Reply to RC1: 'Comment on essd-2026-350', Anonymous Referee #1

The manuscript presents Caravan-CMIP6, a bias-corrected CMIP6 climate projection dataset prepared for ten large-sample hydrological (CAMELS-style) datasets. The dataset is expected to be valuable for the hydrological community, particularly for climate change impact studies requiring consistent meteorological forcing across multiple catchments. The dataset itself appears useful, and both the raw and bias-corrected products are made available. However, I found that many methodological choices are insufficiently explained, several implementation decisions lack scientific justification, and the manuscript relies too heavily on previous publications for essential methodological details. As an ESSD paper, the manuscript should serve as the primary documentation of the dataset and therefore should be more self-contained. I recommend **major revision** before the manuscript can be considered for publication.

Thank you for this thorough review and clear instructions for revision. The review is reproduced below in black text and my responses are provided below in blue. Text to be added to the manuscript is italicized.

General comments

1. While active voice and first-person perspective are acceptable for a single-author paper in ESSD, the current manuscript relies very heavily on personal narrative phrasing throughout (e.g., "I provide...", "I bias-correct...", "I produce...", "I required..."). At times, this shifts the focus away from the dataset and methodology and onto the author's workflow. I recommend varying the sentence structure by focusing more on the dataset, processing steps, and analytical framework to improve the scientific flow and readability.

   I will revise the language to focus more on the dataset/methodology.

2. Throughout the manuscript, many methodological descriptions are replaced by references to previous studies (e.g., model selection, weighting strategy, bias correction, PET calculation). While citing previous work is appropriate, the manuscript should provide concise but sufficient descriptions of the methodology so that readers can understand how Caravan-CMIP6 was constructed without repeatedly consulting multiple publications.

   I will expand the description of the model selection at line 77:

   The models are chosen following (Mahony et al., 2022) who selected an ensemble based on many criteria, including that the models must: be representative of the distribution of equilibrium climate sensitivity in CMIP6, contain tasmax and tasmin, *contain at least three historical runs, and not contain large biases over North America. They also selected only one model per institution and excluded models that shared components. Here, several modifications were made to Mahony's ensemble, according to the following criteria:*

For the weighting strategy, see my response below under "Model selection" comment 3.

For the PET calculation, I assume the reviewer refers to Morton's PET calculation, since the Penman-Monteith and Hargreaves equations are given. I will expand the description of Morton's ET/PET as follows:

Morton's point potential evapotranspiration, wet-environment potential evapotranspiration, and areal actual evapotranspiration are calculated according to the procedure outlined by Morton (1983) as modified by Zajaczkowski and Jeffrey (2020) for the Scientific Information for Land Owners dataset used by CAMELS-AUS-v2 (Fowler et al., 2025). Morton's Complementary Relationship for Areal Evapotranspiration (CRAE) model assumes that 'areal actual' and 'point potential' evapotranspiration have a complementary (inverse) relationship, and sum to two times the 'wet-environment potential' evapotranspiration. The CRAE algorithm includes an iterative solver for an equilibrium temperature at which "the energy-balance equation and the vapour transfer equation for a moist surface give the same result (Morton, 1983) with a convergence criterion of 0.1 °C. Note that each component is bias-corrected separately, so the complementary relationship is not strictly preserved in the bias-corrected data.

3. A summary table describing the final Caravan-CMIP6 dataset would greatly improve readability. Such a table could summarize the included CAMELS datasets, number of catchments, reference observational datasets used for bias correction, variables provided, derived variables, temporal resolution, temporal coverage, and spatial representation.

   Table R1 will be added to the appendix of the manuscript. It includes CAMELS datasets, number of catchments, reference observational datasets used for bias correction, and variables provided. The table also indicates which variables are derived and which are provided as standard climate model output. The temporal resolution & coverage and spatial representation are the same for all of the data (1850-2100, daily data, catchment averages).

Table R1: Summary of included data from the ten large-sample hydrometeorological datasets. The 'derived' column indicates whether the variable is present in the CMIP6 output data or whether it is derived from other variables.

Dataset		**Caravan**	**CAMELS**	**CAMELS-AUS-v2**	**CAMELS-BR**	**CAMELS-CH**	**CAMELS-CL**	**CAMELS-COL**	**CAMELS-DE**	**CAMELS-GB-v2**	**CAMELS-IND**
Number of catchments		24870	671	561	897	331	516	347	1582	671	472
Variable	**Derived?**										
pr	No	ERA5-Land	Daymet, Maurer, NLDAS	AGCD, SILO	BR-DWGD v3.2.3, CHIRPS V2.0, CPC, ERA5-Land, MSWEP v2.8	MeteoSwiss2023, PREVAH	CR2METv1.3, CHIRPS V2.0, MSWEP v1.1	CHIRPS V2.0	DWD-HYRAS	CEH-GEAR, HadUK-GRID	IMD
tas	No	ERA5-Land	Maurer, NLDAS		ERA5-Land	MeteoSwiss2023, PREVAH	CR2METv1.3	MSWX	DWD-HYRAS	CHESS-Met, HadUK-GRID	Srivastava2009
tasmax	No	ERA5-Land	Daymet	AGCD, SILO	CPC, ERA5-Land, BR-DWGD v3.2.3	MeteoSwiss2023	CR2METv1.3	MSWX	DWD-HYRAS		Srivastava2009

tasmin	No	ERA5-Land	Daymet	AGCD, SILO	CPC, ERA5-Land, BR-DWGD v3.2.3	MeteoSwiss2023	CR2METv1.3	MSWX	DWD-HYRAS		Srivastava2009
pet	Yes	derived (Penman-Monteith)		SILO (Penman-Monteith), SILO (Morton)	ERA5-Land, GLEAM v4.2a, BR-DWGD v3.2.3	PREVAH	derived (Hargreaves)	MSWX (Hargreaves)	derived (Hargreaves)	Hydro-PE, CHESS-PE	Singer2021, GLEAM v3
ps	Yes	ERA5-Land									
psl	No			SILO							
rlds	No										IMDAA
rls	Yes	ERA5-Land									
rsds	No		Daymet, Maurer, NLDAS	SILO		PREVAH			DWD-HYRAS		IMDAA
rss	Yes	ERA5-Land									
sfcWind	Yes	ERA5-Land				PREVAH					IMDAA
hurs	No					PREVAH			DWD-HYRAS		IMDAA
tdps	Yes	ERA5-Land									

