~~HESS Opinions:~~ Large-sample hydrology – A few camels or a whole caravan?

Franziska ~~Maria~~ Clerc-Schwarzenbach[1], Giovanni Selleri[2], Mattia Neri[2], Elena Toth[2], Ilja van Meerveld[1], Jan Seibert[1]

[1]Department of Geography, University of Zurich, Zurich, 8057, Switzerland
[2]Department of Civil, Chemical, Environmental, and Materials Engineering, University of Bologna, Bologna, 40136, Italy

Correspondence to: Franziska Clerc (franziska.clerc@geo.uzh.ch)

Abstract. Large-sample datasets containing hydrometeorological time series and catchment attributes for hundreds of catchments in a country, many of them known as "Camels" (catchment attributes and meteorology for large-sample studies), have revolutionized hydrological modelling and enabled comparative analyses. The Caravan dataset is a compilation of several ("Camels" and other) large-sample datasets with uniform attribute names and data structure. This simplifies large-sample hydrology across regions, continents, or the globe. However, the use of the Caravan dataset instead of the original Camels or other large-sample datasets may affect model results and the conclusions derived thereof. For the Caravan dataset, the meteorological forcing data are based on ERA5-Land reanalysis data. Here, we describe the differences between the original precipitation, temperature, and potential evapotranspiration (E_{pot}) data for 1252 catchments in the CAMELS-US, CAMELS-BR, and CAMELS-GB datasets and the forcing data for these catchments in the Caravan dataset. The E_{pot} in the Caravan dataset is unrealistically high for many catchments but there are, not surprisingly, also considerable differences in the precipitation data. We show that the use of the forcing data from the Caravan dataset impairs hydrological model calibration for the vast majority of catchments, i.e., there is a drop in the calibration performance when using the forcing data from the Caravan dataset compared to the original Camels datasets. This drop is mainly due to the differences in the precipitation data. Therefore, we suggest extending the Caravan dataset with the forcing data included in the original Camels datasets wherever possible, so that users can choose which forcing data they want to use, or at least indicating clearly that the forcing data in Caravan come with a data quality loss and using the original datasets is recommended. Moreover, we suggest not using the E_{pot} data (and derived catchment attributes, such as the aridity index) from the Caravan dataset and replacing these with (or based on) alternative E_{pot} estimates.

1 Large-sample datasets as a game changer in hydrological modelling studies

Starting with the CAMELS (Catchment Attributes and MEteorology for Large-sample Studies) dataset for the US (Addor et al., 2017; Newman et al., 2015; in this paper referred to as CAMELS-US), large-sample datasets have been developed for several other countries (e.g., CAMELS-CL for Chile (Alvarez-Garreton et al., 2018), CAMELS-BR (Chagas et al., 2020a) and

CABra (Almagro et al., 2021) for Brazil, CAMELS-GB for Great Britain (Coxon et al., 2020a), or CAMELS-CH for Switzerland (Höge et al., 2023)). We refer to these datasets with the time series of hydrometeorological measurements and information on catchment attributes for hundreds of catchments as the Camels datasets. Because computational power had increased and cloud-computing had advanced when these datasets became available, hydrological models can now be run for hundreds of catchments in a reasonable timeframe. The Camels datasets offer new opportunities for catchment model- and comparison studies because they minimize the effort that is needed to compile and check hydrometeorological data from different datasets. This is a great progress, not only for individual studies, but also for the comparability of different modelling approaches because comparisons are easier when different research groups use the same data for the same sets of catchments.

The Camels datasets have so far been used for different purposes. Examples are the exploration of the predictability of hydrologic signatures (Addor et al., 2018), the use thereof to cluster similar catchments and to explore their behaviour (Jehn et al., 2020), and to analyse the influence of catchment characteristics on runoff processes (Mathai and Mujumdar, 2022; McMillan et al., 2022). The datasets have also been used to conceptualize models, e.g., to determine subsurface flow contributions to the hydrograph (Ranjram and Craig, 2022), to assess the value of limited or alternative data for regionalization (Pool et al., 2019, 2021) or hydrological model calibration (Meyer Oliveira et al., 2023), and to test the influence of changes in the meteorological forcing data on model performance (van Beusekom et al., 2022; Deng et al., 2024). They have, furthermore, been used to train long short-term memory models (Gauch et al., 2021; Kratzert et al., 2024; Lees et al., 2021).

The Caravan dataset (Kratzert et al., 2023a) goes further than the Camels datasets. As indicated by the name, referring to a group of camels, it is a compilation of (subsets of) large-sample datasets released earlier. When the Caravan dataset was released, it included the national datasets CAMELS-US (Addor et al., 2017b), CAMELS-BR (Chagas et al., 2020a), CAMELS-GB (Coxon et al., 2020a), CAMELS-CL (Alvarez-Garreton et al., 2018), and CAMELS-AUS (Fowler et al., 2021), the North American dataset HYSETS (Arsenault et al., 2020), and the Central European dataset LamaH-CE (Klingler et al., 2021). The Caravan dataset not only combined parts of these existing datasets but also solved issues related to the lack of comparability among the different datasets, and the lack of an index referring to human impacts for some of the datasets (Addor et al., 2020). The use of the globally available ERA5-Land (European ReAnalysis) data (Muñoz-Sabater et al., 2021) for all catchments in the Caravan dataset furthermore allows the extension of the dataset with catchments for which streamflow but no meteorological data are available. With this possibility, Caravan allows catchments in underrepresented (climatic) regions to be included in a well-known large-sample dataset. This is positive and may be a first step towards a more equal representation of different regions and biogeoclimatic zones in hydrological research. Because of the use of reanalysis data for the forcing data, it is easier to update the Caravan dataset with additional forcing data, or a new version thereof, than when station data are used. Another advantage of the Caravan dataset as a standard resource for catchment data is that some of the catchments added by the community are not available as individual Camels datasets, i.e., the attributes and hydrometeorological time series for these catchments can only be accessed via the Caravan dataset. Thanks to the open code and software, the Caravan dataset can be extended by the community. The number of catchments in the Caravan dataset had already grown to almost 13,000 (not counting dupli-

cates) in February 2024. Acquiring data from Caravan is the easiest way to get started for large-sample model studies. How-
ever, we argue that using the Caravan data set instead of the individual Camels datasets may have disadvantages despite the
65 obvious advantage of the convenience of using one large dataset instead of the individual datasets.

Following the Caravan philosophy of using the same data source for all catchments for all climatic variables, the meteorolog-
ical forcing data in the original Camels datasets were replaced by reanalysis data from ERA5-Land. ERA5-Land (Muñoz-
Sabater et al., 2021) is a component of the Copernicus Climate Change Service (C3S). With ERA5-Land, global time series
of the water and energy cycle over land are described with 50 different variables. Compared to the earlier products ERA5 (31
70 km; Hersbach et al., 2020) and ERA-Interim (80 km; Dee et al., 2011), the spatial resolution (9 km) and the representation of
the water cycle improved for ERA5-Land (Muñoz-Sabater et al., 2021). However, ERA5-Land tends to overestimate potential
evapotranspiration (E_{pot}) significantly considerably (Klingler et al., 2021; Xu et al., 2024). E_{pot} is computed differently in
ERA5-Land than in ERA5 (as per rectification in the (ERA5-Land: data documentation, (2024) on 18 November 2021). In
ERA5, vegetated land is set to "crops/mixed farming" and it is assumed that there is no soil moisture limitation for the com-
75 putation of E_{pot}. In ERA5-Land, evaporation from an open water surface (i.e., pan evaporation) is computed. The atmosphere
is assumed to be unaffected by the evaporation for both ERA5 and ERA5-Land.

In this paper, we describe the differences between the meteorological forcing data for the catchments in three Camels datasets
(CAMELS-US, CAMELS-BR, and CAMELS-GB) and the ERA5-Land data in the Caravan dataset. We, furthermore, assess
the consequences of the substitution of largely station-based data in the Camels datasets by the reanalysis data in the Caravan
80 dataset on the calibration performance of a bucket-type rainfall-runoff model. It is important to raise awareness of these dif-
ferences, and their consequences on model results because the well-organized data structure and ease of access make it very
tempting to use the Caravan dataset instead of the original Camels datasets, especially when conducting studies across multiple
countries or geographic regions.

2   Caravan forcing data based on ERA5-Land

85 In the original Camels (and other large-sample) datasets, the forcing data were selected with respect to the data availability for
the region of interest. They were mainly based on station data, but for some regions, they also included satellite data or rea-
nalysis data (Table 1). In most cases, several forcing data time series were included to allow the user to choose the most suitable
one or to allow a comparison between different data inputs. When a catchment is added to Caravan, all forcing data are replaced
with data from the ERA5-Land reanalysis dataset (Muñoz-Sabater et al., 2021).

90 Several studies have assessed the ERA5-Land reanalysis data by comparing it to station data. ERA5-Land temperature and
precipitation data were found to better match the observations for flatter regions than for regions with complex terrain (Almeida
and Coelho, 2023; Gomis-Cebolla et al., 2023; Tan et al., 2023). Temperature data from ERA5-Land were considered to be
good for Portugal (Almeida and Coelho, 2023), northeastern Brazil (Araújo et al., 2022), the Chinese Qilian mountains (Zhao

and He, 2022), and Italy (Vanella et al., 2022). For Turkey, ERA5-Land underestimated the daily temperature, but represen-
tated temperature trends well (Yilmaz, 2023). For the Kelantan basin in Malaysia, the daily maximum temperatures were
underestimated and the daily minimum temperatures were overestimated (Tan et al., 2023). In their evaluation of ERA5-Land
data for Italy, (Vanella et al., (2022) found that the variables of ERA5-Land can be used to estimate evapotranspiration. Re-
garding precipitation, (Gomis-Cebolla et al., (2023) found that ERA5-Land represented the spatial and temporal precipitation
patterns well for Spain, but also that there were some difficulties in representing complex precipitation patterns. They further-
more found that ERA5-Land tended to overestimate light precipitation events and underestimated heavier precipitation. This
was also observed for the Tibetan Plateau (Wu et al., 2023) and the Kelantan basin in Malaysia (Tan et al., 2023). For the
Tibetan Plateau, the overestimation of light precipitation led to an overestimation of annual precipitation (Wu et al., 2023).
ERA5-Land also overestimated precipitation for China (Xie et al., 2022), but there were regional differences. ERA5-Land
represented precipitation for northeastern China better than for southwestern China (Xie et al., 2022).

