, CAMELS temperate 0.65, MOPEX arid 0.58, MOPEX continental 0.41, MOPEX temperate 0.68), with a minor change of 0.35 to 0.34 in CAMELS continental daily to monthly. All basins within each region demonstrate minimal variability in the mean and proportional aggregation. The range in variance is slightly wider for daily and annual aggregations in all regions for both CAMELS and MOPEX. The range in annual variance increases when moving to an annual aggregation except for MOPEX continental, which remains at a normalized value of 1.29 between seasonal and annual. This suggests interannual variability may be more pronounced than intra-seasonal fluctuations, attributed to the accumulation of extreme precipitation events or shifting between wet and dry. The range of standard deviation values mimics the variability in variance values with the smallest ranges for monthly and seasonal aggregations and minor increases at daily and annual aggregations for all regions except MOPEX continental (seasonal and annual 0.73). Differences in precipitation patterns can become more apparent over longer periods of time. Precipitation variability has been shown to increase over longer time scales under a warming climate (Pendergrass et al., 2017; Zhang et al., 2021), The distribution in all regions tends to become more Gaussian as the aggregation increases from daily to annual, which is to be expected.

The minimum and maximum median, mean, variance, standard deviation, and skewness values for mean temperature in degrees Celcius along with the range values are shown in Table 5. Overall, temperature variability in each climate region decreases with temporal aggregation with daily values showing the highest variability and annual values the lowest for both CAMELS and MOPEX datasets. There is minimal variability in the central tendency in all climate regions, with a slightly narrower spread in the mean values compared to the median values. In continental regions, the minimum seasonal median value for all CAMELS basins is 0.81 ˚C and 0.96 ˚C for MOPEX basins, which is due to a few colder than normal winters in a watershed located in Montana. Variance in mean temperature is smallest in annual aggregations for all regions because it is based on annual averages, which smooths the extreme values. In contrast, the variability in skewness is greatest at annual aggregations in both CAMELS and MOPEX. Aggregation at an annual scale reduces variance among mean temperature values but at the same time, fewer data points increase the sensitivity to extremes, which can shift the distribution.

Table 5. Minimum (min) and maximum (max) median, mean, variance, standard deviation, and skewness for MOPEX (M) and CAMELS (C) mean daily, monthly, seasonal, and mean temperature (TAIR) in degree Celsius. Values are based on all basins within a climate region. Range is maximum minus minimum.