vp	Yes		Daymet, Maurer, NLDAS	AGCD, SILO							
vpDeficit	Yes			SILO							

AGCD: (Jones et al., 2009), **BR-DWGD v3.2.3:** (Xavier et al., 2022), **CEH-GEAR:** (Tanguy et al., 2021), **CHESS-Met:** (Robinson et al., 2017),**CHESS-PE:** (Robinson et al., 2023a), **CHIRPS V2.0:** (Funk et al., 2015), **CPC:** (Chen and Xie, 2008; CPC Global Unified Temperature, 2025), **CR2METv1.3:** (DGA, 2017), **Daymet:** (Thronton et al., 2012), **DWD-HYRAS:** (HYRAS – Hydrometeorologische Rasterdaten, 2024; Razafimaharo et al., 2020), **ERA5-Land:** (Muñoz Sabater, 2019), **GLEAM v3:** (Martens et al., 2017; Miralles et al., 2011), **GLEAM v4.2a:** (Miralles et al., 2025), **HadUK-GRID:** (Hollis et al., 2019), **Hydro-PE:** (Brown et al., 2023; Robinson et al., 2023b), **IMD**: (Pai et al., 2014), **IMDAA** : (Rani et al., 2021), **Maurer** : (Maurer et al., 2002), **MeteoSwiss2023:** (Federal Office of Meteorology and Climatology MeteoSwiss, 2023), **PREVAH**: (Viviroli et al., 2007), **MSWEP v1.1:** (Beck et al., 2017), **MSWEP v2.8:** (Beck et al., 2019), **MSWX:** (Araghi et al., 2023; Beck et al., 2022), **NLDAS:** (Xia et al., 2012), **Singer2021:** (Singer et al., 2021), **SILO:** (Jeffrey et al., 2001), **Srivastava2009:** (Srivastava et al., 2009).

Introduction

1. The statement describing Caravan-CMIP6 as covering "CAMELS-style datasets" should be clarified. It is not clear whether "CAMELS-style" refers to the catchment concept, temporal resolution, spatial representation, variable availability, or another characteristic.

   I will refer to these datasets instead as large-sample hydrometeorological datasets, as defined in the first paragraph of the introduction.

Methodology

Dataset description

1. Table S1 summarizes the original CAMELS datasets, whereas some variables are excluded in the present work. A summary of the datasets and variables actually included in Caravan-CMIP6 would improve clarity.

   I will provide this table, as described in General Comment 3.

2. The exclusion of snow-water equivalent appears to be primarily motivated by implementation difficulties associated with the bias-correction algorithm. Additional scientific justification should be provided together with discussion of the implications for users working in snow-dominated catchments.

   More justification will be added. There are two interrelated issues here: First, the Caravan data (ERA5-Land data) includes glaciers as a constant 10,000 mm of SWE. The climate models do not. Second, the ERA5-Land data contain temporal discontinuities.

   The first issue causes a problem for the bias correction algorithm because the observed and modelled data are fundamentally not reporting the same variable. This causes instabilities in the bias-corrected future snow. For example, imagine bias-correcting the following data point, which represents a specific quantile of the observed historical, modelled historical, and modelled projected data for a period in the summer:

$x_{o,h} = 7000 \text{ mm}$ (observed, historical) – this could be a catchment that is 70% glacier and the rest snow-free.

$x_{m,h} = 0.05 \text{ mm}$ (modelled, historical)

$x_{m,p} = 0.1 \text{ mm}$ (modelled, projected)

The bias-corrected value is thus calculated by multiplying the observed, historical data by the change in the modelled data:

$$\hat{x}_{m,p} = \frac{x_{m,p}}{x_{m,h}} * x_{o,h} = 14{,}000 \text{ mm}$$

This gives the unrealistic result that future SWE will be 14 m, because the very small values in the modelled data doubled.

Second, there are temporal discontinuities in the ERA5-Land data, which are thus also errors in the Caravan data. For example, see Figure R1 below, which shows the snow-water-equivalent data for camelscl_11532000, a partially glaciated catchment in southern Chile. The snow depth time series contains temporal discontinuities around 1990 and 2002, when certain ERA5-Land cells suddenly become (or cease to be) glaciers.

These discontinuities make it impossible to solve the first issue by simply subtracting the glacier contribution to SWE in the ERA5-Land before performing the bias correction.

Implications for users working in snow-dominated catchments: most process-based hydrologic models compute snow depth water equivalent internally, and do not rely on this as an external input. The omission of this variable does not prevent these models from being used to perform projections with Caravan-CMIP6.

Machine learning models *can* be trained using snow water equivalent as a predictor; such models would not be able to perform projections using Caravan-CMIP6.

More discussion will be added, including Figure R1 as an example in the appendix.

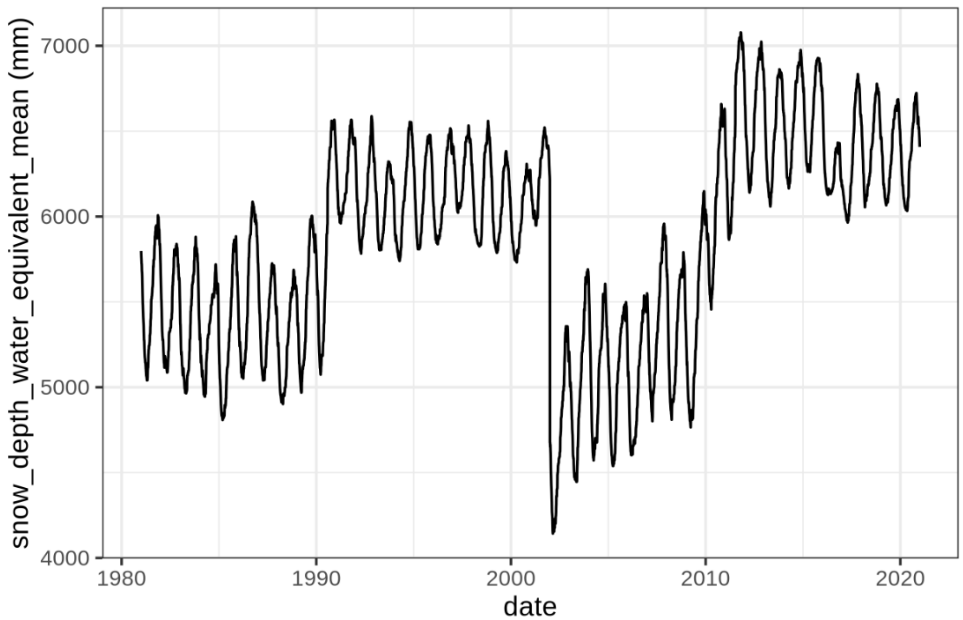

Figure R1: Snow-water-equivalent from CARAVAN for camelscl_11532000. Temporal discontinuities occur around 1990 and 2002, when certain grid cells within the catchment abruptly become (or cease to be) glaciers.