A number of previous studies have analysed the advantages and disadvantages of the gridded products of the ERA family,
when used as forcing data in hydrological models. For example, Beck et al. (2017) included ERA-Interim data in a comparison
of different precipitation products with gauge data. They found a reasonable agreement between the ERA-Interim data and the
gauged data for all regions of the world, except northern South America, Africa, Central Asia and Southeast Asia. Essou et al.
(2016, 2017) compared different reanalysis products (including ERA-Interim) for North America and found that the datasets
had similar temperature data, but that there was a bias in precipitation for the humid continental and subtropical regions (i.e.,
for the eastern part of the US) and that this led to a deterioration in model performance (Essou et al., 2016). However, the
reanalysis data performed better than gridded data for large and mountainous catchments, where the density of weather stations
is low (Essou et al., 2017). Based on these findings, they suggested using reanalyses as meteorological forcing data when
observational data are missing or limited. Similarly, Tarek et al. (2020) tested ERA5 temperature and precipitation data for
hydrological modelling in North America. They found a clear improvement in model performance compared to ERA-Interim
data, and that model performances were similar to those achieved with observational data, except for the eastern half of the
US. They concluded that ERA5 data are useful, especially when observational data are lacking. Baez-Villanueva et al. (2021)
compared ERA5 precipitation data and three other precipitation products for Chile and found a similar model performance for
ERA5 data and some of the gauge-corrected precipitation products. However, they also reported some difficulties with ERA5
data for snow-dominated catchments.

3 Assessment of the differences between Camels and Caravan forcing data

3.1 Choice of catchments and climate variables

We compared the precipitation, temperature, and potential evapotranspiration (E_{pot}) data for 1252 catchments in the Caravan
dataset with the original forcing data from the CAMELS-US, CAMELS-BR, and CAMELS-GB datasets. We chose precipita-
tion, temperature, and E_{pot} data for the comparisons because they are the most relevant for hydrological modelling. From the

different Camels forcing datasets, we chose those with the highest spatial resolution (see Table 1~~Table 1~~), except for precipitation for the Brazilian catchments (as described below). The period for the comparisons ranged from April 1983 to March 2013 for the Brazilian catchments (southern hemisphere) and from October 1983 to September 2013 for the catchments in the US and Great Britain (northern hemisphere) to account for the differences in the water year for the two hemispheres.

130 For each catchment we compared the mean annual precipitation, the mean daily temperature and the mean annual E_{pot}. We only compared the mean annual values, even though there are other components of the time series, such as the timing of the rainfall events, that are also crucial for hydrological modelling. To account for differences in the data other than the mean values, we used a hydrological modelling approach (see Sect. 4) that implicitly takes into account all the features of the forcing time series through the simulation of streamflow. For hydrological modelling, the temperature data "per se" (i.e., when not

135 considered as the driver of E_{pot}) are mainly relevant for snow-related processes, i.e., to determine if precipitation is falling as snow (and is thus stored in the catchment) and if the precipitation that accumulated as snow is melting. Hence, the accuracy of the temperature data is relevant for only a few days per year for catchments where snow is an essential component of the water balance. In other words, the temperature plays a minor role for hydrological modelling compared to the accuracy of the precipitation or E_{pot} data (cf. Tarek et al., 2020). Still, we compared the temperature data for all catchments and focused on the

140 mean daily temperature (rather than, for example, the number of days with temperatures below or above 0 °C).

When we compared the two datasets, we always subtracted the value from the Camels dataset from the value from the Caravan dataset (i.e., a positive difference indicates a larger value for the Caravan dataset and a negative difference indicates a smaller value for the Caravan dataset). To determine the relative differences (i.e., for the mean annual precipitation and E_{pot}), we divided this difference by the value from the Camels dataset and report it as a percentage. As the catchment characteristics that

145 depend on the meteorological data also differ for the Caravan and Camels datasets, we furthermore compared the differences in the aridity index (E_{pot}/P).

3.2   Choice of Camels forcing data

The Camels forcing data we used for comparison had a spatial resolution of 1 km for CAMELS-US and CAMELS-GB, and a coarser resolution for CAMELS-BR (Table 1~~Table 1~~). For the US catchments, we used the Daymet v2 data (Thornton et al.,

150 2014; see also Thornton et al., 2021) for precipitation and temperature (the mean daily temperature was estimated from the average of the daily minimum and maximum temperature). As E_{pot} data are not available in the CAMELS-US dataset, we calculated E_{pot} with the Priestley-Taylor formula (Priestley and Taylor, 1972) based on the input data from Daymet v2. This is in line with the suggestion by Newman et al. (2015) and similar to the approach used in earlier studies with CAMELS-US data (e.g., Seibert and Vis (2016) and Addor et al. (2018)). As input data for the E_{pot} calculations, we used the elevation and latitude

155 of each catchment, and the time series of the day of the year, day length, minimum and maximum temperature, vapor pressure, and solar radiation. The Priestley-Taylor coefficient was set to 1.26 (cf. Priestley and Taylor, 1972) for all catchments. For the catchments in Brazil (BR), we used the MSWEP v2.2 precipitation data (Beck et al., 2019), the CPC temperature data (NOAA, 2019), and the GLEAM v3.3a E_{pot} data (Martens et al., 2017; Miralles et al., 2011), which are based on the Priestley-Taylor

formula with satellite-derived radiation and air temperature data. We chose MSWEP v2.2 data for the precipitation instead of
160 CHIRPS (Funk et al., 2015) because the MSWEP v2.2 daily time series are based on a data point every three hours, and the
ones from CHIRPS are based on one data point every five days, disaggregated to daily values via reanalysis. For the catchments
in Great Britain (GB), we used the CEH-GEAR precipitation data (Keller et al., 2015; Tanguy et al., 2016), the CHESS-met
temperature data (Robinson et al., 2017a), and the CHESS-PE E_{pot} data (Robinson et al., 2016, 2017b), which are based on the
Penman-Monteith formula, with meteorological data obtained from stations.

165

Table 1: Meteorological source datasets from the Camels datasets and the Caravan dataset used for comparison.

Region	Variable(s)	Dataset	Spatial resolution	References
US	Precipitation, Temperature	Daymet v2 (based on station data)	1 km	Thornton et al. (2014, 2021)
BR	Precipitation	MSWEP v2.2 (based on station-, satellite-, and reanalysis data)	0.1 °	Beck et al. (2019)
BR	Temperature	CPC (based on station data)	0.5 °	NOAA (2019)
BR	E_{pot}	GLEAM v3.3a (Priestley-Taylor method based on satellite data)	0.25 °	Martens et al. (2017), Miralles et al. (2011)
GB	Precipitation	CEH-GEAR (based on station data)	1 km	Keller et al. (2015), Tanguy et al. (2016)
GB	Temperature	CHESS-met (based on station data)	1 km	Robinson et al. (2017a)
GB	E_{pot}	CHESS-PE (Penman-Monteith based on CHESS-met data)	1 km	Robinson et al. (2016, 2017b)
US, BR, GB	Precipitation, Temperature, E_{pot}	ERA5-Land (Penman-Monteith based on reanalysis data)	9 km	Muñoz-Sabater et al. (2021)

3.3 Differences between ERA5-Land data in the Caravan dataset and forcing data in the Camels datasets

3.3.1 Differences in mean annual precipitation

170 The mean annual precipitation in the Caravan dataset differed between -53 % and 101 % from the one in the Camels dataset,
i.e., taking the Camels data as a reference, the mean annual precipitation was underestimated by up to 53 % and overestimated
by up to 101 % in the Caravan dataset. For 583 of the 1252 catchments (47 %), the deviation was within ±10 %, and for 968
catchments (77 %), it was within ±20 % (Fig. 1). The mean annual precipitation in the Caravan dataset was lower than in the

CAMELS-US dataset for the catchments in the eastern part of the US and on the West Coast. For some catchments in the

175   centre of the US, the mean annual precipitation in the Caravan dataset was much higher (>40 %) than in the CAMELS-US dataset. For the southern part of Brazil, the mean annual precipitation in the Caravan dataset was almost consistently higher (and sometimes much higher) than in the CAMELS-BR data, while for the northern part of Brazil, it tended to be lower than in the CAMELS-BR dataset. For the catchments in the eastern part of Great Britain, the mean annual precipitation was slightly higher in the Caravan dataset than in the CAMELS-GB dataset, while for the catchments in the western part of Great Britain,

180   the mean annual precipitation was lower in the Caravan dataset than in the CAMELS-GB dataset.

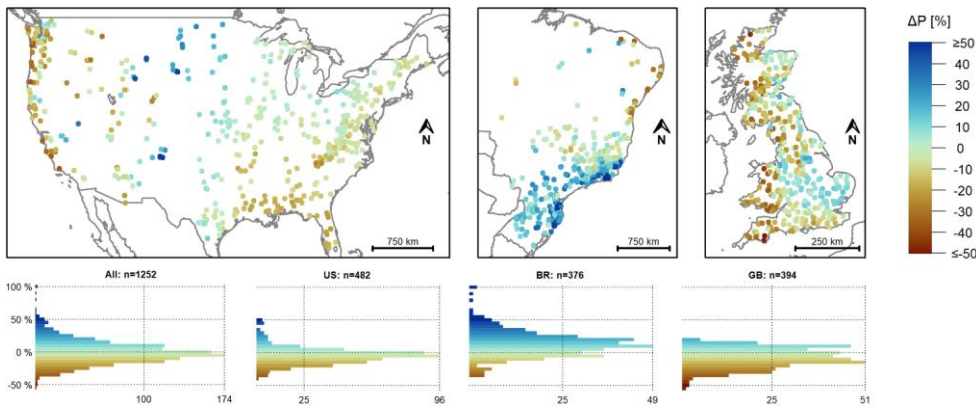



Figure 1: Relative difference in the mean annual precipitation (calculated for a 30-year period: 1983-2013) for each catchment in the Caravan dataset compared to the mean annual precipitation for each catchment in the Camels datasets. The brown colours
185   **indicate less precipitation in the Caravan dataset, the blue colours indicate more precipitation in the Caravan dataset than in the Camels datasets. Note that the colour scale was cut at ±50 % but the histograms cover the full range of differences (at 5 % intervals). For one catchment, the difference was less than -50 % and for twelve catchments, it was more than 50 %. The scale bars refer to the map centre and are different for each country. The base maps with the country outlines were obtained from Natural Earth (natu-ralearthdata.com).**

190

3.3.2   Differences in mean daily temperature

The mean daily temperature data in the Caravan and Camels datasets were relatively similar. In the most extreme cases, the mean daily temperature in the Caravan dataset was 4 °C less (i.e., colder) and 2.8 °C higher (i.e., warmer) than in the Camels datasets. For 961 of the 1252 catchments (77 %), the temperature difference was less than ±1 °C (Fig. 2). For the catchments

195   in the eastern part and the southern part of the West Coast of the US, the mean daily temperature in the Caravan dataset tended to be slightly higher than in the CAMELS-US dataset. For the catchments in the Pacific Northwest and most of the western US, the mean daily temperature in the Caravan dataset was lower than in the CAMELS-US dataset. In the snow-dominated

Rocky Mountain region, the mean daily temperature in the Caravan dataset was up to 2.8 °C lower than in the CAMELS-US dataset. For Brazil, the mean daily temperature in the Caravan dataset was almost always lower than in the CAMELS-BR dataset (i.e., it was higher for only eight catchments), and this difference was often substantial. For 246 Brazilian catchments (65 %), the mean temperature differed by at least -1 °C. For the catchments in Great Britain, the temperature data were similar, with differences between the two datasets varying between -0.9 °C and 0.5 °C.