	TAIR	Median			Mean			Variance			Standard Deviation			Skewness		
	˚C	Min	Max	Range	Min	Max	Range	Min	Max	Range	Min	Max	Range	Min	Max	Range
ARID CAMELS	Day	8.91	22.77	13.86	8.70	21.54	12.84	48.08	139.15	91.07	6.93	11.8	4.87	-0.68	0.12	0.80
	Month	9.08	21.98	12.90	8.64	21.51	12.87	36.97	113.08	76.11	6.03	10.63	4.60	-0.22	0.14	0.36
	Season	8.91	21.97	13.06	8.63	21.5	12.87	30.14	92.47	62.33	5.49	9.62	4.13	-0.24	0.13	0.37
	Year	8.74	21.57	12.83	8.70	21.54	12.84	0.30	1.14	0.84	0.54	1.07	0.53	-0.30	0.51	0.81
ARID MOPEX	Day	9.08	22.81	13.73	9.08	21.57	12.49	43.84	123.74	79.90	6.62	11.12	4.50	-0.68	0.15	0.83
	Month	9.12	22.00	12.88	9.02	21.54	12.52	36.49	103.1	66.61	6.04	10.15	4.11	-0.23	0.19	0.42
	Season	9.00	21.90	12.90	9.02	21.54	12.52	29.49	84.07	54.58	5.43	9.17	3.74	-0.23	0.16	0.39
	Year	9.04	21.59	12.55	9.08	21.57	12.49	0.27	0.86	0.59	0.52	0.93	0.41	-0.33	0.45	0.78
CONT CAMELS	Day	1.26	14.38	13.12	0.38	13.48	13.10	75.41	141.07	65.66	8.68	11.88	3.20	-0.40	-0.19	0.21
	Month	1.04	13.92	12.88	0.33	13.43	13.10	59.41	119.00	59.59	7.71	10.91	3.20	-0.22	0.08	0.30
	Season	0.81	13.54	12.73	0.33	13.43	13.10	49.13	99.53	50.40	7.01	9.98	2.97	-0.23	0.10	0.33
	Year	1.47	13.38	11.91	0.38	13.48	13.10	0.32	0.96	0.64	0.57	0.98	0.41	-0.18	0.22	0.40
CONT MOPEX	Day	1.16	14.09	12.93	1.45	13.32	11.87	73.49	125.15	51.66	8.57	11.19	2.62	-0.33	-0.04	0.29
	Month	1.34	13.75	12.41	1.40	13.27	11.87	59.02	107.77	48.75	7.68	10.38	2.70	-0.14	0.12	0.26
	Season	0.96	13.29	12.33	1.40	13.27	11.87	48.77	89.26	40.49	6.98	9.45	2.47	-0.14	0.16	0.30
	Year	1.52	13.27	11.75	1.45	13.32	11.87	0.31	0.72	0.41	0.56	0.85	0.29	-0.19	0.34	0.53
TEMP CAMELS	Day	2.72	20.03	17.31	2.73	19.03	16.30	52.24	102.48	50.24	7.36	10.12	2.76	-0.53	-0.03	0.50
	Month	2.36	19.04	16.68	2.68	19.00	16.32	41.80	82.09	40.29	6.46	9.06	2.60	-0.14	0.11	0.25
	Season	2.57	19.23	16.66	2.69	18.99	16.30	34.08	67.80	33.72	5.84	8.23	2.39	-0.12	0.12	0.24
	Year	2.87	18.98	16.11	2.73	19.03	16.30	0.28	0.88	0.60	0.53	0.94	0.41	-0.28	0.50	0.78
TEMP MOPEX	Day	2.75	20.06	17.31	3.11	19.06	15.95	53.08	92.52	39.44	7.29	9.62	2.33	-0.51	0.04	0.55
	Month	2.54	19.16	16.62	3.07	19.03	15.96	41.59	77.30	35.71	6.45	8.79	2.34	-0.15	0.12	0.27
	Season	2.83	19.26	16.43	3.06	19.03	15.97	34.05	64.06	30.01	5.84	8.00	2.16	-0.13	0.15	0.28
	Year	3.23	19.06	15.83	3.11	19.06	15.95	0.25	0.59	0.34	0.50	0.77	0.27	-0.29	0.55	0.84

The coefficient of variation (CV) was calculated for each catchment on all temporal scales for precipitation (Fig. 6). Daily precipitation shows considerably high variation, with CAMELS mean CV values of 3.28, 2.39, and 2.12 and MOPEX mean CV values of 3.23, 2.42, and 2.22 in arid, continental, and temperate regions respectively (Fig. 6a). Considerably high variation is still observed on monthly scales (Fig. 6b) but decreases to moderate variability for seasonal temporal aggregations for all regions, and low variability, less than one, on an annual scale. The normalized ranges for precipitation variance in Table 4 indicate that annual totals are the most variable while the CV demonstrates decreasing variability from a daily to annual scale. While both are measures of variability, they differ in how they express dispersion and their sensitivity to scale. Variance is unit dependent and is sensitive to magnitude while the CV is normalized relative to the mean. This suggests that at short time scales, precipitation is more event-driven whereas at longer scales, climate patterns dominate. Temperature demonstrates a consistent decrease in variability from daily to annual temporal aggregation for all regions and is not shown.

Manuscript resumes with the precipitation coefficient of variation figure, now Figure 6.

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?.

We agree and have removed the sentence.

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.

Thank you for your comment. We have modified the portion of the text to now read:

For average total annual precipitation, arid regions exhibit the highest variability within each individual dataset, however, their overall temporal mean values remain similar between the two datasets

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

Thank you for your suggestion. Our objective is to present the results as concisely as possible and reserve our speculations for the discussion section. See discussion beginning on line 668. We hope to maintain the structure of our paper. We also added significantly to our machine learning section and 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.

Thank you for this comment. We have reworked Section 3.3 to explain the reasoning behind machine learning validation. We also included additional analyses to provide further support for this form of validation. The section now reads as follows:

Machine learning (ML) techniques, such as linear regression, random forest, gradient boosting, and support vector regression, offer a valuable alternative to physically based models by capturing relationships between input and output variables. While they do not rely on detailed hydrological processes, these models can still provide robust predictions and allow for comparative analysis of different datasets (Herrera et al., 2022). Using ML models as a proxy is increasingly common in hydrological research, as these models can efficiently handle high-dimensional data and learn intricate patterns without explicitly modelling physical processes (Kratzert et al., 2019). ML models have been shown to perform well in a range of hydrological applications, especially in data-rich contexts. For this study, we employed ML models to evaluate the potential influences of MOPEX and CAMELS precipitation and temperature biases on predicted runoff.