3. The manuscript states that spatial median, minimum, maximum, and standard deviation cannot be provided because of the coarse CMIP6 resolution. However, it is unclear what information is provided instead.

   Spatial means are provided. This will be clarified.

4. The sentence "I required each model and variant to report nine variables" should be clarified. It is not clear what is meant by "report" and how the model selection was performed.

   This will be clarified. ‘Report’ was a poor word choice – I required the model output to be available on the Earth System Grid Foundation federated nodes.

5. The statement "replacing the central 10 years within each 30-year window" is difficult to follow and should be explained more clearly.

   I will revise as follows, and include Figure R2 in the appendix

Following Cannon et al. (2022) the bias correction is applied using sliding windows both across and within years. Across years, I use 30-year sliding windows from 1850 to 2014 (historical experiment) and 2015 to 2100 (ssp126, ssp245, and ssp585 scenarios). Within years I divide each month into two halves and bias correct each half month separately, using a sliding window of three half months. This results in a sliding block of data of 30 years by three half months (see Figure R2). The bias-correction algorithm sees all these data but only the central 10 years and central half month are saved at each step.

Using equal-length slices in the historical (reference) and future periods ensures there is a one-to-one mapping between the sorted historical and projected values (Cannon, 2015). Extrapolation beyond the historical range is dealt with by preserving the relative or absolute changes in these values.

		2015-2019	2020-2029	2030-2039	2040-2049	2050-2059	2060-2069	2070-2079	2080-2089	2090-2100
Jan	1-15									
	16-31									
Feb	1-15									
	16-28									
Mar	1-15									
	16-31									
Apr	1-15									
	16-30									
May	1-15									
	16-31									
Jun	1-15									
	16-30									
Jul	1-15									
	16-31									
Aug	1-15									
	16-31									
Sep	1-15									
	16-30									
Oct	1-15									
	16-31									
Nov	1-15									
	16-30									
Dec	1-15									
	16-31									

Figure R2: illustration of sliding windows both within and across years. In this example the data to be bias-corrected are the values from May 16-31 over the decade 2040-2049. However, data from May 1-June 15 and from 2030-2059 are also included in the bias-correction algorithm, which improves the temporal consistency within years and across decades.

6. The temporal characteristics of the dataset should be introduced explicitly, including temporal resolution, temporal coverage, treatment of different CMIP6 calendars, and any harmonization procedures.
   At line 45 I will add:
   The data are provided at daily resolution from 1850-2100, in the native calendar of each climate model, which include the proleptic_gregorian, no _leap, and 360-day calendars defined by the CF conventions (Eaton et al., 2024).

Model selection

1. The discussion of model selection lacks coherence. The manuscript first discusses ensemble selection, then model modifications, and then abruptly moves to the "hot model" problem and model weighting. Consider separating these topics into distinct subsections.

   I will move the ECS/TCR discussion to a new section entitled "Usage recommendations and limitations" which will follow the Methods.

2. The discussion of ECS/TCR and the "hot model" issue appears disconnected from the construction of the dataset. It is not immediately clear whether this section describes methodological decisions used to build Caravan-CMIP6 or recommendations for future users.

   See previous comment.

3. The model weights provided in Table 1 require additional explanation. Their meaning, derivation, interpretation, and intended application should be described.

   At line 119 I will add:

   Users may take the weighted ensemble mean (or weighted quantiles) of projected impacts derived from the full ensemble from all models. Weighting ensures that models that represent unlikely climate change pathways have a smaller influence on results.

   I will also add a section in the appendix describing the model weighting strategy.

   Cannon (2024) proposed a method of estimating weights for an ensemble of climate models, based on matching the distribution of ECS values in the models to the assessed (theoretical) distribution of ECS (Sherwood et al., 2020). This method aims to optimize model weights to minimize the differences between the weighted model quantiles and the assessed quantiles of ECS. The method uses a quasi-Newton optimization algorithm (R Core team, 2025).

The optimization algorithm is run 50 times, and the average weight is taken. For each run a different (random) set of initial weights are provided, and a small amount of Gaussian noise is added to the ECS estimates for each model, $\epsilon \sim N(0, 0.1)$.

Figure R3 shows the unweighted and weighted model distribution of ECS, compared to the assessed distribution.

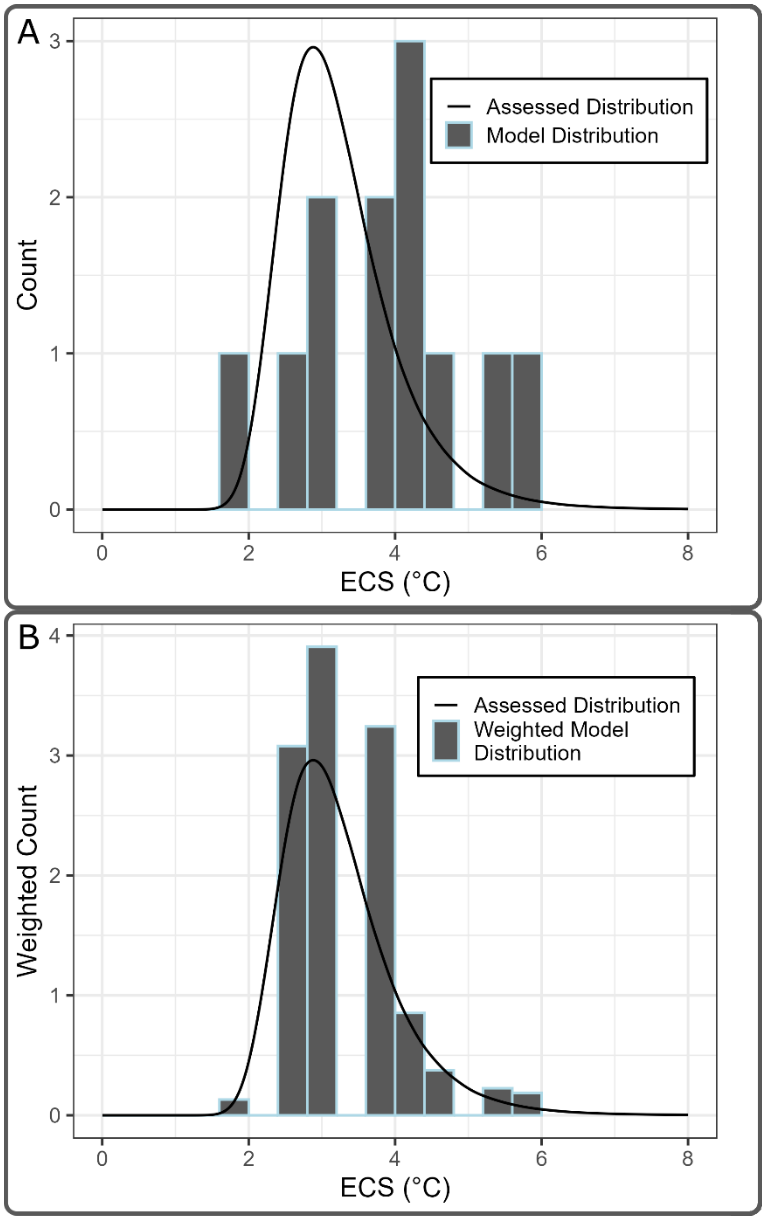

Figure R3: Effect of weighting on the model distribution of ECS. The unweighted ensemble model distribution of ECS is provided in (A), along with the assessed distribution of ECS from Sherwood et al. (2020). The model distribution is notably 'hot': it is shifted to the right of the assessed distribution. The weighted ensemble of ECS (B) matches the assessed distribution more closely.