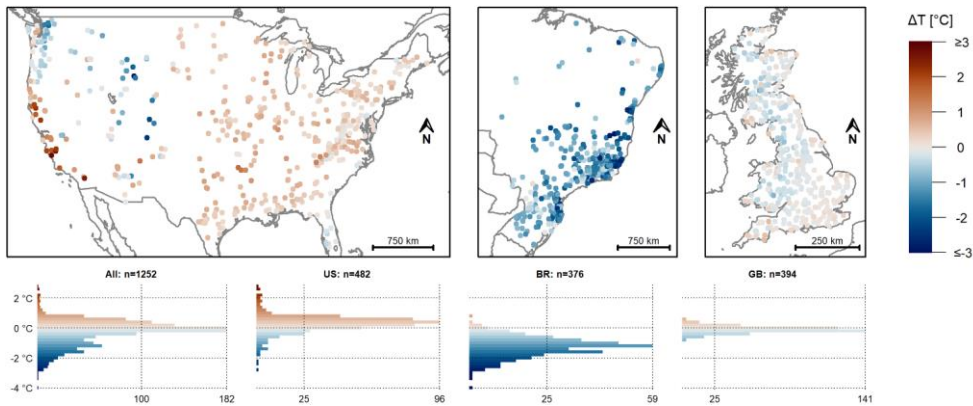



Figure 2: Difference in the mean daily temperature (calculated for a 30-year period: 1983-2013) for each catchment in the Caravan dataset and the Camels datasets. The blue colours indicate a lower mean daily temperature in the Caravan dataset, the red colours indicate a higher mean daily temperature in the Caravan dataset than in the Camels datasets. Note that the colour scale was cut at ±3 °C, but the histograms cover the full range of values (at 0.2 °C intervals). For three catchments, the difference was below -3 °C. For none of the catchments, the difference was higher than 3 °C.

3.3.3 Differences in mean annual potential evapotranspiration

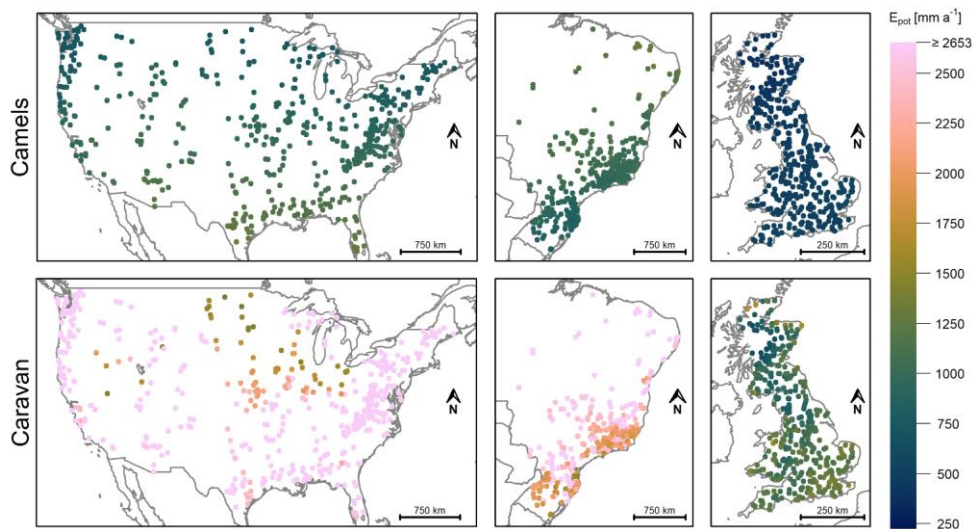
The E_{pot} data derived from ERA5-Land in the Caravan dataset are unrealistically high for most catchments in the US, Brazil, and Great Britain (Fig. 3), confirming the results of Klingler et al. (2021) for Central Europe and Xu et al. (2024) for China. The minimum mean annual E_{pot} in the Caravan dataset was higher than the maximum mean annual E_{pot} in the Camels datasets for each of the three regions, i.e., the ranges of the E_{pot} data did not overlap. The relative differences between the mean annual E_{pot} in the Caravan dataset and the mean annual E_{pot} in the Camels datasets varied between 46 % and 913 % (median: 462 %) for the US catchments, between 58 % and 523 % (median: 121 %) for the Brazilian catchments, and between 52 % and 337 % (median: 120 %) for the catchments in Great Britain.

220

Figure 3: Mean annual E$_{pot}$ (calculated for a 30-year period: 1983-2013) for the Camels datasets (Brazil, Great Britain) or calculated with the data from the Camels dataset (US) (top) and for the Caravan dataset (bottom). Note that the colour scale ends at twice the maximum E$_{pot}$ value reported in the Camels datasets. The number of catchments for which the E$_{pot}$ in the Caravan dataset was higher than this cutoff value (2653 mm a^{-1}, shown in light pink) was 385 for the US (80 % of the US catchments), 115 for Brazil (31
225 **%), and 0 for Great Britain.**

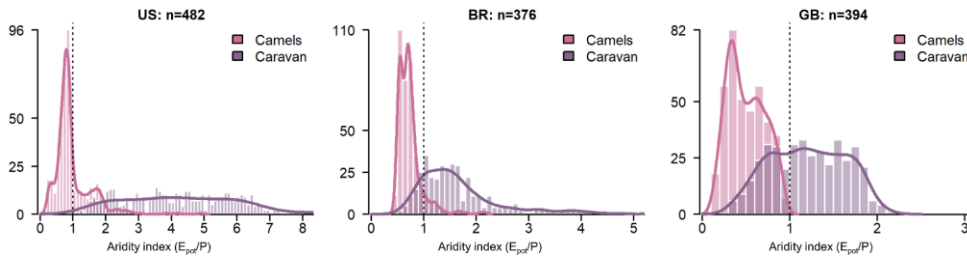
Even though the use of E$_{pot}$ from the ERA5-Land data is consistent with the other variables in the Caravan dataset, the high (and often unrealistic) E$_{pot}$ values are problematic. Kratzert et al. (2023a) mention the high E$_{pot}$ values in the Caravan paper in a table caption. However, hydrologists using the Caravan dataset under the assumption that the data are ready for use may end
230 up with wrong conclusions. The high E$_{pot}$ values do not only influence model simulation results (see Sect. 4.2) but also the catchment attributes based on these values. For the 30 years considered here, the mean annual E$_{pot}$ was larger than the mean annual precipitation (i.e., the aridity index was larger than 1.0) for 1059 of the 1252 catchments (85 %) based on the Caravan data whereas this was the case for only 167 catchments (13 %) based on the Camels data (Fig. 4).

235

Figure 4: Histograms of the aridity index values based on the mean annual evapotranspiration (E$_{pot}$) and precipitation (P) from the Camels and Caravan datasets (calculated for a 30-year period: 1983-2013). Note that 39 US catchments (8 %), and 4 Brazilian catchments (1 %) were not included in the histograms because the aridity index values for the Caravan data plot beyond the x-axis limits. The maximum calculated aridity index values were 20.2 for the US, 8.1 for Brazil, and 2.2 for Great Britain.

240

To provide a possible alternative, we calculated time series of E$_{pot}$ using the formula given by Adam et al. (2006) based on Droogers and Allen (2002). This formula is based on the Hargreaves formula (Hargreaves and Samani, 1982) and was used for one of the E$_{pot}$ products included in the CAMELS-AUS dataset (Fowler et al., 2021). The relatively low data requirement for this method allowed us to calculate E$_{pot}$ time series based on the ERA5-Land precipitation and temperature data only, i.e.,

245 not violating the philosophy of Caravan to use only globally available data. More specifically, it takes only the location and temperature into account, and additionally adjusts the E$_{pot}$ estimates based on the monthly precipitation as a proxy for humidity. As input data, we used the latitude of each catchment, as well as the time series of the day of the year, daily mean temperature, the difference between the mean daily maximum temperature and the mean daily minimum temperature for each month, and the monthly precipitation sums (see the data repository linked in the data availability statement for the calculations). We refer

250 to this E$_{pot}$ data as "Hargreaves E$_{pot}$".

The Hargreaves E$_{pot}$ data resulted in a mean annual E$_{pot}$ that was similar to the one of the Camels datasets (Table 1~~Table 1~~). For the US, the ratio between the mean annual Hargreaves E$_{pot}$ and the mean annual E$_{pot}$ in the Camels datasets varied between 0.6 and 1.4 (median: 0.9). This range was 0.6 to 1.3 for the catchments in Brazil (median: 1.0), and 0.5 to 1.1 for the catchments in Great Britain (median: 0.9). The catchments in the US and Great Britain for which the Hargreaves E$_{pot}$ values were (too)

255 low were mainly located at the higher latitudes. As a comparison, the ratio between the mean annual E$_{pot}$ in the Caravan dataset and the mean annual E$_{pot}$ in the Camels datasets varied between 1.5 and 10.1 (median: 5.6) for the US, between 1.6 and 6.2 (median: 2.2) for Brazil, and between 1.5 and 4.4 (median: 2.2) for Great Britain (Fig. 5).

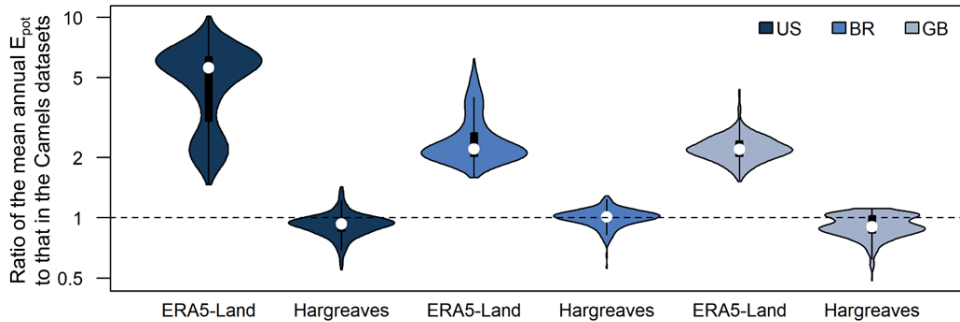



Figure 5: Violin plots showing the ratio between the mean annual E_{pot} of either the ERA5-Land data in the Caravan dataset or the Hargreaves E_{pot} data based on input data from the Caravan dataset and the mean annual E_{pot} in the Camels datasets for the catchments in the US, Brazil and Great Britain. The Camels E_{pot} refer to the E_{pot} data calculated with the Priestley-Taylor equation for the CAMELS-US dataset and the E_{pot} data included in CAMELS-BR and CAMELS-GB datasets (see Table 1). Note that the y-axis is logarithmic.