Hydrologic models rely on parameterization and assumptions about physical processes while ML models learn directly from data, reducing dependence on prior assumptions and allowing for a purely data-driven evaluation (Nearing et al., 2021). ML models can highlight inconsistencies or biases in input datasets by comparing their predictive performance across datasets. If one dataset consistently leads to better predictions, it may indicate better representativeness or higher quality. Traditional hydrologic models typically require extensive calibration and long run times, especially for larger scale applications, but ML models, once trained, can make predictions rapidly and do not require manual calibration (Kratzert et al., 2019). ML models can also be trained separately on different temporal scales, allowing for direct comparisons without modifying model structures. By evaluating performance metrics across datasets, ML provides an objective assessment of whether precipitation and temperature inputs are sufficient to capture runoff variability (Yokoo et al., 2022).

Four different ML models were implemented in R to estimate runoff from precipitation and mean air temperature using the *e1071* (Meyer et al., 2024), *gbm* (Ridgeway et al., 2024), *randomForest* (Breiman et al., 2024), and *caret* (Kuhn et al., 2024) packages. Linear regression models the

relationship between a dependent variable and one or more independent variables by fitting a linear equation (Xu and& Liang, 2021). Random forest is an ensemble learning method that constructs multiple decision trees and averages their predictions to improve accuracy and reduce overfitting (Breiman, 2001). Gradient boosting builds models sequentially, optimizing for errors in previous iterations by combining weak learners to create a stronger predictive model (Xu and& Liang, 2021). Support vector regression (SVR) maps input data into a higher-dimensional space and finds the ideal hyperplane, separating the data points into different classes, and minimizes prediction error while maintaining generalization (Shmilovici, 2023). These models provide a diverse approach to estimating runoff, ranging from simple linear relationships to more complex, non-linear learning techniques.

MOPEX and CAMELS precipitation and temperature values were used as input to predict runoff at daily, monthly, seasonal, and annual time scales. Precipitation and temperature data were transformed into common scales using min-max normalization. Datasets were then split into training and test sets, with 80 % of the data allotted to training and 20 % to testing. Rather than partitioning the data into multiple subsets, each ML model was run 10 times, resampling and randomly splitting into testing and training sets (Domingos, 2012). Predicted runoff values were then compared to actual observed runoff to assess model accuracy using root mean square error (RMSE), MAE, R2, and bias as performance metrics. Model results were then compared across MOPEX and CAMELS datasets to determine their consistency, assess whether they provide compatible inputs for runoff estimation, and the influence of potential systematic biases in the input data.

SVR was also able to compare MOPEX and CAMELS datasets as a simple binary classification problem using the *e1071* (Meyer et al., 2023) R package. The two datasets were merged into a composite dataset for each climate region and temporal aggregation, and each was identified by either a zero (CAMELS) or one (MOPEX), representing the target variable. The composite dataset was then split into training and test sets, with 75 % of the data allotted to training and 25 % to testing. Data were randomly selected to avoid any potential bias due to formatting, etc. SVR models were trained on the composite datasets to classify the binary label (MOPEX or CAMELS) using precipitation, temperature, and evapotranspiration values as predictor variables. Classification was performed separately for all three climate regions at daily, monthly, seasonal, and water year aggregations. If the datasets are similar, then the model should have difficulty differentiating between them, yielding a classification probability near 50 %, akin to a random guess

Section 4.4, Validation, lines 631 to 633 were removed. The section is now as follows:

Hydrologic models are used to simulate real world processes and range from simple conceptual models to complex physically based models. Choosing a suitable model is highly dependent on the purpose and scale. The input data required depends on the spatial and temporal distributions evaluated in a model, but precipitation and temperature are fundamental. Inherent biases in input data can skew modelling results. Machine learning (ML) was used instead of hydrologic models (i.e. SWAT, VIC, SAC-SMA) because ML models provide a data-driven, model-agnostic approach that focuses on the relationships between inputs and outputs without relying on predefined process-based assumptions (Herrera et al., 2022). Four machine learning models were used to predict runoff at daily, monthly, seasonal, and annual scales for MOPEX and CAMELS. The objective is not

to determine model suitability, rather evaluate the performance of each dataset. The RMSE, MAE, R2, and bias of predicted versus observed runoff serve as dataset comparisons.