Data processing

1. Section 2.2 is entitled "Raw CMIP6 model output extraction", although the section mainly describes basin-scale spatial aggregation rather than raw data extraction. A more appropriate section title would improve clarity.

   This section will be renamed "Basin-scale spatial aggregation of raw CMIP6 model output"

2. Basin averaging inevitably modifies the native spatial representation of the climate models, particularly for small catchments and heterogeneous terrain. This limitation should be acknowledged and discussed.

   I agree that basin averaging can reduce the information content of the data, but Caravan and all the CAMELS also provide basin-averaged meteorology. This limitation is thus not introduced by Caravan-CMIP6. I will add the following:

   Note that many of the catchments are much smaller than the grid resolutions of the CMIP6 models, which range from approximately 50 km to 500 km. The raw CMIP6 data may thus be poorly representative of the conditions within individual catchments.

3. Throughout Section 2.3, implementation details are often introduced before explaining why they are required. For example, basin elevation is introduced before explaining that it is needed to estimate surface pressure. Similar issues occur throughout the methodology. Reorganizing the workflow by first explaining the motivation and then describing the implementation would substantially improve readability.

   This section will be revised to ensure implementation details follow the motivation.

Bias correction

1. The rationale for selecting Quantile Delta Mapping (QDM) requires further discussion. QDM is a well-established bias-correction method but is not without limitations. In particular, it does not preserve multivariate dependence and, without multivariate extensions, does not preserve spatial and temporal dependence structures. The manuscript should explain why QDM was selected over other available approaches and discuss its limitations more explicitly.

   I will add the following near line 135:

   Many bias-correction algorithms exist, and their future performance is impossible to validate (Teutschbein and Seibert, 2012). Dinh and Aires (2023) reviewed available *univariate statistical bias-correction algorithms, and evaluated them based on (a) how well the bias corrected data for the historical reference period match the observed data, and (b) how well the bias-correction preserves the mean change signal in the climate model data. They found that the Equidistant Quantile Matching algorithm developed by Li et al. (2010) achieved both goals reasonably well. Cannon et al. (2015) showed that the additive form of Quantile Delta Mapping (QDM) is equivalent to Equidistant Quantile Matching, and QDM also supports a multiplicative correction. Cannon also showed that QDM is better at preserving changes in extreme values than two other commonly-used algorithms. QDM has been widely used and validated since its development, and mature software packages exist for its implementation (Cannon, 2024a).*

A multivariate version of QDM has also been developed (Multivariate Quantile Mapping Bias Correction, MBCn) (Cannon, 2018). MBCn preserves the dependence between variables, which can be important for hydrologic simulations. However, MBCn was not chosen for Caravan-CMIP6 for two reasons. First, MBCn is an iterative technique that proceeds until convergence criteria are met. This is computationally expensive, especially when many variables are being simultaneously corrected. Second, using a multivariate technique increases the risk of overfitting (Cannon, 2018) and also creates the possibility for errors in one variable to negatively affect the results for another variable. For example, if there are errors in the observed historical surface wind data, all of the bias-corrected variables would be potentially compromised.

2. The statement that variables "are not always provided in the same form" or "are not in the optimal form for bias correction" is unclear. Please clarify what is meant by "form" and whether these transformations are required because of differences between CMIP6 outputs and the reference datasets or because of assumptions of the selected bias-correction method.

   Both are true. For temperature, it is preferable to bias-correct the mean temperature and the diurnal range separately, to preserve the relationship between the mean, maximum, and minimum temperature (Thrasher et al., 2012). This corresponds to data ‘not in the optimal form for bias correction’

   For other variables (*eg.* rss, rls, tdps, ps, pet) the variables must be derived from CMIP6 output. This corresponds to data ‘not always provided in the same form’

   This section will be revised to clarify this distinction and the language. The section will be renamed “*Derivation of additional variables from raw climate model outputs*”

3. The temperature correction workflow is difficult to follow because the notation for bias-corrected variables is introduced before the bias-correction procedure itself is described.

   This comment is incorrect. The bias correction is described in section 2.3, while the workflow is in 2.3.1.

4. Earlier sections describe strict model selection criteria, whereas Section 2.3 repeatedly states that some variables are "not always provided". This creates unnecessary confusion. Please clarify that this refers to different representations of equivalent variables rather than missing variables.

   This will be clarified.

5. Several derived variables (dew point temperature, surface pressure, PET, etc.) are introduced without first explaining why they are required or how they relate to the objectives of Caravan-CMIP6.

   These variables will be introduced with reference to which CAMELS dataset(s) they are included in, and also with reference to Table R1.

6. The treatment of wind deserves further discussion. The manuscript acknowledges the difficulty of bias correcting the wind vector components but ultimately bias corrects only wind speed. Please clarify whether the released dataset contains only wind speed or also the vector components, and discuss the implications of not preserving wind direction.

   The released dataset only includes wind speed. The implication is that models that require x- and y- components of wind speed cannot be run on these data, and this will be explicitly stated. That being said, I am not aware of any lumped rainfall-runoff models that require wind vectors.

7. The derivation of PET products requires stronger scientific justification. Several methodological choices appear to be motivated primarily by implementation convenience (e.g., assumptions regarding Priestley-Taylor parameters or replacing cloud-cover-based radiation calculations because cloud cover was not downloaded). These choices should be supported by scientific references or discussion demonstrating that the approximations are appropriate and their implications are acceptable.

   For the Priestley-Taylor parameters, as noted in the manuscript, I assume G=0 and alpha =1. These are scaling factors which affect the multiplicative bias of the 'raw'

Priestly-Taylor PET. The multiplicative bias is fixed anyway by the bias correction, so there is no need to set these values. This will be made even more explicit.