Of course, there are a variety of other ways to obtain daily E_{pot} values for the Caravan dataset and to provide an alternative to the current E_{pot} data in the Caravan dataset, e.g., the Hargreaves-Samani equation without the adjustment for humidity (Hargreaves and Samani, 1982) or the Thornthwaite equation (Thornthwaite, 1948) with a scaling for daily values. While it is an open question which method leads to the best results, the Hargreaves-based method used here provides a straightforward solution to avoid the problematic ERA5-Land-based Caravan E_{pot} data.

4   Effect of the differences in the forcing data on hydrological model results

4.1   Description of modelling experiments

To assess the overall effect of the differences in the forcing data for the Camels and the Caravan datasets on hydrological model performance, we conducted a series of modelling experiments. Even though a compensational effect of the model parameters can be expected, i.e., to adjust for possibly inaccurate or biased forcing data, we consider the model performances (i.e., how well the streamflow observations could be represented with a certain combination of forcing data) as an aggregated measure for data quality.

We calibrated the bucket-type HBV model (Bergström, 1992; Lindström et al., 1997) in the version HBV-light (Seibert and Vis, 2012) with a genetic algorithm (Seibert, 2000), optimizing the Kling-Gupta efficiency (KGE; Gupta et al., 2009) for the daily streamflow simulations. A detailed description of the model routines can be found elsewhere (e.g., Seibert and Vis, 2012). We created seven different combinations of forcing data, varying the data source for the precipitation, temperature, and E_{pot} time series (Table 2Table 2), and calibrated the model for each of these datasets. We did this for each of the 1252 catchments

for which we also compared the forcing data (see Sect. 3.3). These are all catchments from CAMELS-US, CAMELS-BR, and CAMELS-GB that were included in the Caravan dataset, except for 14 catchments from CAMELS-GB for which more than

285   20 % of the streamflow data were missing for the simulation period. We divided each catchment into elevation zones of 200 m, whereby each elevation zone had to make up at least 5 % of the catchment area (if not, the elevation zones were merged with the neighbouring elevation zone). This division is relevant for the snow routine of the HBV model. We used the EarthEnv-DEM90 digital elevation model (Robinson et al., 2014) and the shapefiles contained in the Caravan dataset to derive the elevation zones.

290   For the catchments in the US and Great Britain, we used 1 October 1988 to 30 September 2013 as the simulation period, and for the catchments in Brazil, we used 1 April 1988 to 31 March 2013 as the simulation period. The preceding five years were used as a warm-up period. Note that we did not distinguish between a calibration and validation period (i.e., we used the simulation period for calibration and evaluation) because we are interested in the influence of the different data types on model performance (cf. Tarek et al., 2020).

295   To account for equifinality, we calibrated the model for each scenario and catchment 100 times. From these 100 optimized parameter sets and their corresponding simulated hydrographs, we calculated the ensemble mean hydrograph based on the arithmetic average of the 100 simulated streamflow values for each day. We compared this simulated hydrograph to the observed hydrograph to obtain one KGE value per data scenario for each catchment.

300   **Table 2: Overview of the seven combinations of calibration data used for the different scenarios of the modelling experiment. In addition to the forcing data from the Caravan and the Camels datasets (see Sect. 3.2 and ~~Table 1~~Table 1 for details), we also used the Hargreaves-based E_{pot} values based on Caravan data as an alternative to the unrealistically high E_{pot} data in the Caravan dataset (see Sect. 3.3.3).**

Scenario	Scenario description	Precipitation	Temperature	E_{pot}
I	Camels	Camels	Camels	Camels
II	Caravan	Caravan	Caravan	Caravan
III	Camels, but with Caravan precipitation data	Caravan	Camels	Camels
IV	Camels, but with Caravan temperature data	Camels	Caravan	Camels
V	Camels, but with Caravan E_{pot} data	Camels	Camels	Caravan
VI	Camels, but with Hargreaves E_{pot} data	Camels	Camels	Hargreaves
VII	Caravan, but with Hargreaves E_{pot} data	Caravan	Caravan	Hargreaves

305   ## 4.2   Results

4.2.1   Model performances with Camels and Caravan data

Using the Camels forcing data for model calibration (scenario I) led to good model performances for most catchments (Fig. 6, Fig. 7). For the US catchments, the KGE ranged from 0.12 to 0.96 (median: 0.85) and for 20 of the 482 catchments (4 %), it was below 0.6. For the Brazilian catchments, the KGE ranged from -0.85 to 0.94 (median: 0.77); it was negative for two

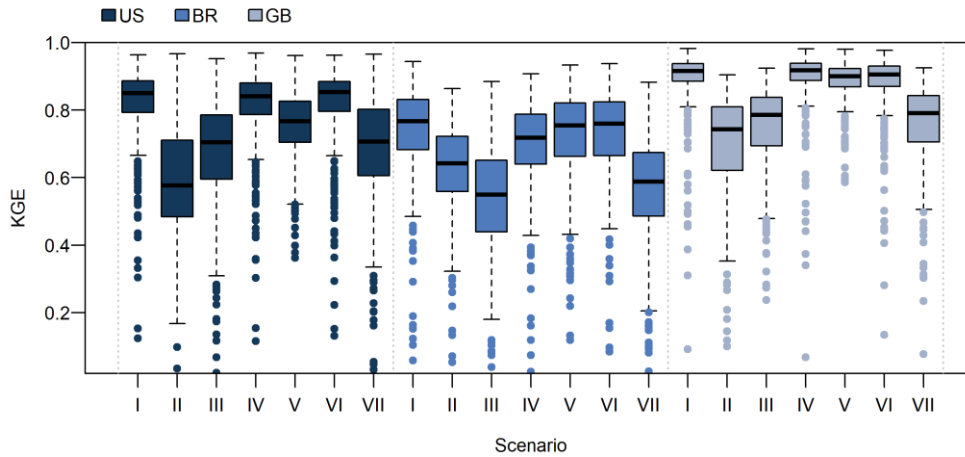
310 catchments and below 0.6 for 52 of the 376 catchments (14 %). For the catchments in Great Britain, the KGE ranged from -2.27 to 0.98 (median: 0.92); it was negative for three catchments, and below 0.6 for 13 of the 394 catchments (3 %). For the five catchments with a negative KGE, the simulated streamflow was higher than the observed streamflow but the observed streamflow was less than expected based on the precipitation and E_{pot} data.

315

Figure 6: Boxplots illustrating the model performances (KGE values for the ensemble mean hydrograph) for all scenarios (see ~~Table 2~~Table 2 for a description) for all catchments in the US (n=482), Brazil (BR; n=376), and Great Britain (GB; n=394). The lower limit of each box represents the 25th percentile, the upper limit the 75th percentile, and the line the median. The whiskers end at the most extreme data point within 1.5 times the interquartile range. The dots represent outliers. Note that the y-axis was limited to positive
320 **KGE values. The KGE values were negative for 46 cases.**

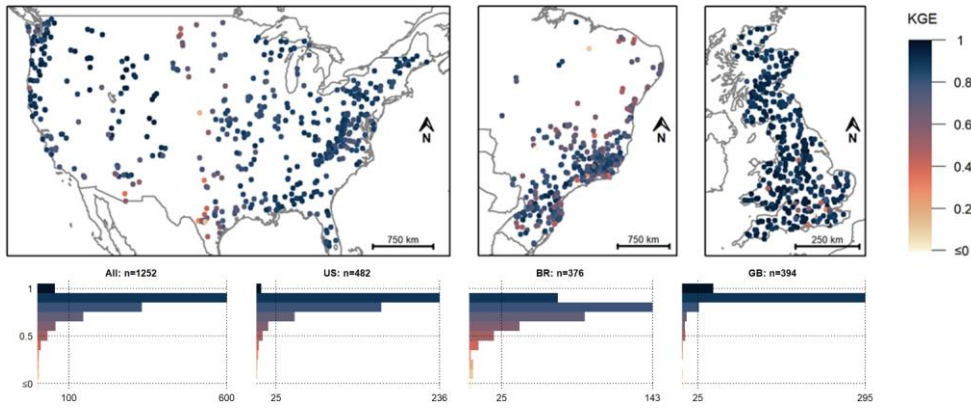



Figure 7: Model performance (KGE values) for scenario I (Camels forcing data) for the catchments in the US, Brazil and Great Britain for the period April 1988 to March 2013 (Brazil) or October 1988 to September 2013 (US, Great Britain). Note that the lower limit of the scale was cut at 0. The KGE was negative for five catchments. The KGE values were rounded to one decimal for the histograms.

Compared to calibration with the Camels data, calibration with the Caravan data (scenario II) decreased the KGE for 1134 of 1252 catchments (91 %; Fig. 6, Fig. 8, Table 3~~Table 3~~). The KGE for the calibration with Caravan data was below 0.6 for 488 of the 1252 catchments (39 %, i.e., 403 catchments more than for scenario I (Camels data)). However, the Caravan forcing data led to a positive KGE for all catchments, i.e., for the five catchments in Brazil and Great Britain for which the KGE for the calibration with the Camels data was negative, calibration with Caravan data resulted in positive KGE values. For these five catchments, the simulated streamflow was overestimated with the Camels forcing data and lower for the Caravan forcing data, and thus more similar to the observed streamflow.

Table 3: Effect of differences in all forcing data (i.e., comparison of scenarios II and I), precipitation data (scenarios III and I), temperature data (scenarios IV and I), E_{pot} data (scenarios V and I) from the Camels and Caravan datasets on model performance (i.e., KGE values), as well as effect of using Hargreaves E_{pot} data instead of the E_{pot} data from the Camels datasets (scenarios VI and I), or the Caravan dataset (scenarios VII and II) on model performance for all catchments together and each region. The stars indicate the statistical significance of the one-sided Wilcoxon test: ** indicates a p-value <0.001, * indicates a p-value <0.01. For all tests except the effect of using the Hargreaves-based E_{pot} data instead of the E_{pot} data from the Caravan dataset, we tested for a significant decrease in model performance; for the latter, we tested for a significant increase in model performance (last column).