On a daily scale, CAMELS has a slightly lower RMSE and MAE than MOPEX for all regions and a better R2, although the values are quite low, less than 0.3, shown in Table 12. A good fit is not expected with daily data which will have multiple zero values for precipitation.

Table 12. Machine learning model metrics for predicted versus observed total daily runoff using total daily precipitation and mean daily temperature data as inputs for CAMELS (C) and MOPEX (M).

Day	ML Model	RMSE C	RMSE M	MAE C	MAE M	R2 C	R2 M	Bias C	Bias M
ARID	Linear Regression	0.73	0.82	0.21	0.21	0.15	0.08	-0.01	0.00
	Random Forest	0.70	0.84	0.20	0.22	0.21	0.04	-0.01	0.01
	Gradient Boosting	0.69	0.82	0.20	0.21	0.25	0.09	0.00	0.00
	SVR	0.77	0.85	0.17	0.17	0.12	0.06	-0.06	-0.06
CONT	Linear Regression	1.99	2.09	0.89	0.93	0.21	0.06	0.00	-0.01
	Random Forest	1.95	2.14	0.85	0.95	0.25	0.05	0.01	0.01
	Gradient Boosting	1.89	2.06	0.83	0.90	0.29	0.09	-0.01	-0.01
	SVR	1.99	2.15	0.73	0.77	0.29	0.08	-0.39	-0.46
TEMP	Linear Regression	2.62	2.83	1.41	1.48	0.17	0.06	0.00	-0.01
	Random Forest	2.60	2.87	1.38	1.48	0.19	0.07	0.01	0.01
	Gradient Boosting	2.51	2.77	1.34	1.41	0.24	0.10	-0.01	-0.01
	SVR	2.63	2.88	1.21	1.26	0.23	0.11	-0.60	-0.69

At the monthly aggregation, Table 13, CAMELS narrowly outperforms MOPEX with lower RMSE and MAE values. The R2 values are extremely similar between datasets in all regions, and both exhibit the same positive biases with all ML models except for SVR, which underpredicts runoff and results in negative biases for both datasets. The results indicate that the predictive performance of the models is very similar across both datasets, suggesting a high degree of consistency between them.

Table 13. Machine learning model metrics for predicted versus observed total monthly runoff using total monthly precipitation and mean monthly temperature data as inputs for CAMELS (C) and MOPEX (M).

Month	ML Model	RMSE C	RMSE M	MAE C	MAE M	R2 C	R2 M	Bias C	Bias M
ARID	Linear Regression	7.73	10.64	5.12	5.94	0.37	0.37	0.62	0.64
	Random Forest	8.08	10.82	3.82	5.32	0.31	0.37	0.12	0.30
	Gradient Boosting	8.15	11.13	4.00	5.54	0.33	0.35	0.37	0.44
	SVR	7.00	10.65	3.48	4.34	0.36	0.39	-0.91	-1.42
CONT	Linear Regression	23.42	23.45	17.16	16.95	0.40	0.41	0.42	0.31
	Random Forest	22.61	22.81	16.05	16.03	0.45	0.44	0.48	0.29
	Gradient Boosting	21.49	21.87	15.40	15.56	0.50	0.48	0.45	0.51
	SVR	21.69	22.02	14.80	14.93	0.50	0.49	-2.74	-2.99
TEMP	Linear Regression	41.55	42.09	28.96	29.27	0.32	0.31	0.74	0.78
	Random Forest	41.01	40.97	28.13	28.18	0.36	0.36	1.39	1.27
	Gradient Boosting	38.73	38.91	26.52	26.84	0.41	0.41	1.21	0.75
	SVR	38.71	39.21	25.29	25.85	0.42	0.41	-5.00	-5.20

Seasonally, Table 14, the main discrepancies between the datasets are in continental regions, where CAMELS runoff predictions are lower than those from MOPEX by approximately 4 to 5 mm. This difference, while evident, is relatively small and may not have significant implications for broader regional or long-term studies. For instance, seasonal runoff values in continental regions range from 0.3 mm in one basin (JJA 1988) to 423.88 mm (MAM 1996) in another basin. This effect of these biases would be more pronounced for basins with very little runoff in a specific season, but this issue is not unique to these datasets. Any dataset used on such a fine, basin-specific scale may exhibit similar biases.

