Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR034875

Assessment of the Value of Remotely Sensed Surface Water Extent Data for the Calibration of a Lumped Hydrological Model

Aline Meyer Oliveira[1], H. J. (Ilja) van Meerveld[1], Marc Vis[1], and Jan Seibert[1,2]

[1]Department of Geography, University of Zurich, Zurich, Switzerland, [2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Abstract For many catchments, there is insufficient field data to calibrate the hydrological models that are needed to answer water resources management questions. One way to overcome this lack of data is to use remotely sensed data. In this study, we assess whether Landsat-based surface water extent observations can inform the calibration of a lumped bucket-type model for Brazilian catchments. We first performed synthetic experiments with daily, monthly, and limited monthly data (April–October), assuming a perfect monotonic relation between streamflow and stream width. The median relative performance was 0.35 for daily data and 0.17 for monthly data, where values above 0 imply an improvement in model performance compared to the lower benchmark. This indicates that the limited temporal resolution of remotely sensed data is not an impediment for model calibration. In a second step, we used real remotely sensed water extent data for calibration. For only 76 of the 671 sites the remotely sensed water extent was large and variable enough to be used for model calibration. For 30% of these sites, calibration with the actual remotely sensed water extent data led to a model fit that was better than the lower benchmark (i.e., relative performance >0). Model performance increased with river width and variation therein. This indicates that the coarse spatial resolution of the freely-available, long time series of water extent used in this study hampered model calibration. We, therefore, expect that newer higher-resolution imagery will be helpful for model calibration for more sites, especially when time series length increases.

Plain Language Summary Hydrological models are important for water resources management. The parameters for these models are estimated in a calibration process. Usually, calibration is based on streamflow data from gauging stations. However, for many catchments there are no streamflow data and therefore the calibration of hydrological models is difficult. In this study, we tested whether satellite data that shows the area that is covered by water can be used to calibrate the parameters of a hydrological model for Brazilian catchments. First, we tested if satellite data would be useful if the water extent was perfectly correlated to streamflow and available for every day, month, or month for half of the year due to cloud cover. For two thirds of the catchments, daily observations would be helpful for model calibration, but both monthly data sets were also informative. When we used actual satellite images to calibrate the model for a subset of 76 large rivers, only 30% of them benefitted from these data. This is probably due to inaccuracies in the water extent from satellite images and its coarse spatial resolution. We expect that newer higher-resolution satellite data will be more useful for model calibration, especially when they become available for longer time periods.

Key Points:

- Synthetic (i.e., perfect) daily water extent time series were informative for model calibration for two thirds of the Brazilian study catchments
- Reduction of the temporal resolution to monthly time series did not limit the value of the synthetic water extent data for model calibration
- Actual remotely sensed water extent data was helpful for calibration for only one third of the subset of 76 catchments with large rivers

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Meyer Oliveira,
aline.meyer@geo.uzh.ch

Citation:

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

Received 15 MAR 2023
Accepted 13 OCT 2023

Author Contributions:

Conceptualization: Aline Meyer Oliveira, H. J. (Ilja) van Meerveld, Jan Seibert
Formal analysis: Aline Meyer Oliveira, H. J. (Ilja) van Meerveld, Marc Vis, Jan Seibert
Funding acquisition: H. J. (Ilja) van Meerveld
Investigation: Aline Meyer Oliveira, H. J. (Ilja) van Meerveld
Methodology: Aline Meyer Oliveira
Resources: H. J. (Ilja) van Meerveld
Software: Marc Vis

1. Introduction

Hydrological models aim to represent the flow and storage of water in catchments to answer questions related to water management (e.g., Qin et al., 2013; Seibert & Bergström, 2022), or to predict the impacts of climate change (Driessen et al., 2010; Sorribas et al., 2016) or land use change (Montenegro & Ragab, 2010) on streamflow. They can also be used to fill gaps in hydrological monitoring in space and time, or to estimate water fluxes in unmonitored regions (Bergström, 2006; Hrachowitz et al., 2013). The parameters used to calculate the fluxes in hydrological models usually need to be calibrated. This is typically done by maximizing the agreement between the observed and simulated streamflow (i.e., optimizing model fit). However, streamflow data are available for only a few locations, and many regions are poorly gauged (Hrachowitz et al., 2013; Ruhi et al., 2018). However, alternative data sets (e.g., stream level) can be valuable for hydrological model calibration as well (Etter et al., 2020; Seibert & Vis, 2016; van Meerveld et al., 2017).