Effect of differences in: Comparison of scenarios:		All data II-I	Precip. III-I	Temp. IV-I	E_{pot} V-I	E_{pot} VI-I	E_{pot} VII-II
Median ΔKGE and significance	All (n=1252)	-0.17**	-0.14**	-0.00**	-0.02**	-0.00**	0.04**
	US (n=482)	-0.25**	-0.14**	-0.00**	-0.05**	0.00	0.10**
	BR (n=376)	-0.11**	-0.19**	-0.03**	-0.00*	-0.01**	-0.02
	GB (n=394)	-0.17**	-0.11**	0.00	-0.02**	-0.01**	0.08**
Number and percentage of catchments with ΔKGE > 0.1	All (n=1252)	39 (3 %)	23 (2 %)	3 (0 %)	61 (5 %)	9 (1 %)	423 (34 %)
	US (n=482)	7 (1 %)	3 (1 %)	2 (0 %)	9 (2 %)	0 (0 %)	245 (51 %)
	BR (n=376)	17 (5 %)	10 (3 %)	1 (0 %)	28 (7 %)	2 (1 %)	17 (5 %)
	GB (n=394)	15 (4 %)	10 (3 %)	0 (0 %)	24 (6 %)	7 (2 %)	161 (41 %)
Number and percentage of catchments with ΔKGE < 0	All (n=1252)	1134 (91 %)	1169 (93 %)	757 (60 %)	855 (68 %)	786 (63 %)	433 (35 %)
	US (n=482)	434 (90 %)	443 (92 %)	293 (61 %)	385 (80 %)	205 (43 %)	84 (17 %)
	BR (n=376)	333 (89 %)	349 (93 %)	339 (90 %)	200 (53 %)	275 (73 %)	255 (68 %)
	GB (n=394)	367 (93 %)	377 (96 %)	125 (32 %)	270 (69 %)	306 (78 %)	94 (24 %)
Number and percentage of catchments with ΔKGE < -0.1	All (n=1252)	873 (70 %)	770 (62 %)	34 (3 %)	236 (19 %)	26 (2 %)	172 (14 %)
	US (n=482)	375 (78 %)	287 (60 %)	0 (0 %)	160 (33 %)	4 (1 %)	21 (4 %)
	BR (n=376)	208 (55 %)	265 (70 %)	34 (9 %)	50 (13 %)	16 (4 %)	115 (31 %)
	GB (n=394)	290 (74 %)	218 (55 %)	0 (0 %)	26 (7 %)	6 (2 %)	36 (9 %)

For the catchments in the US, the KGE mainly decreased for the catchments east of the 100° W meridian and along the West Coast. For the remainder of the western part of the US, the KGE did not change considerably. For the catchments in Brazil, the KGE tended to decrease most for the more southern catchments, but there were also some catchments in the eastern part of Brazil for which the KGE decreased quite strongly. The KGE increased for a few Brazilian catchments. For the catchments along the western coast of Great Britain, the KGE decreased strongly. The decrease was less strong for catchments in the southern part. For some catchments in ~~Southern~~ southern England, the KGE increased (Fig. 8). This included a cluster of catchments for which the KGE was comparably low when calibrated with the Camels data (scenario I, Fig. 7).

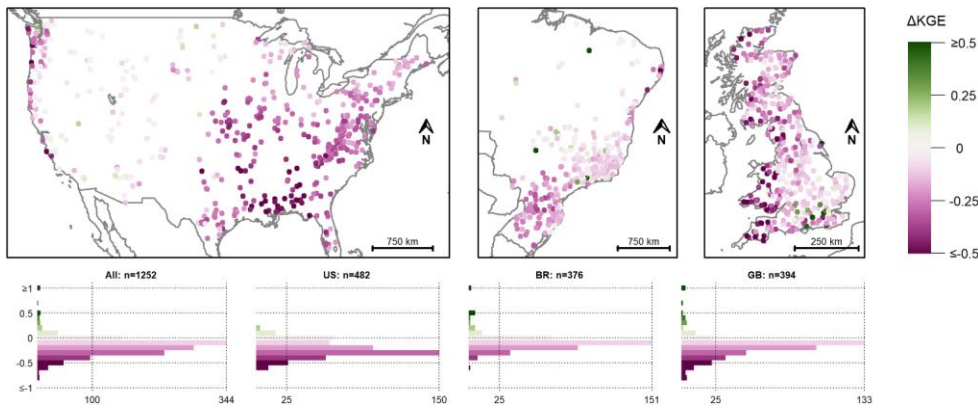



Figure 8: Difference in model performance when using the Caravan forcing data (scenario II) and the Camels forcing data (scenario I). The pink colours indicate a lower KGE when calibrating with the Caravan data, the green colours indicate a higher KGE when calibrating with the Caravan data. Note that the colour scale was cut at a difference in KGE of ±0.5 and that the y-axes of the histograms were cut at a difference in KGE of ±1. The ΔKGE values were rounded to one digit for the histograms.

4.2.2   Effect of differences in precipitation data

Model performances were higher when Camels precipitation data were used for model calibration than when Caravan precipitation data were used (Fig. 6). Using the Caravan precipitation data (scenario III) instead of the Camels precipitation data (scenario I) decreased the KGE for 1169 of the 1252 catchments (93 %) (Table 3~~Table 3~~). The pattern of the effect of the Caravan precipitation data on the KGE values was similar to the pattern of the effect of all the Caravan forcing data (Fig. 8). Indeed, the median difference between the KGE achieved with scenario II and scenario III for all 1252 catchments was -0.03, i.e., scenario III performed only slightly better than scenario II. The median difference was -0.09 for the US catchments, 0.06 for the Brazilian catchments (where scenario II performed better than scenario III, see Fig. 6 and Table 3~~Table 3~~), and -0.07 for the catchments in Great Britain. In other words, the difference in the precipitation data explained most of the effect of replacing the forcing data from the Camels datasets with the forcing data from the Caravan dataset. Furthermore, the effect of the difference in the precipitation data was larger than the effect of the difference in the temperature data and also the effect of the large difference in the E_{pot} data (see Sect. 4.2.3 and 4.2.4).

4.2.3 Effect of differences in temperature data

370 The effect of using temperature data from the Caravan dataset (scenario IV) instead of temperature data from the Camels datasets (scenario I) was comparably small (Fig. 6). However, when considering all 1252 catchments, as well as when considering only the US catchments or only the Brazilian catchments, the KGE values still decreased significantly in scenario IV compared to scenario I (Table 3Table 3; p<0.001). Only in Great Britain, where the mean daily temperature data in the Caravan dataset were very similar to the mean daily temperature data in the CAMELS-GB dataset for most catchments (Fig. 2), no

375 significant decrease in the KGE values was found when scenario IV was compared to scenario I (p=1.0), i.e., replacing the temperature data from the CAMELS-GB dataset with the temperature data from the Caravan dataset did not have a significant effect. There was no indication that the replacement of the temperature data had a stronger influence on the KGE for snow-dominated (mountainous) catchments than other catchments, as it may have been expected.

4.2.4 Effect of differences in E_pot potential evapotranspiration data

380 Using the E_{pot} data from the Caravan dataset (scenario V) instead of the E_{pot} data from the Camels datasets (scenario I), significantly decreased the KGE (Table 3Table 3; p<0.01 for the Brazilian catchments and p<0.001 for the catchments in the US and Great Britain, or when taking all catchments together). The decrease was particularly pronounced for the catchments in the US, where the differences between the mean annual E_{pot} from the Caravan dataset and the mean annual E_{pot} from the CAMELS-US dataset were especially large (Fig. 3, Fig. 5). However, compared to the KGE decrease when all forcing data

385 were taken from the Caravan dataset (scenario II) or when only precipitation data were taken from the Caravan dataset (scenario III), the effect of the unrealistic E_{pot} data from the Caravan dataset was relatively small (Fig. 6).

The model performance drop compared to scenario I tended to be smaller when the model was calibrated with the Hargreaves E_{pot} data (scenario VI), than when the model was calibrated with the E_{pot} data from the Caravan dataset (scenario V). For the US catchments, there was no significant decrease in KGE when the E_{pot} data calculated with the Priestley-Taylor equation for

390 the CAMELS-US dataset were replaced with the Hargreaves E_{pot} data (compare scenario VI to scenario I; p=0.987; Table 3Table 3), while this was the case when replacing the E_{pot} data from the Camels datasets with the Hargreaves E_{pot} data for Brazil and Great Britain (p<0.001).

Similarly, we tested whether replacing the ERA5-Land E_{pot} data in the Caravan dataset with the Hargreaves E_{pot} data (scenario VII) significantly improved the model performance compared to when all forcing data from the Caravan dataset were used

395 (scenario II). This was indeed the case (p<0.001) when either all catchments, all US catchments, or all catchments in Great Britain were considered (Table 3Table 3, last column). However, for the Brazilian catchments, the effect was the opposite, i.e., the unrealistic E_{pot} data from the Caravan dataset led to significantly better results than the alternative Hargreaves E_{pot} data (p<0.001).

A positive effect of the Hargreaves E_{pot} data instead of the E_{pot} data from the Caravan dataset on the model performance could be observed especially for regions in which the use of Caravan forcing data (scenario II) instead of Camels forcing data (scenario I) had a strong negative impact (Fig. 8, Fig. 9). In the US, this was mainly the case for the catchments in the eastern part of the country and along the West Coast. The few catchments in Brazil for which the model performance increased due to the Hargreaves E_{pot} data were located in the southern part of the country, as well as along the eastern coast. In Great Britain, the increases tended to be stronger in the western part of the country (Fig. 9).

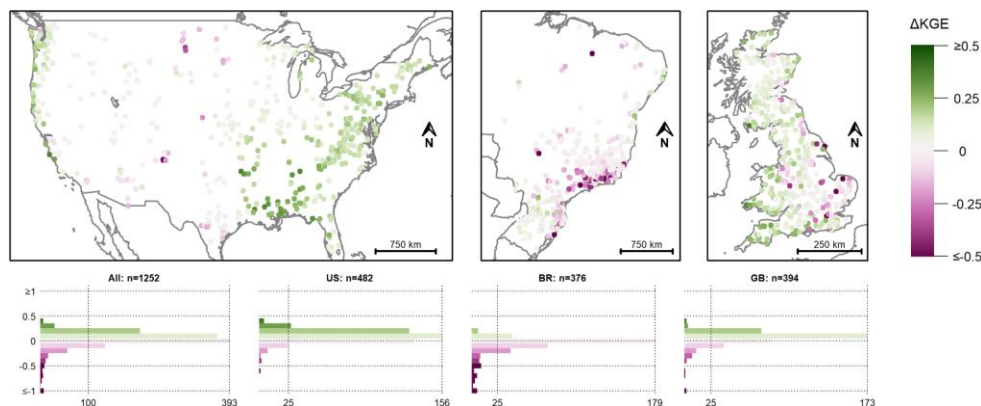



Figure 9: Difference in KGE values for the model calibration with precipitation and temperature data from the Caravan dataset and the Hargreaves-based E_{pot} data (scenario VII) and the calibration when all Caravan forcing data were used (scenario II). The pink colours indicate a lower KGE value when the Hargreaves-based E_{pot} data were used compared to the calibration with all Caravan data, the green colours indicate a higher KGE value with the Hargreaves-based E_{pot} data. Note that the colour scale was cut at a difference in KGE of ±0.5 and that the y-axes of the histograms were cut at a difference in KGE of ±1. The ΔKGE values were rounded to one decimal for the histograms.