Table 14. Machine learning model metrics for predicted versus observed total seasonal runoff using total seasonal precipitation and mean seasonal temperature data as inputs for CAMELS (C) and MOPEX (M).

Season	ML Model	RMSE C	RMSE M	MAE C	MAE M	R2 C	R2 M	Bias C	Bias M
ARID	Linear Regression	21.60	22.66	14.52	14.65	0.37	0.31	0.48	0.41
	Random Forest	16.96	20.25	10.36	11.27	0.61	0.55	1.29	1.02
	Gradient Boosting	18.25	21.17	11.26	12.15	0.59	0.54	2.27	1.32
	SVR	18.63	21.33	9.27	9.49	0.53	0.39	-2.49	-4.35
CONT	Linear Regression	59.11	51.78	43.00	38.57	0.44	0.43	-4.41	-0.12
	Random Forest	54.75	54.91	39.47	40.55	0.52	0.58	-4.45	-0.99
	Gradient Boosting	53.19	50.00	39.01	36.76	0.55	0.47	-5.00	0.07
	SVR	53.93	50.94	37.02	35.79	0.55	0.47	-9.69	-4.22
TEMP	Linear Regression	97.33	96.28	70.64	70.71	0.31	0.33	0.66	0.46
	Random Forest	92.10	92.02	67.26	65.56	0.40	0.40	2.23	2.28
	Gradient Boosting	87.83	88.08	66.09	63.93	0.44	0.44	2.54	1.92
	SVR	91.48	91.78	64.44	64.54	0.41	0.41	-7.17	-9.29

The differences in precipitation and temperature between MOPEX and CAMELS become more relevant depending on the scale and objective of the study. For daily or single-month analyses, as well as for very specific seasons, the datasets may not be directly comparable. However, as with any modelling approach, results come with inherent uncertainty, which should be acknowledged when presenting findings. Model results should be accompanied by an uncertainty estimate, reflecting potential biases or discrepancies. Bias correction is an essential part of any modelling process, typically done during the calibration phase (Lehner et al., 2023b). In this context, the warm bias in MOPEX and the wet bias in CAMELS are important only when focusing on very fine, basin-specific or daily temporal scales. As evidenced by the subsequent analyses, on larger temporal or spatial scales, these biases are less likely to significantly affect the conclusions, making these two datasets comparable for general hydrological or climate studies. At an annual scale, Table 15, MOPEX and CAMELS have improved R2 and the same predicted runoff biases despite the overall warm MOPEX temperature biases and wet CAMELS precipitation biases present in the data. The similarity in predicted runoff demonstrates the compatibility between the two datasets and that no corrections to the raw data are required at an annual scale.

Table 15. Machine learning model metrics for predicted versus observed total annual runoff using total annual precipitation and mean annual temperature data as inputs for CAMELS (C) and MOPEX (M).

Year	ML Model	RMSE C	RMSE M	MAE C	MAE M	R2 C	R2 M	Bias C	Bias M
ARID	Linear Regression	33.57	32.52	28.63	26.97	0.79	0.78	17.72	19.03
	Random Forest	31.82	34.20	23.87	25.44	0.58	0.60	10.86	13.29
	Gradient Boosting	35.33	30.54	26.30	22.16	0.68	0.83	17.85	21.22
	SVR	20.18	21.24	15.95	16.53	0.81	0.81	4.27	8.06
CONT	Linear Regression	85.78	91.05	77.12	77.10	0.69	0.67	-6.11	-9.69
	Random Forest	91.53	94.40	82.25	77.77	0.64	0.65	-10.76	-10.76
	Gradient Boosting	89.53	95.91	83.76	80.24	0.64	0.64	-15.48	-15.15
	SVR	80.13	85.16	74.19	70.84	0.73	0.73	-11.10	-15.48
TEMP	Linear Regression	106.89	113.85	82.26	91.08	0.88	0.87	-19.61	-19.52
	Random Forest	121.34	126.85	92.93	103.58	0.85	0.83	-17.52	-23.33
	Gradient Boosting	109.74	121.89	85.29	95.15	0.87	0.84	-20.21	-28.71
	SVR	101.84	118.68	79.35	94.11	0.89	0.86	-19.59	-33.60

Total annual observed runoff is plotted against the predicted runoff for all four ML models in Fig. 27 with a 1:1 reference line. In all regions, MOPEX and CAMELS exhibit similar visual patterns and alignment of the points. In arid regions (Fig. 27a, b), both datasets show distinct clusters of low and high runoff values, reflecting greater variability and defined wet and dry periods. In contrast, continental (Fig. 27c, d) and temperate regions (Fig. 27e, f) display a more even distribution of runoff throughout the year, with both datasets capturing this behavior.