For Morton's ET calculation, I will rephrase:

However, (Zajaczkowski and Jeffrey, 2020) estimated net radiation using cloud oktas, sunshine hours duration, and extraterrestrial radiation because observations of net shortwave and longwave radiation were not available. Since net shortwave and longwave radiation are available from CMIP6 models, these variables are used directly.

8. The physical limits imposed after bias correction should be better justified. It would also be useful to report how frequently these limits are activated and whether they have any measurable impact on the resulting dataset.

   The limits on humidity, vapour pressure, and dewpoint temperature are all based on the maximum moisture carrying capacity of air; the limits on temperature only require that the relationship between minimum, mean, and maximum temperature is not mathematically impossible; the limit on radiation is a highly conservative upper bound set by the sum of radiation arriving at the earth's upper atmosphere; the limit on precipitation is considerably higher than the maximum value ever observed on earth.

   The number of times each limit is activated is included in the metadata of each file. I will include some summary statistics in the text.

9. The observational or reference datasets used for bias correction across the different CAMELS datasets should be summarized more clearly.

   These are now included in Table R1.

Results and Discussion

1. The discussion states that the bias correction is performed separately for each half-month, although this temporal framework has not been introduced previously. This creates confusion regarding the implementation of the bias-correction procedure.

   The 'half-month' framework is introduced at Line 138, in the Methods. Also, see 'Dataset Description' comment 5, where I propose changes to the text. Hopefully the revised text makes this point clearer.

2. The description of return levels is difficult to follow. The rationale for estimating return levels using annual block maxima and empirical ranking should be better justified rather than being presented mainly as a matter of simplicity.

These points will be clarified.

Annual block maxima are a common, and simple, approach used to detect changes in hydrologic extremes (Slater et al., 2021). The simplicity is not only a matter of convenience, but also of reducing the number of subjective decisions involved in the analysis. This leads to a more transparent evaluation of the dataset.

Empirical ranking is chosen because the question Figures 2 and 3 are asking is how well the bias correction preserves changes in extremes. Thus, I am interested in whether the *actual* changes in extremes in the bias-corrected data are similar to the *actual* changes in the raw data. If I were to fit an extreme value function, it would be difficult to tell whether the bias correction or the extreme value function was to blame for any discrepancies.

3. The concepts of ratio and interval variables should be clearly defined before discussing the results.

   This concept is introduced at Line 141. I will remind the reader of this as follows:

   Figure 2 shows a comparison of the raw and bias-corrected (QDM) change in the mean and 10-year return period maximum values for the four ratio variables (for which the relative changes in the climate models are meant to be preserved), for all climate models and all catchments within the Caravan dataset. Figure 3 shows the same comparison, for the mean and 10-year return period minimum and maximum values, for the interval variables (for which the absolute changes in the climate models are meant to be preserved).

4. Figures 2 and 3 evaluate only selected variables. Please explain why the remaining variables are not presented or provide justification for the selected subset.

   Figures 2 and 3 evaluate all the variables in the Caravan dataset. Variables in each of the CAMELS datasets were also inspected but not included in the manuscript for brevity. The Caravan variables are broadly representative of the overall performance.

5. Figures 4 and 5 present projections only for the 3°C warming level, although four warming levels are introduced earlier in the manuscript. Please explain why only one warming level is shown.

   This is to keep the figure from having too many subplots. The objective of Figures 4 and 5 is to illustrate the potential value of the ensemble for projecting changes in floods or droughts and their uncertainty. As such, a single global warming level is sufficient, and that showing the differences between models is more informative than showing the differences between warming levels.

6. The current evaluation focuses primarily on whether QDM preserves the projected climate change signal by comparing the raw and bias-corrected CMIP6 outputs (Figures 2 and 3). While this is an important characteristic of QDM, it does not demonstrate how well the bias-corrected historical simulations reproduce the reference observations. Since Caravan-CMIP6 is intended to provide bias-corrected meteorological forcing, I recommend including a concise validation of the historical period against the observational datasets used for bias correction. Representative performance statistics (e.g., bias, RMSE, correlation, distributional agreement, or seasonal climatology) for selected variables and representative catchments or regions would provide users with greater confidence in the quality of the released dataset and complement the current evaluation.

   I will add the following section, "***Match to Observed Reference Data***"

   Match to Observed Reference Data

   Apart from preserving changes in variables from the climate model simulations, the other goal of bias correction is to match the statistical properties of the observed reference data. This section evaluates how well the bias-corrected data match the overall distribution and the seasonality of the observed data for the reference period of 1981-2010. The Caravan data are used, but the results are representative of the behaviour for the CAMELS datasets as well.

   Distribution

   Quantile Delta Mapping is designed to bias-correct each quantile of the modelled data to exactly match the observed data (Cannon et al., 2015). However, the use of sliding windows to perform the bias correction can lead to deviations from this ideal behaviour. Here I verify that the observed distribution for the reference period is reproduced by the bias-corrected data.

   The first column of Figure R4 shows the agreement between the observations and bias-corrected simulations for 1981-2010. The left column shows the Wasserstein Distance between the two distributions for all catchments and climate models (n=1.12 million), for each variable in Caravan. The Wasserstein Distance (Vaserstein, 1969), also called the 'Earth Mover's Distance', is the minimum 'cost' required to transform one distribution to another, and has the same units as the underlying data. It was calculated with the transport package in R (Schuhmacher et al., 2024). The location of the median, 99.9th percentile, and maximum values are indicated, and columns two to four show the observed and bias-corrected histograms for the catchments corresponding to these percentiles.

In general, the distributional agreement is excellent, even for the worst case for each variable (the maximum Wasserstein Distance). However, for rss and petfao56, the worst case shows noticeable disagreement at the upper end of the distribution. These worst cases correspond to the same catchment and climate model (camels_04043050: Trap Rock River Near Lake Linden, MI, climate model MRI-ESM2-0, variant r2i1p1f1). The error originates from an error in the climate model rss variable around the North American Great Lakes, which will be catalogued in "Appendix A: Miscellaneous Notes".

For sfcWind, the worst match is for the gauge GRDC_3618700 and the climate model MIROC6. This is a gauge in northern Brazil on the Cotingo River, and large biases in the raw climate model data are to blame for this poor match. The raw data indicate very low 10-metre wind speeds, never exceeding 0.07 m/s in the historical period and include a strong trend towards increasing wind speeds, from a mean of 5.5E-5 in the 1850's to a mean of 2.6E-2 in the 2000's (a 500-fold increase). The small values and the unrealistic trend over the reference period are the cause of the poor performance of the bias correction algorithm for this case.