4.3   Discussion of the difference in model performances for Camels and Caravan data

Streamflow modelling with the forcing data included in the three Camels datasets worked well for most catchments. An unsuitable model structure, errors in the Camels data, or human impacts on streamflow are possible explanations for the poor model performances for some of the catchments. For example, the catchments in the arid regions of the US for which the model performances were low were identified as more difficult to model in earlier studies as well (Knoben et al., 2020; Kollat et al., 2012). Based on the comparison of different models, Knoben et al. (2020) found that there are model structures that can simulate the discharge in these catchments successfully. Similarly, the low model performances for some catchments in south-eastern Great Britain may be attributed to complex groundwater systems (as identified earlier by Lane et al., 2019;

Seibert et al., 2018). A more suitable model structure accounting for subsurface losses would lead to a better model performance (Kiraz et al., 2023). However, looking at the model results as an aggregated measure of data quality, the good model fits indicate a high data quality of the data in the CAMELS-US, CAMELS-BR, and CAMELS-GB datasets.

425 The overall deterioration in model performances for calibration with the Caravan dataset indicates that the quality of the forcing data from ERA5-Land is less lower than the quality of the data that are available for the US, Brazil, and Great Britain. As the ERA5-Land data are coarser than most data in the Camels datasets (Table 1Table 1), this was to a certain extent expected. Furthermore, the negative effect of the Caravan forcing data on model performance may be smaller for models that are less sensitive to errors in the input data and can adapt more flexibly. However, a user who decides to use the Caravan dataset

430 instead of different Camels datasets out of convenience may not be aware of the considerable degradation of the input data and the potentially severe effects on the model performanceof the severity of these differences.

Even though the E_{pot} data in the Caravan dataset are unrealistically high for many catchments (Fig. 3), the analysis of the isolated effects of the Caravan forcing data showed that differences in the precipitation (Fig. 1) were responsible for most of the decrease in model performance for the Caravan forcing data (Fig. 6, Table 3Table 3). As precipitation is the main driver

435 of streamflow, the strong influence of precipitation is not surprising. That the model performance dropped so much indicates more than a small bias and rather a lower plausibility of the reanalysis-based precipitation data (cf. Beck et al., 2017; Tarek et al., 2020; Wang et al., 2023a).

The spatial differences in how the model performance was affected by the Caravan precipitation data may be related to both the spatial patterns in the errors and the catchment characteristics. For example, the Caravan precipitation data led to a much

440 stronger deterioration in model performance for catchments in the eastern part of the US than in the western part. This pattern was also observed in earlier studies that tested the value of reanalysis data for hydrological modelling in North America (Essou et al., 2016; Tarek et al., 2020). Essou et al. (2016) mainly attributed this issue to the convective summer storms in the eastern part of the US that are poorly represented in the reanalysis data.

These results mean that one should be cautious when using the Caravan precipitation dataset instead of more reliable (e.g.,

445 station-based) precipitation data because the conclusions may be affected by the lower data quality of the forcing data in the Caravan dataset. In our opinion, ERA5-Land precipitation data should only be used for catchments for which there are no alternative data (so that these catchments can still be included in large-sample studies). This is in line with the conclusions of Essou et al. (2016, 2017) and Tarek et al. (2020), who stated that reanalysis data can serve as a proxy for meteorological data for regions with little or no weather station data.

450 Considering the large bias of the Caravan E_{pot} data (Fig. 3, Fig. 5), the effect on the model performance was surprisingly small and clearly smaller than the effect of the precipitation data (Fig. 6, Table 3Table 3: III-I versus V-I). This is in line with earlier studies that showed that E_{pot} data affect model performance less than precipitation data (Oudin et al., 2006; Paturel et al., 1995) because the model can compensate for a systematic overestimation of E_{pot}. Thus, an overestimation of E_{pot} is less severe than an underestimation (cf. Jayathilake and Smith, 2022). Indeed, additional sensitivity analyses with artificially biased E_{pot} data,

455 not shown here for the sake of brevity, showed that tThe HBV model compensated for the overestimated E_{pot} data from the

Caravan dataset mainly by adjusting the values of the parameters of the soil routine to reduce evapotranspiration. This allowed the model to simulate an actual evapotranspiration that was more realistic and of a similar order of magnitude as the actual evapotranspiration simulated with the E_{pot} data from the Camels datasets. Thus, even though the model performance may have not changed considerably, the processes were represented differently due to the compensation. This is problematic, especially when the calibrated parameter values are subsequently used to characterize a catchment (cf. Bouaziz et al., 2022). Nevertheless, the overestimation of E_{pot} in the Caravan data still reduced the model performance for many catchments. Even if the model performance did not decrease too strongly, it is likely that the catchment processes were represented less accurately as the too-high E_{pot} data needed to be compensated for.

The few cases for which the model performance was better with the E_{pot} data from the Caravan dataset can either be attributed to even more (but exceptional) erroneous Camels data or to compensation effects of biased variables (cf. Wang et al., 2023b). Possible explanations for the catchments for which the unrealistically high E_{pot} data led to an increase in model performance may be a wrong representation of the processes that coincidentally led to a better model performance (Kirchner, 2006) or errors in the water balance data for the Camels dataset and thus an improvement thanks to the high (but still wrong) E_{pot} data. A compensation of the wrong water balance with overestimated E_{pot} data may also explain why many catchments in Brazil did not profit from the more realistic Hargreaves-based E_{pot} data (cf. Fig. 9).

The low sensitivity of the hydrological model to the wrong E_{pot} data indicates that validating meteorological forcing data with a hydrological model approach, as we did in this study, may not be the most suitable way to investigate the quality of E_{pot} data but works fine for precipitation data. Thus, other approaches or simple plausibility tests may be more useful for the validation of E_{pot} data and the indices calculated thereof.

5   Suggestions for use of the Caravan dataset

For the vast majority of the catchments, using the forcing data from the Caravan dataset deteriorates model results and impacts the conclusions drawn from them. In our opinion, the model performances were affected so strongly by the use of the reanalysis data in the Caravan dataset that it cannot be considered an inconsequential trade-off between the use of homogeneous data and a drop in model performance. Even though we agree that the use of ERA5-Land data for all catchments has advantages, such as comparability and the possibility to extend the Caravan dataset to other catchments, the loss in data quality for this stand-ardization is a hefty price tag.

Because the Caravan dataset is easy to acquire, well-organized, and offers opportunities for catchments in underrepresented regions to be included in large-sample studies in hydrology, there are clear advantages of using reanalysis data for some studies, and in particular for catchments for which the forcing data would otherwise not be available. The use of the Caravan dataset as the standard resource for large-sample hydrology would also facilitate the comparison of model results. However, the quality of the meteorological data that are used for hydrological model calibration is lower for the Caravan dataset than for the original Camels datasets. Thus, in our opinion, the Caravan forcing time series are not the most suitable dataset for all studies, in

particular for catchments for which higher quality data are available. Therefore, we provide two suggestions to improve the Caravan dataset.

490 **5.1   Extension with forcing data from the original datasets**

To make researchers aware that they are using lower quality data when downloading the data from the Caravan dataset (than when they would use the Camels datasets), we suggest extending the Caravan dataset by also adding the forcing data that were originally included in the national and regional large-sample datasets when these are available. In this way, users would be able to decide if either global comparability or the use of the best possible data is more important for their study. Including

495 both data types in Caravan would also lead to more transparency regarding the differences between the forcing data in the Camels datasets and the reanalysis data in the Caravan dataset. For catchments for which no other data are available than those from ERA5-Land, i.e., for which the ERA5-Land data are state of the art, no extension would be necessary. Of course, users already have this choice, since the Camels datasets are always available in their own repositories. Still, it would be much more convenient for the users finding them in Caravan, for facilitating their use and the comparison with the ERA5-Land data.

500 ~~Alternatively, a clear warning on the loss of data quality due to the standardization of the meteorological forcing data in Caravan is needed to avoid that the Caravan dataset is used instead of the original Camels datasets without the user being aware of the consequences.~~

Until the Caravan dataset is extended in such a way, we highly recommend that users assess thoroughly if they want to use the Caravan dataset or if they prefer the data that were originally included in the Camels datasets. Especially if a study is limited

505 to catchments for which better data are available, it may be valuable to go the more tedious way and download the different Camels datasets separately. Even though the Camels data are also not perfect, their quality is better than the one of the standardized data currently available in the Caravan dataset.

There are, of course, also situations in which the global comparability (and thus the reliance on input data that was generated uniformly for all catchments) is most important. In such cases using the Caravan forcing data is the best possible solution (at

510 least currently), and we suggest using the Caravan data as it is (for all variables, except E_{pot}, see Sect. 5.2), even though this may mean a loss in data quality and model performance (which is larger for some catchments than for others). However, in most applications, we think that it is better to use the best possible data, as one would do in every other situation in life.

~~Alternatively~~As an alternative or addition to the extension of the Caravan dataset with the original Camels data, a clear warning on the loss of data quality due to the standardization of the meteorological forcing data in the Caravan dataset is needed to

515 avoid that the Caravan dataset is used instead of the original Camels datasets without the user being aware of the consequences. Such a warning would avoid duplicating already existing data and still enable the user to make an informed decision.

It can be considered a general lesson learned from this study that new large-sample datasets need to clearly state their advantages compared to already existing datasets, but also inform users about possible drawbacks. With ERA6, the next generation of reanalysis data is currently being developed. Considering the development of reanalysis data so far, it is expected that

520 the quality will increase. This could change the appropriateness of reanalysis data as forcing data in hydrological models.

However, if limitations of a new dataset are already known on beforehand, a disclaimer section in the accompanying publication should be added and the users should be informed about the limitations in the database itself. Furthermore, if issues with some of the data only become clear at a later point in time, this information should be added to the database. With that, it can be promoted that the right datasets are used for the right purposes.

5.2 Replacement of the ERA5-Land derived potential evapotranspiration data

The comparison of the E_{pot} data included in the Caravan dataset with the E_{pot} data from the CAMELS-US, CAMELS-BR, and CAMELS-GB datasets showed that the Caravan E_{pot} data are systematically too high and are not reliable for any hydrological application. Because hydrological models can cope with some errors in the E_{pot} input data (Andréassian et al., 2004; Bai et al., 2016; Oudin et al., 2006), we expect that this large difference is mainly problematic for the attributes based on these E_{pot} data, such as the aridity index (see Fig. 4). Therefore, we suggest replacing the E_{pot} data from ERA5-Land with an alternative method and recalculating the values of the catchment attributes that include the E_{pot} data. The Hargreaves-based approach (see Sect. 3.3.3) is a possible alternative for the E_{pot} data that could be included in Caravan. The advantages are that they are realistic and can be calculated based on the other ERA5-Land derived data (temperature and precipitation) that are already in the Caravan dataset. However, there are other methods to estimate E_{pot} as well and different global datasets containing E_{pot} estimates, such as the dataset presented by Singer et al. (2021) resulting from the application of the FAO's Penman-Monteith equation based on ERA5-Land meteorological variables. With Caravan being a community effort, making a suitable choice for new Caravan E_{pot} data can be considered a task of the large-sample hydrology community. Aside from replacing the current E_{pot} data with other globally available E_{pot} data, our suggestion of including the forcing data from the original Camels datasets where possible as an alternative to the standardized global data (see Sect. 5.1) also applies for E_{pot} data.