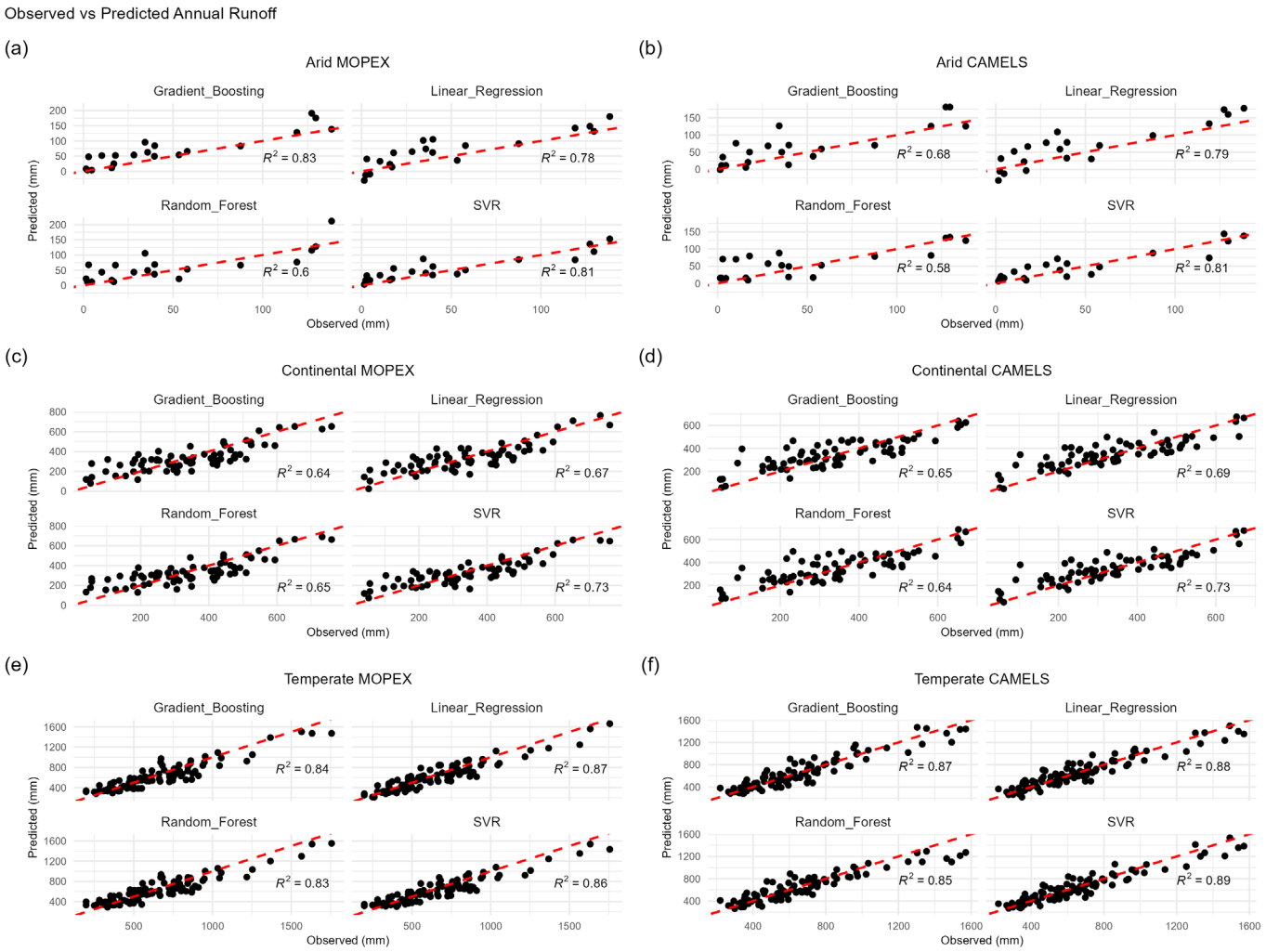

Figure 27. Observed versus predicted total annual runoff (mm) for each ML model and each climate region. The dashed red line represents the 1:1 reference line.