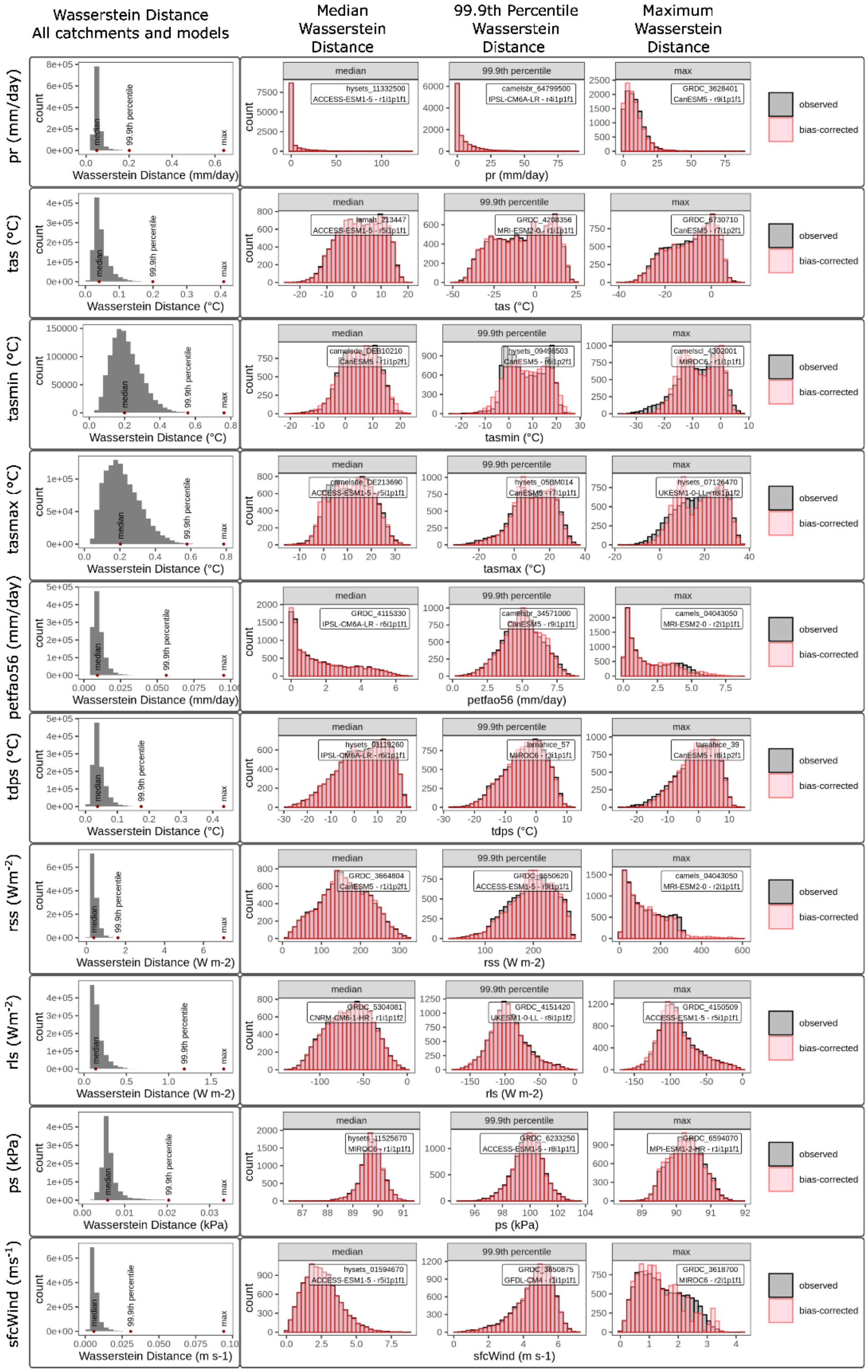

Figure R4: Match of bias-corrected data to observations for the reference period. The left column shows the distribution of the Wasserstein Distance (a measure of distributional similarity) for all catchments and climate models in the Caravan dataset. Columns 2-4 show the observed and bias-corrected distributions of each variable for the catchment/climate model at the median, 99.9th percentile, and maximum of the Wasserstein Distance.

Seasonality

The bias correction procedure is also intended to ensure that the seasonality of the bias-corrected data matches the observed reference seasonality. This is achieved by bias-correcting each half-month separately.

Here I evaluate the match between the observed and bias-corrected monthly climatological mean values of each of the ten variables in Caravan. The first column of Figure R5 shows the coefficient of determination (R^2). Similar to Figure R4, Figure R5 indicates the median performance, the performance achieved by 99.9% of models and catchments (0.1th percentile R^2), and the worst-performing case (minimum R^2), and shows these examples in columns two to four.

The bias-correction produces good matches to the observed seasonality in almost all cases. However, precipitation is the most challenging variable for which to match the seasonality, because its episodic nature and the use of sliding windows creates the potential for the bias-correction to 'bleed' across the half-month boundaries. This produces the jitter observable in row one, columns three and four, for catchments that do not have a strongly seasonal precipitation regime.