6 Conclusions

Currently, the Caravan dataset is the most comprehensive large-sample dataset available in hydrology. It provides the community with hydrometeorological information and catchment attributes for many catchments in the world and offers the opportunity to extend the dataset with catchments for which streamflow data (but potentially no meteorological data) are available. It, furthermore, allows the forcing data to be comparably easily updated. Therefore, the Caravan dataset brings large-sample hydrology to the next level. However, there are considerable differences between the forcing data included in the Caravan dataset and the forcing data in the original large-sample datasets, as shown here for the CAMELS-US, CAMELS-BR, and CAMELS-GB datasets. The goal of this opinion paper is to make researchers aware of these differences and to show that these differences cause a reduction in model performance for most catchments. The impact of the lower quality data on model results may lead to wrong conclusions, for example regarding the suitability of a model or its parametrization. It can also affect conclusions regarding the suitability of regionalization approaches and the value of data for calibration of otherwise ungauged catchments. Therefore, we suggest that the standardized global forcing data in the Caravan dataset is are extended with the

higher quality forcing data from the original data sources where available. We also suggest using other E_{pot} data, e.g., calculated from the temperature data included in the Caravan dataset, as the ERA5-Land E_{pot} data are unrealistically high for many catchments. Even though this does not affect the model calibration results as much as the differences in the precipitation data, it can lead to wrong parameterizations and affects the catchment attributes (and thus catchment comparisons). We are sure that these relatively easy changes will increase the value of the Caravan dataset further and support its establishment as the main resource for large-sample hydrology.

7   Code and data availability

The Caravan dataset (Kratzert et al., 2023b) is available from Zenodo, with doi: 10.5281/zenodo.7944025. The CAMELS-US dataset (Addor et al., 2017a; Newman et al., 2014) is available from https://ral.ucar.edu/solutions/products/camels (last accessed: March 6, 2024). The CAMELS-BR dataset (Chagas et al., 2020b) is available from Zenodo, doi: 10.5281/zenodo.3964745. The CAMELS-GB dataset (Coxon et al., 2020b) is available from the NERC Environmental Information Data Centre, doi: 10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9.

The HBV model in the version HBV-light is available from https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model/HBV-Download.html (last accessed: March 6, 2024).

An R script for the calculation of the Hargreaves-based E_{pot} values is available from Zenodo, doi: 10.5281/zenodo.10784701. All colourmaps used in this paper are scientific colourmaps from Crameri (2023), accessed via the R package "scico", version 1.5.0 (Pedersen and Crameri, 2023). Other R packages used for this study are "circlize", version 0.4.15 (Gu et al., 2014); "hydroGOF", version 0.4-0 (Zambrano-Bigiarini, 2023); "vioplot", version 0.4.0 (Adler et al., 2022); "rworldmap", version 1.3-6 (South, 2011); "rworldxtra", version 1.01 (South, 2012); and "maps", version 3.4.1 (Becker et al., 2022).

8   Author contribution

FCS: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing – original draft preparation; Writing – review and editing

GS: Conceptualization; Data curation; Formal analysis; Investigation; Validation; Writing – review and editing

MN: Supervision; Writing – review and editing

ET: Supervision; Writing – review and editing

IvM: Conceptualization; Methodology; Resources; Supervision; Writing – review and editing

JS: Conceptualization; Methodology; Resources; Supervision; Writing – review and editing

9 Competing interests

ET and JS are members of the editorial board of Hydrology and Earth System Sciences.

10 Acknowledgements

We thank Marc Vis for providing the scripts to create the catchment files for the HBV model, for his advice on parallel computing, and previous discussions on hydrological modelling. We thank the Science IT team of the University of Zurich (www.s3it.uzh.ch) for providing the infrastructure for cloud computing. We thank Nans Addor and Frederik Kratzert for their helpful comments on an earlier version of this manuscript. We also thank the editor and the two reviewers for their constructive comments.

11 References

Adam, J. C., Clark, E. A., Lettenmaier, D. P., and Wood, E. F.: Correction of global precipitation products for orographic effects, J. Clim., 19, 15–38, https://doi.org/10.1175/JCLI3604.1, 2006.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: Catchment attributes for large-sample studies, Boulder, CO: UCAR/NCAR [dataset], https://doi.org/10.5065/D6G73C3Q, 2017a.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: Catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017b.

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., and Clark, M. P.: A ranking of hydrological signatures based on their predictability in space, Water Resour. Res., 54, 8792–8812, https://doi.org/10.1029/2018WR022606, 2018.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrol. Sci. J., 65, 712–725, https://doi.org/10.1080/02626667.2019.1683182, 2020.

Adler, D., Kelly, S. T., Elliott, T., and Adamson, J.: vioplot: violin plot, R package version 0.4.0 [software], https://github.com/TomKellyGenetics/vioplot, 2022.

Almagro, A., Oliveira, P. T. S., Meira Neto, A. A., Roy, T., and Troch, P.: CABra: A novel large-sample dataset for Brazilian catchments, Hydrol. Earth Syst. Sci., 25, 3105–3135, https://doi.org/10.5194/hess-25-3105-2021, 2021.

Almeida, M. and Coelho, P.: A first assessment of ERA5 and ERA5-Land reanalysis air temperature in Portugal, Int. J. Climatol., 43, 6643–6663, https://doi.org/10.1002/joc.8225, 2023.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies - Chile dataset, Hydrol. Earth Syst. Sci., 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Andréassian, V., Perrin, C., and Michel, C.: Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models, J. Hydrol., 286, 19–35, https://doi.org/10.1016/j.jhydrol.2003.09.030, 2004.

Araújo, C. S. P. D., Silva, I. A. C. E., Ippolito, M., and Almeida, C. D. G. C. D.: Evaluation of air temperature estimated by ERA5-Land reanalysis using surface data in Pernambuco, Brazil, Environ. Monit. Assess., 194, 381, https://doi.org/10.1007/s10661-022-10047-2, 2022.

Arsenault, R., Brissette, F., Martel, J. L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., and Poulin, A.: A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds, Sci. Data, 7, 243, https://doi.org/10.1038/s41597-020-00583-2, 2020.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Mendoza, P. A., McNamara, I., Beck, H. E., Thurner, J., Nauditt, A., Ribbe, L., and Thinh, N. X.: On the selection of precipitation products for the regionalisation of hydrological model parameters, Hydrol. Earth Syst. Sci., 25, 5805–5837, https://doi.org/10.5194/hess-25-5805-2021, 2021.

Bai, P., Liu, X., Yang, T., Li, F., Liang, K., Hu, S., and Liu, C.: Assessment of the influences of different potential evapotranspiration inputs on the performance of monthly hydrological models under different climatic conditions, J. Hydrometeorol., 17, 2259–2274, https://doi.org/10.1175/JHM-D-15-0202.1, 2016.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrol. Earth Syst. Sci., 21, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017, 2017.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., McVicar, T. R., and Adler, R. F.: MSWEP v2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment, Bull. Am. Meteorol. Soc., 100, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1, 2019.

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A.: maps: draw geographical maps, R package version 3.4.1 [software], https://CRAN.R-project.org/package=maps, 2022.

Bergström, S.: The HBV Model — its structure and applications, SMHI Reports RH, Norrköping, Sweden, 1992.

van Beusekom, A. E., Hay, L. E., Bennett, A. R., Choi, Y. D., Clark, M. P., Goodall, J. L., Li, Z., Maghami, I., Nijssen, B., and Wood, A. W.: Hydrologic model sensitivity to temporal aggregation of meteorological forcing data: A case study for the contiguous United States, J. Hydrometeorol., 23, 167–183, https://doi.org/10.1175/JHM-D-21-0111.1, 2022.

Bouaziz, L. J. E., Aalbers, E. E., Weerts, A. H., Hegnauer, M., Buiteveld, H., Lammersen, R., Stam, J., Sprokkereef, E., Savenije, H. H. G., and Hrachowitz, M.: Ecosystem adaptation to climate change: the sensitivity of hydrological predictions to time-dynamic model parameters, Hydrol. Earth Syst. Sci., 26, 1295–1318, https://doi.org/10.5194/hess-26-1295-2022, 2022.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth Syst. Sci. Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020a.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil - link to files, Zenodo [dataset], https://doi.org/10.5281/zenodo.3964745, 2020b.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: Hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth Syst. Sci. Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020a.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB), NERC Environmental Information Data Centre [dataset], https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020b.

Crameri, F.: Scientific colour maps, Zenodo [software], https://doi.org/10.5281/zenodo.8409685, 2023.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Deng, C., Zou, J., and Wang, W.: Assimilation of remotely sensed evapotranspiration products for streamflow simulation based on the CAMELS data sets, J. Hydrol., 629, 130574, https://doi.org/10.1016/j.jhydrol.2023.130574, 2024.

Droogers, P. and Allen, R. G.: Estimating reference evapotranspiration under inaccurate data conditions, Irrig. Drain. Syst., 16, 33–45, https://doi.org/10.1023/A:1015508322413, 2002.

ERA5-Land: data documentation: https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation#ERA5Land:datadocumentation-Knownissues, last access: 26 June 2024.

Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F., and Poulin, A.: Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling?, J. Hydrometeorol., 17, 1929–1950, https://doi.org/10.1175/JHM-D-15-0138.1, 2016.

Essou, G. R. C., Brissette, F., and Lucas-Picher, P.: The use of reanalyses and gridded observations as weather input data for a hydrological model: comparison of performances of simulated river flows based on the density of weather stations, J. Hydrometeorol., 18, 497–513, https://doi.org/10.1175/JHM-D-16-0088.1, 2017.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia, Earth Syst. Sci. Data, 13, 3847–3867, https://doi.org/10.5194/essd-13-3847-2021, 2021.

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.

Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: How limited training data affects streamflow prediction, Environ. Model. Softw., 135, 104926, https://doi.org/10.1016/j.envsoft.2020.104926, 2021.

Gomis-Cebolla, J., Rattayova, V., Salazar-Galán, S., and Francés, F.: Evaluation of ERA5 and ERA5-Land reanalysis precipitation datasets over Spain (1951–2020), Atmospheric Res., 284, 106606, https://doi.org/10.1016/j.atmosres.2023.106606, 2023.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B.: circlize implements and enhances circular visualization in R, Bioinformatics, 30, 2811–2812, https://doi.org/10.1093/bioinformatics/btu393, 2014.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hargreaves, G. H. and Samani, Z. A.: Estimating potential evapotranspiration, J. Irrig. Drain. Div., 108, 225–230, https://doi.org/10.1061/JRCEA4.0001390, 1982.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth Syst. Sci. Data, 15, 5755–5784, https://doi.org/10.5194/essd-15-5755-2023, 2023.