In addition to predicting runoff, machine learning was used to differentiate between MOPEX and CAMELS. Data were separated by climate region and then daily, monthly, seasonal, and annual precipitation, temperature, and water balance derived evapotranspiration were used for classifications. The support vector machine performed binary classification, assigning the standardized values to either MOPEX or CAMELS.

Section 4.4 resumes with original line at 634 beginning with "The classification accuracy values shown …"

The Conclusions, Section 6 numbered points were revised to read:

1) The relevance of differences between MOPEX and CAMELS depends on the study's scale and purpose.

2) Daily pairwise comparisons are not recommended due to the variability in extreme precipitation event measurements. However, both datasets capture similar patterns and basin behavior, e.g. when evaluating the number of rainy days or dry days per year. Comparison improves significantly with monthly, seasonal, and annual aggregations. Despite temperature and precipitation biases, MOPEX and CAMELS show similar predicted runoff at the annual scale, requiring no raw data corrections. Monthly, seasonal, and annual values are comparable, as their differences are within expected uncertainty ranges.
3) Compatibility is constrained by basin water balance and requires basin averaged values, i.e. ET values from model output CAMELS time series must be used with caution, and often cannot be reconciled with MOPEX or other water-balance based estimates
4) All modelling results should include uncertainty estimates. Bias correction is typically performed during calibration, addressing dataset specific biases.

Additional References added to manuscript

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Carter, E., Hain, C., Anderson, M., & Steinschneider, S. (2018). A water balance-based, spatiotemporal evaluation of terrestrial evapotranspiration products across the contiguous United States. *Journal of Hydrometeorology*, *19*(5), 891–905. https://doi.org/10.1175/JHM-D-17-0186.1

Domingos, P. (2012). A few useful things to know about machine learning. In *Communications of the ACM* (Vol. 55, Issue 10, pp. 78–87). https://doi.org/10.1145/2347736.2347755

Herrera, P. A., Marazuela, M. A., & Hofmann, T. (2022). Parameter estimation and uncertainty analysis in hydrological modeling. In *Wiley Interdisciplinary Reviews: Water* (Vol. 9, Issue 1). John Wiley and Sons Inc. https://doi.org/10.1002/wat2.1569

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Lehner, F., Nadeem, I., & Formayer, H. (2023). Evaluating skills and issues of quantile-based bias adjustment for climate change scenarios. *Advances in Statistical Climatology, Meteorology and Oceanography*, *9*(1), 29–44. https://doi.org/10.5194/ascmo-9-29-2023

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, *50*(2), 1131–1151. https://doi.org/10.1002/2013WR014581

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? In *Water Resources Research* (Vol. 57, Issue 3). Blackwell Publishing Ltd. https://doi.org/10.1029/2020WR028091

Pendergrass, A. G., Knutti, R., Lehner, F., Deser, C., & Sanderson, B. M. (2017). Precipitation variability increases in a warmer climate. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-17966-y

Pierce, D. W., Cayan, D. R., Maurer, E. P., Abatzoglou, J. T., & Hegewisch, K. C. (2015). *Improved Bias Correction Techniques for Hydrological Simulations of Climate Change**. https://doi.org/10.1175/JHM-D-14-0236.s1

Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. In *Wiley Interdisciplinary Reviews: Water* (Vol. 8, Issue 5). John Wiley and Sons Inc. https://doi.org/10.1002/wat2.1533

Yokoo, K., Ishida, K., Ercan, A., Tu, T., Nagasato, T., Kiyama, M., & Amagasaki, M. (2022). Capabilities of deep learning models on learning physical relationships: Case of rainfall-runoff modeling with LSTM. *Science of the Total Environment*, *802*. https://doi.org/10.1016/j.scitotenv.2021.149876

Zhang, W., Furtado, K., Wu, P., Zhou, T., Chadwick, R., Marzin, C., Rostron, J., & Sexton, D. (2021). Increasing precipitation variability on daily-to-multiyear time scales in a warmer world. In *Sci. Adv* (Vol. 7). https://www.science.org