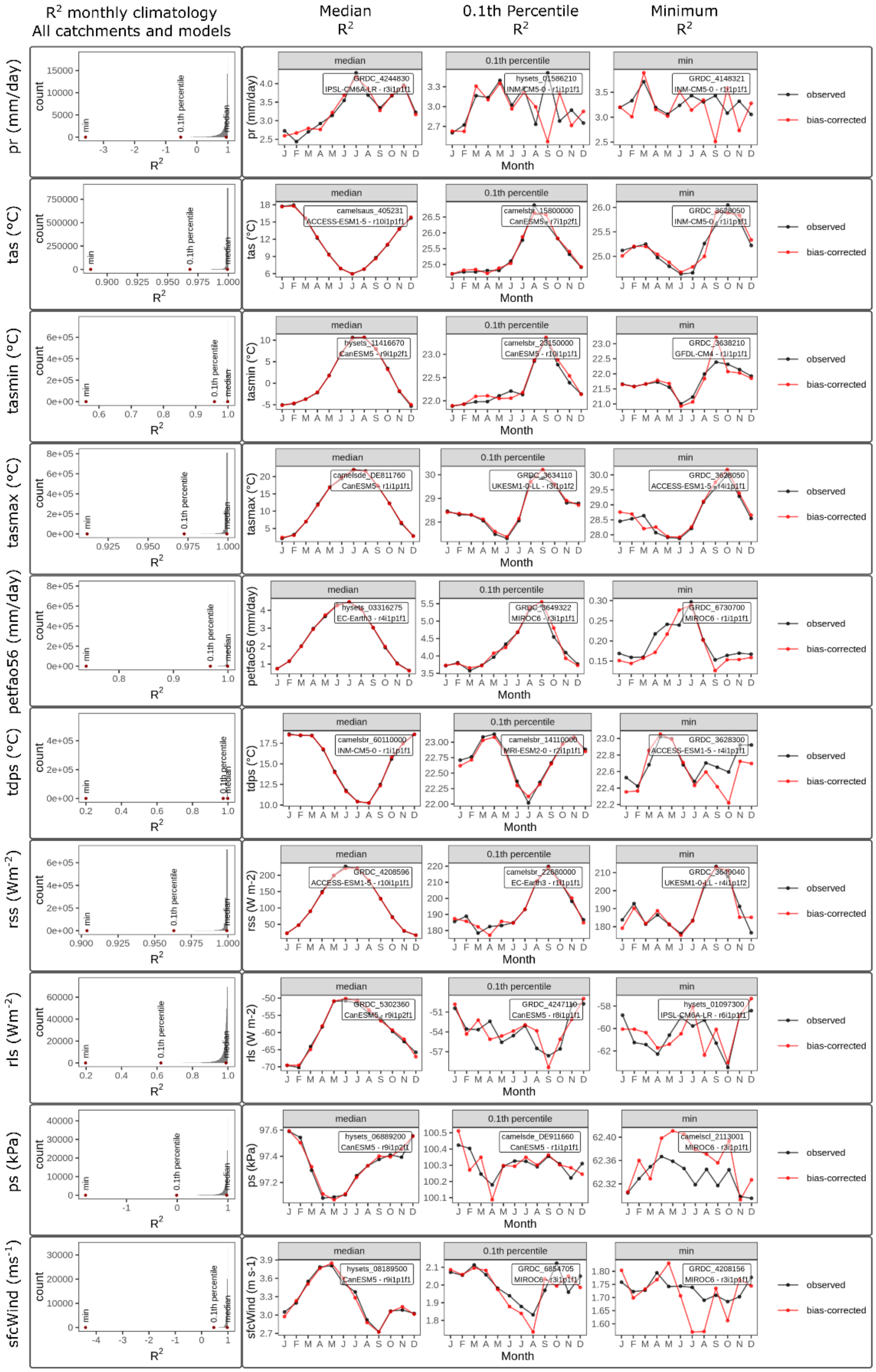

Figure R5: Match of bias-corrected data to seasonality of observed data. The first column shows the coefficient of determination R^2 between the observed and bias-corrected monthly climatological (long-term) means. Columns two to four show examples of the median, 0.1th percentile, and minimum (worst) R^2.

Conclusions

1. The opening paragraph discusses projected streamflow changes, although streamflow itself has not been analysed in this study. The conclusion should better reflect the actual scope of the manuscript.

   I will instead refer to 'hydrometeorological projections'.

2. The justification for not adopting multivariate bias correction focuses primarily on computational cost. A stronger scientific justification should be provided, together with discussion of the methodological trade-offs.

   See 'Bias Correction' comment 1.

3. The statement that Caravan-CMIP6 is biased toward Europe and North America has not been analysed or discussed elsewhere in the manuscript. This limitation should either be discussed earlier or omitted from the conclusion.

   This statement will be removed.

Overall, I believe the dataset has significant value for the hydrological community. However, the manuscript requires substantial revision to improve the clarity of the methodology, strengthen the scientific justification for several processing decisions, and provide a more complete evaluation of the released bias-corrected dataset. Addressing these issues would considerably improve the reproducibility and usability of Caravan-CMIP6 as a long-term community resource.

References

Araghi, A., Martinez, C. J., and Olesen, J. E.: Evaluation of MSWX gridded data for modeling of wheat performance across Iran, European Journal of Agronomy, 144, 126769, https://doi.org/10.1016/j.eja.2023.126769, 2023.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21, 589–615, https://doi.org/10.5194/hess-21-589-2017, 2017.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Dijk, A. I. J. M. van, McVicar, T. R., and Adler, R. F.: MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Beck, H. E., Dijk, A. I. J. M. van, Larraondo, P. R., McVicar, T. R., Pan, M., Dutra, E., and Miralles, D. G.: MSWX: Global 3-Hourly 0.1° Bias-Corrected Meteorological Data Including Near-Real-Time Updates and Forecast Ensembles, https://doi.org/10.1175/BAMS-D-21-0145.1, 2022.

Brown, M. J., Robinson, E. L., Kay, A. L., Chapman, R., Bell, V. A., and Blyth, E. M.: Potential evapotranspiration derived from HadUK-Grid 1km gridded climate observations 1969-2022 (Hydro-PE HadUK-Grid), 2023.

Cannon, A. J.: Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables, Clim Dyn, 50, 31–49, https://doi.org/10.1007/s00382-017-3580-6, 2018.

Cannon, A. J.: MBC: Multivariate Bias Correction of Climate Model Outputs, , https://doi.org/10.32614/CRAN.package.MBC, 2024a.

Cannon, A. J.: The Impact of “Hot Models” on a CMIP6 Ensemble Used by Climate Service Providers in Canada: Do Global Constraints Lead to Appreciable Differences in Regional Projections?, Journal of Climate, 37, 2141–2154, https://doi.org/10.1175/JCLI-D-23-0459.1, 2024b.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes?, Journal of Climate, 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, 2015.

Chen, M. and Xie, P.: CPC unified gauge-based analysis of global daily precipitation, Western Pacific Geophysics Meeting, 2008.

DGA: Actualización del Balance Hídrico Nacional, Ministerio de Obras Públicas, Dirección General de Aguas, División de Estudios y Planificación, Santiago, Chile, 2017.

Dinh, T. L. A. and Aires, F.: Revisiting the bias correction of climate models for impact studies, Climatic Change, 176, 140, https://doi.org/10.1007/s10584-023-03597-y, 2023.

HYRAS – Hydrometeorologische Rasterdaten: https://www.dwd.de/DE/leistungen/hyras/hyras.html, last access: 25 March 2024.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Blower, J., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., Hassell, D., Snow, A. D., Kölling, T., Allured, D., Jelenak, A., Soerensen, A. M., Gaultier, L., Herlédan, S., Manzano, F., Bärring, L., Barker, C., and Bartholomew, S. L.: NetCDF Climate and Forecast (CF) Metadata Conventions, https://doi.org/10.5281/zenodo.14275599, 2024.

Federal Office of Meteorology and Climatology MeteoSwiss: Weather Forecast Switzerland, 2023.

Fowler, K. J. A., Zhang, Z., and Hou, X.: CAMELS-AUS v2: updated hydrometeorological time series and landscape attributes for an enlarged set of catchments in Australia, Earth System Science Data, 17, 4079–4095, https://doi.org/10.5194/essd-17-4079-2025, 2025.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, Sci Data, 2, 150066, https://doi.org/10.1038/sdata.2015.66, 2015.