Jayathilake, D. I. and Smith, T.: Identifying the influence of systematic errors in potential evapotranspiration on rainfall-runoff models, J. Hydrol. Eng., 27, 04021047, https://doi.org/10.1061/(ASCE)HE.1943-5584.0002157, 2022.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., and Houska, T.: Using hydrological and climatic catchment clusters to explore drivers of catchment behavior, Hydrol. Earth Syst. Sci., 24, 1081–1100, https://doi.org/10.5194/hess-24-1081-2020, 2020.

Keller, V. D. J., Tanguy, M., Prosdocimi, I., Terry, J. A., Hitt, O., Cole, S. J., Fry, M., Morris, D. G., and Dixon, H.: CEH-GEAR: 1 km resolution daily and monthly areal rainfall estimates for the UK for hydrological and other applications, Earth Syst. Sci. Data, 7, 143–155, https://doi.org/10.5194/essd-7-143-2015, 2015.

Kiraz, M., Coxon, G., and Wagener, T.: A priori selection of hydrological model structures in modular modelling frameworks: application to Great Britain, Hydrol. Sci. J., 68, 2042–2056, https://doi.org/10.1080/02626667.2023.2251968, 2023.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, https://doi.org/10.1029/2005WR004362, 2006.

Klingler, C., Schulz, K., and Herrnegger, M.: LamaH-CE: LArge-saMple dAta for Hydrology and environmental sciences for Central Europe, Earth Syst. Sci. Data, 13, 4529–4565, https://doi.org/10.5194/essd-13-4529-2021, 2021.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, Water Resour. Res., 56, e2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

Kollat, J. B., Reed, P. M., and Wagener, T.: When are multiobjective calibration trade-offs in hydrologic models meaningful?, Water Resour. Res., 48, 2011WR011534, https://doi.org/10.1029/2011WR011534, 2012.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Sci. Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023a.

720 Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Zenodo [dataset], https://doi.org/10.5281/zenodo.7944025, 2023b.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train an LSTM on a single basin, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2023-275, in review 2024.

725 Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., Greene, S., Macleod, C. J. A., and Reaney, S. M.: Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain, Hydrol. Earth Syst. Sci., 23, 4011–4032, https://doi.org/10.5194/hess-23-4011-2019, 2019.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual 730 models, Hydrol. Earth Syst. Sci., 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

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, Geosci. Model Dev., 735 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

Mathai, J. and Mujumdar, P. P.: Use of streamflow indices to identify the catchment drivers of hydrographs, Hydrol. Earth Syst. Sci., 26, 2019–2033, https://doi.org/10.5194/hess-26-2019-2022, 2022.

McMillan, H. K., Gnann, S. J., and Araki, R.: Large scale evaluation of relationships between hydrologic signatures and processes, Water Resour. Res., 58, e2021WR031751, https://doi.org/10.1029/2021WR031751, 2022.

740 Meyer Oliveira, A., van Meerveld, H. J. (Ilja), Vis, M., and Seibert, J.: Assessment of the value of remotely sensed surface water extent data for the calibration of a lumped hydrological model, Water Resour. Res., 59, e2023WR034875, https://doi.org/10.1029/2023WR034875, 2023.

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, Hydrol. Earth Syst. Sci., 15, 453–469, https://doi.org/10.5194/hess-745 15-453-2011, 2011.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J. N.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

750 Newman, A., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, Boulder, CO: UCAR/NCAR [dataset], https://doi.org/10.5065/D6MW2F4D, 2014.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

NOAA: CPC global unified temperature data [dataset]. Provided by NOAA Physical Sciences Laboratory. Available from https://psl.noaa.gov, 2019.

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., and Michel, C.: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, J. Hydrol., 320, 62–83, https://doi.org/10.1016/j.jhydrol.2005.07.016, 2006.

Paturel, J. E., Servat, E., and Vassiliadis, A.: Sensitivity of conceptual rainfall-runoff algorithms to errors in input data-case of the GR2M model, J. Hydrol., 168, 111–125, https://doi.org/10.1016/0022-1694(94)02654-T, 1995.

Pedersen, T. L. and Crameri, F.: scico: colour palettes based on the scientific colour-maps, R package version 1.5.0 [software], https://CRAN.R-project.org/package=scico, 2023.

Pool, S., Viviroli, D., and Seibert, J.: Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States, Water Resour. Res., 55, 363–377, https://doi.org/10.1029/2018WR023855, 2019.

Pool, S., Vis, M., and Seibert, J.: Regionalization for ungauged catchments — lessons learned from a comparative large-sample study, Water Resour. Res., 57, e2021WR030437, https://doi.org/10.1029/2021WR030437, 2021.

Priestley, C. H. B. and Taylor, R. J.: On the assessment of surface heat flux and evaporation using large-scale parameters, Mon. Weather Rev., 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2, 1972.

Ranjram, M. and Craig, J. R.: Upscaling hillslope-scale subsurface flow to inform catchment-scale recession behavior, Water Resour. Res., 58, e2021WR031913, https://doi.org/10.1029/2021WR031913, 2022.

Robinson, E. L., Blyth, E., Clark, D. B., Comyn-Platt, E., Finch, J., and Rudd, A. C.: Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961-2015) [CHESS-PE] [dataset]. NERC Environmental Information Data Centre, https://doi.org/10.5285/8baf805d-39ce-4dac-b224-c926ada353b7, 2016.

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 [dataset]. NERC Environmental Information Data Centre, https://doi.org/10.5285/b745e7b1-626c-4ccc-ac27-56582e77b900, 2017a.

Robinson, E. L., Blyth, E. M., Clark, D. B., Finch, J., and Rudd, A. C.: Trends in atmospheric evaporative demand in Great Britain using high-resolution meteorological data, Hydrol. Earth Syst. Sci., 21, 1189–1224, https://doi.org/10.5194/hess-21-1189-2017, 2017b.

Robinson, N., Regetz, J., and Guralnick, R. P.: EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data, ISPRS J Photogramm, 87, 57–67, https://doi.org/10.1016/j.isprsjprs.2013.11.002, 2014.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, https://doi.org/10.5194/hess-4-215-2000, 2000.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.

790 Seibert, J. and Vis, M. J. P.: How informative are stream level observations in different geographic regions?, Hydrol. Process., 30, 2498–2508, https://doi.org/10.1002/hyp.10887, 2016.

Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H. J.: Upper and lower benchmarks in hydrological modelling, Hydrol. Process., 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018.

Singer, M. B., Asfaw, D. T., Rosolem, R., Cuthbert, M. O., Miralles, D. G., MacLeod, D., Quichimbo, E. A., and Michaelides, 795 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.

South, A.: rworldmap: a new R package for mapping global data, R J., 3, 35, https://doi.org/10.32614/RJ-2011-006, 2011.

South, A.: rworldxtra: country boundaries at high resolution, R package version 1.01 [software], https://CRAN.R-project.org/package=rworldxtra, 2012.

800 Tan, M. L., Armanuos, A. M., Ahmadianfar, I., Demir, V., Heddam, S., Al-Areeq, A. M., Abba, S. I., Halder, B., Cagan Kilinc, H., and Yaseen, Z. M.: Evaluation of NASA POWER and ERA5-Land for estimating tropical precipitation and temperature extremes, J. Hydrol., 624, 129940, https://doi.org/10.1016/j.jhydrol.2023.129940, 2023.

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 (1890-2015) [CEH-GEAR] [dataset]. NERC Environmental Information Data Centre, 805 https://doi.org/10.5285/33604ea0-c238-4488-813d-0ad9ab7c51ca, 2016.

Tarek, M., Brissette, F. P., and Arsenault, R.: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America, Hydrol. Earth Syst. Sci., 24, 2527–2544, https://doi.org/10.5194/hess-24-2527-2020, 2020.

Thornthwaite, C. W.: An approach toward a rational classification of climate, Geogr. Rev., 38, 55, 810 https://doi.org/10.2307/210739, 1948.

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., and Cook, R. B.: Daymet: Daily surface weather data on a 1-km grid for North America, version 2 [dataset]. Oak Ridge National Laboratory Distributed Active Archive Center, https://doi.org/10.3334/ORNLDAAC/1219, 2014.

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S. C., Wei, Y., and Wilson, B. E.: Gridded daily weather data for North 815 America with comprehensive uncertainty quantification, Sci. Data, 8, 190, https://doi.org/10.1038/s41597-021-00973-0, 2021.

Vanella, D., Longo-Minnolo, G., Belfiore, O. R., Ramírez-Cuesta, J. M., Pappalardo, S., Consoli, S., D'Urso, G., Chirico, G. B., Coppola, A., Comegna, A., Toscano, A., Quarta, R., Provenzano, G., Ippolito, M., Castagna, A., and Gandolfi, C.: Comparing the use of ERA5 reanalysis dataset and ground-based agrometeorological data under different climates and topography in Italy, J. Hydrol. Reg. Stud., 42, 101182, https://doi.org/10.1016/j.ejrh.2022.101182, 2022.

820 Wang, J., Zhuo, L., Han, D., Liu, Y., and Rico-Ramirez, M. A.: Hydrological model adaptability to rainfall inputs of varied quality, Water Resour. Res., 59, e2022WR032484, https://doi.org/10.1029/2022WR032484, 2023a.

Wang, J., Zhuo, L., Rico-Ramirez, M. A., Abdelhalim, A., and Han, D.: Interacting effects of precipitation and potential evapotranspiration biases on hydrological modeling, Water Resour. Res., 59, e2022WR033323, https://doi.org/10.1029/2022WR033323, 2023b.

825 Wu, X., Su, J., Ren, W., Lü, H., and Yuan, F.: Statistical comparison and hydrological utility evaluation of ERA5-Land and IMERG precipitation products on the Tibetan Plateau, J. Hydrol., 620, 129384, https://doi.org/10.1016/j.jhydrol.2023.129384, 2023.

Xie, W., Yi, S., Leng, C., Xia, D., Li, M., Zhong, Z., and Ye, J.: The evaluation of IMERG and ERA5-Land daily precipitation over China with considering the influence of gauge data bias, Sci. Rep., 12, 8085, https://doi.org/10.1038/s41598-022-12307-830 0, 2022.

Xu, C., Wang, W., Hu, Y., and Liu, Y.: Evaluation of ERA5, ERA5-Land, GLDAS-2.1, and GLEAM potential evapotranspiration data over mainland China, J. Hydrol. Reg. Stud., 51, 101651, https://doi.org/10.1016/j.ejrh.2023.101651, 2024.

Yilmaz, M.: Accuracy assessment of temperature trends from ERA5 and ERA5-Land, Sci. Total Environ., 856, 159182, https://doi.org/10.1016/j.scitotenv.2022.159182, 2023.

835 Zambrano-Bigiarini, M.: hydroGOF: goodness-of-fit functions for comparison of simulated and observed hydrological time series, R package version 0.4-0, Zenodo [software], https://doi.org/10.5281/zenodo.839854, 2023.

Zhao, P. and He, Z.: A first evaluation of ERA5-Land reanalysis temperature product over the Chinese Qilian mountains, Front. Earth Sci., 10, 907730, https://doi.org/10.3389/feart.2022.907730, 2022.