Hollis, D., McCarthy, M., Kendon, M., Legg, T., and Simpson, I.: HadUK-Grid—A new UK dataset of gridded climate observations, Geoscience Data Journal, 6, 151–159, https://doi.org/10.1002/gdj3.78, 2019.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., and Beswick, A. R.: Using spatial interpolation to construct a comprehensive archive of Australian climate data, Environmental Modelling & Software, 16, 309–330, https://doi.org/10.1016/S1364-8152(01)00008-1, 2001.

Jones, D. A., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, Australian Meteorological and Oceanographic Journal, 58, 233–248, https://doi.org/10.1071/ES09032, 2009.

Li, H., Sheffield, J., and Wood, E. F.: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012882, 2010.

Mahony, C. R., Wang, T., Hamann, A., and Cannon, A. J.: A global climate model ensemble for downscaled monthly climate normals over North America, Intl Journal of Climatology, 42, 5871–5891, https://doi.org/10.1002/joc.7566, 2022.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States*, J. Climate, 15, 3237–3251, https://doi.org/10.1175/1520-0442(2002)015%3C3237:ALTHBD%3E2.0.CO;2, 2002.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. a. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

Miralles, D. G., Bonte, O., Koppa, A., Baez-Villanueva, O. M., Tronquo, E., Zhong, F., Beck, H. E., Hulsman, P., Dorigo, W., Verhoest, N. E. C., and Haghdoost, S.: GLEAM4: global land evaporation and soil moisture dataset at 0.1° resolution from 1980 to near present, Sci Data, 12, 416, https://doi.org/10.1038/s41597-025-04610-y, 2025.

Morton, F. I.: Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, Journal of Hydrology, 66, 1–76, https://doi.org/10.1016/0022-1694(83)90177-4, .

Muñoz Sabater, J.: ERA5 Land hourly time-series data from 1950 to present, https://doi.org/10.24381/cds.e2161bac, 2019.

CPC Global Unified Temperature: https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html, last access: 27 May 2025.

Pai, D. S., Rajeevan, M., Sreejith, O. P., Mukhopadhyay, B., and Satbha, N. S.: Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region, MAUSAM, 65, 1–18, https://doi.org/10.54302/mausam.v65i1.851, 2014.

R Core team: R: A Language and Environment for Statistical Computing, 2025.

Rani, S. I., Arulalan, T., George, J. P., Rajagopal, E. N., Renshaw, R., Maycock, A., Barker, D. M., and Rajeevan, M.: IMDAA: High-Resolution Satellite-Era Reanalysis for the Indian Monsoon Region, https://doi.org/10.1175/JCLI-D-20-0412.1, 2021.

Razafimaharo, C., Krähenmann, S., Höpp, S., Rauthe, M., and Deutschländer, T.: New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS), Theor Appl Climatol, 142, 1531–1553, https://doi.org/10.1007/s00704-020-03388-w, 2020.

Robinson, E. L., Blyth, E., Clark, D. B., Comyn-Platt, E., Finch, J., and Rudd, A. C.: Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961-2015) [CHESS-met] v1.2, 2017.

Robinson, E. L., Blyth, E. M., Clark, D. B., Comyn-Platt, E., Rudd, A. C., and Wiggins, M.: Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2019) [CHESS-PE], 2023a.

Robinson, E. L., Brown, M. J., Kay, A. L., Lane, R. A., Chapman, R., Bell, V. A., and Blyth, E. M.: Hydro-PE: gridded datasets of historical and future Penman–Monteith potential evaporation for the United Kingdom, Earth System Science Data, 15, 4433–4461, https://doi.org/10.5194/essd-15-4433-2023, 2023b.

Schuhmacher, D., diagrams), B. B. (aha and power, Bonneel (networkflow), N., shortlist), C. G. (simplex and, Hartmann (semidiscrete1), V., integration), F. H. (transport_track and networkflow, Schmitzer (shielding), B., Schrieber (subsampling), J., and Wilm (wpp), T.: transport: Computation of Optimal Transport Plans and Wasserstein Distances, 2024.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Reviews of Geophysics, 58, e2019RG000678, https://doi.org/10.1029/2019RG000678, 2020.

Singer, M. B., Asfaw, D. T., Rosolem, R., Cuthbert, M. O., Miralles, D. G., MacLeod, D., Quichimbo, E. A., and Michaelides, K.: Hourly potential evapotranspiration at 0.1° resolution for the global land surface from 1981-present, Sci Data, 8, 224, https://doi.org/10.1038/s41597-021-01003-9, 2021.

Slater, L. J., Anderson, B., Buechel, M., Dadson, S., Han, S., Harrigan, S., Kelder, T., Kowal, K., Lees, T., Matthews, T., Murphy, C., and Wilby, R. L.: Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management, Hydrology and Earth System Sciences, 25, 3897–3935, https://doi.org/10.5194/hess-25-3897-2021, 2021.

Srivastava, A. K., Rajeevan, M., and Kshirsagar, S. R.: Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region, Atmospheric Science Letters, 10, 249–254, https://doi.org/10.1002/asl.232, 2009.

Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D. G., and Keller, V. D. J.: Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019) [CEH-GEAR], 2021.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of Hydrology, 456–457, 12–29, https://doi.org/10.1016/j.jhydrol.2012.05.052, 2012.

Thrasher, B., Maurer, E. P., McKellar, C., and Duffy, P. B.: Technical Note: Bias correcting climate model simulated daily temperature extremes with quantile mapping, Hydrology and Earth System Sciences, 16, 3309–3314, https://doi.org/10.5194/hess-16-3309-2012, 2012.

Thronton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., and Cook, R. B.: Daymet: Daily surface weather on a 1km grid for North America, 1980–2012, 2012.

Vaserstein, L. N.: Markov processes over denumerable products of spaces, describing large systems of automata, Problemy Peredači Informacii, 5, 64–72, 1969.

Viviroli, D., Gurtz, J., and Zappa, M.: The Hydrological Modelling System PREVAH Part I – Overview and Selected Applications, University of Berne, Switzerland, https://doi.org/https://doi.org/10.4480/GB2022.P40, 2007.

Xavier, A. C., Scanlon, B. R., King, C. W., and Alves, A. I.: New improved Brazilian daily weather gridded data (1961–2020), International Journal of Climatology, 42, 8390–8404, https://doi.org/10.1002/joc.7731, 2022.

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Lettenmaier, D., Koren, V., Duan, Q., Mo, K., Fan, Y., and Mocko, D.: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011JD016048, 2012.

Zajaczkowski, J. and Jeffrey, S.: Potential evaporation and evapotranspiration data provided by SILO, Department of Environment and Science, Queensland Government, 2020.