Supervision: H. J. (Ilja) van Meerveld, Jan Seibert
Validation: Aline Meyer Oliveira, Marc Vis
Writing – original draft: Aline Meyer Oliveira
Writing – review & editing: Aline Meyer Oliveira, H. J. (Ilja) van Meerveld, Marc Vis, Jan Seibert

Remote sensing is one way to overcome limitations in hydrological field data, as large rivers can be observed from space (Lettenmaier et al., 2015). The Landsat mission, whose first satellite was launched in 1972, now has acquired 50 years of data, with locations typically being observed every 16–18 days. This temporal resolution is sufficient to capture the changing flow conditions in large rivers (with drainage area larger than 1,000 km^2) (Allen et al., 2020). Indeed, remotely sensed water extent imagery has been used to retrieve streamflow, by applying a hydraulic geometry framework that relates streamflow to river width via a power-law relation (e.g., $w = aQ^b$, where w is width, Q is streamflow, and a and b are parameters) (Frasson et al., 2019; Junqueira et al., 2021; Pavelsky, 2014; Pôssa et al., 2020). For example, W. C. Sun et al. (2010) and W. Sun et al. (2015) extracted river width from Synthetic Aperture Radar (SAR) images (12.5 m resolution) to calibrate a hydrological model for two large (545,000 km^2 and 411,000 km^2) catchments in Asia. They concluded that river width is a helpful proxy for streamflow in basins without gauging stations. Meyer Oliveira et al. (2021) similarly used SAR images to calibrate a hydrologic-hydraulic model for the 236,000 km^2 Purus river in the Amazon. They found that the use of SAR data led to a significant improvement in the simulation of the flood extent for the validation period, even though the improvement in the simulation of the streamflow was relatively small. W. Sun et al. (2018) used commercial high-resolution remotely sensed river width data for the simulation of a 33,000 km^2 catchment in China and also concluded that the proposed framework was suitable for ungauged basins. Revilla-Romero et al. (2015) used remotely sensed water extent data from the Global Flood Detection System to calibrate the LISFLOOD model and found that for 21 out of 30 sites (with catchment areas ranging from 27,650 to 4.7 million km^2), these data were useful to estimate streamflow.

The conversion of the remotely sensed water extent data to an estimated streamflow requires additional parameters (a and b in case of the power law relation mentioned above) (Bjerklie et al., 2003; Gleason & Durand, 2020; Lin et al., 2023), which can negatively affect parameter (and thus model simulation) uncertainty. The retrieval of streamflow from remotely sensed water extent observations, furthermore, depends on the adopted method. So far, it is unclear to what extent the temporal and spatial resolution of the remotely sensed data contributes to the final model performance (Allen et al., 2020; Liu et al., 2015). In addition, the previous studies only simulated streamflow for one or a handful of very large rivers. As a result, it is not yet clear for which catchments remotely sensed water extent data is informative for model calibration.

Therefore, in this study, we applied a different approach and used Landsat-based remotely sensed water extent data directly in model calibration to investigate if and to what degree, water extent observations can inform the calibration of a lumped bucket-type hydrological model for catchments in Brazil. The Global Surface Water (GSW) data set (Pekel et al., 2016) provides monthly water extent data derived from Landsat imagery. It is thus readily available for hydrologic modelers and practitioners. Although the resolution of Landsat data is much coarser than for some of the newer satellite products (e.g., CubeSat, QuickBird, RapidEye), we used it here because it is freely available. Furthermore, the long time series of the Landsat data means that it is more likely to include extreme flood and drought events than the shorter time series from newer satellites. We assessed the potential of the monthly water extent data derived from Landsat imagery for 671 catchments in the CAMELS-BR data set (Chagas et al., 2020). We used a systematic approach with both synthetic (i.e., perfect) data and actual remotely sensed water extent data to assess the influence of the temporal resolution and the uncertainty in the water extent data (e.g., due to spatial resolution) on model performance separately. The synthetic data was used to determine whether monthly stream width data would be useful for model calibration if it were perfectly related to streamflow, and if the effect of cloud cover (and thus a reduction of the amount of data available) would affect model performance. Afterward, we assessed the true value of Landsat-derived water extent data for model calibration to determine the effect of uncertainty in the relation between water extent and streamflow on model calibration, and for which catchments these actual remotely sensed data are informative for model calibration. More specifically, we addressed the following research questions:

1. Is the temporal resolution of Landsat imagery sufficient for model calibration if it is perfectly correlated to streamflow?
2. How informative are (actual) remotely sensed water extent data for model calibration for catchments in Brazil?
3. For which types of rivers and catchments are remotely sensed water extent data most informative for model calibration?

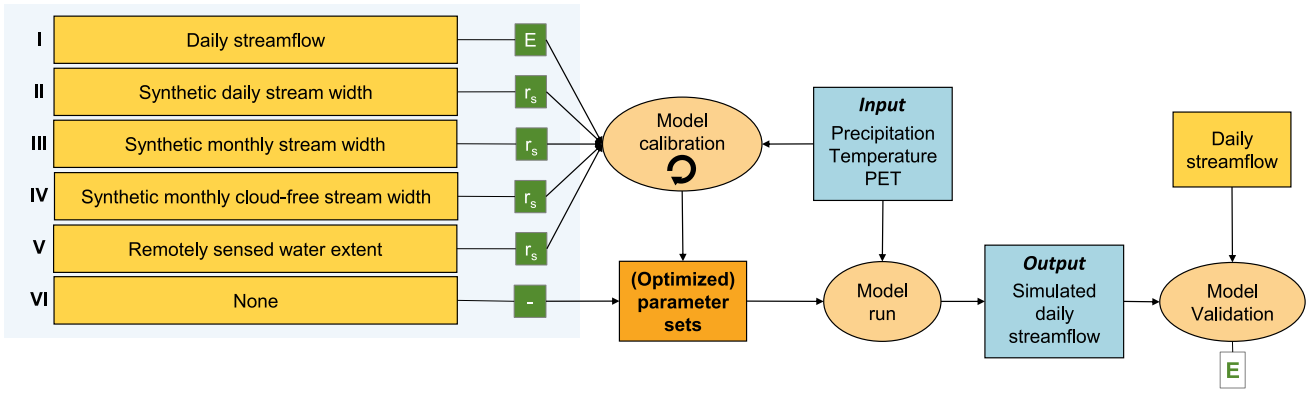

Figure 1. Overview of the approach used in this study, with the different data sets used for model calibration (I–VI in the yellow boxes), the different objective functions (E, non-parametric Kling Gupta efficiency; r_s, Spearman rank correlation; in the green boxes), and the input to the model and the outputs (blue boxes). The different data sets and their temporal resolution are described in Section 2.3.2. PET: potential evapotranspiration.

2. Methods

2.1. Study Design

In this study, we used a systematic approach to determine the value of remotely sensed stream width data for model calibration. We first used the streamflow data from the CAMELS-BR data set (Chagas et al., 2020) in a synthetic experiment approach to determine the influence of the temporal resolution of stream width data if it was available for every catchment and perfectly correlated with streamflow. We calibrated the HBV (Hydrologiska Byrans Vattenavdelning) model (Bergström, 1976; Seibert & Bergström, 2022) on different subsets of the data (daily, monthly, or monthly for the dry season months only) using the Spearman rank correlation (r_s) as the objective function and validated the model on the observed daily streamflow (II–IV in Figure 1; see Section 2.3.2. Model calibration and data sets). These synthetic experiments allowed us to assess the effect of a lack of information on the streamflow volume and the effects of lower temporal resolution data on model calibration performance. Afterward, we determined the actual remotely sensed water extent for all the gauging stations in the CAMELS-BR data set (see Section 2.4 Water extent extraction in Google Earth Engine). For the gauging station sites for which there was enough variation in the water extent, we used these data in model calibration (V in Figure 1) and validated the model again using the observed streamflow data. This step allowed us to determine the effect of uncertainties in the remotely sensed water extent data on model calibration. For each catchment, we compared the model performance to an upper benchmark, that is, calibration based on daily streamflow data (I in Figure 1) and a lower benchmark, that is, the ensemble mean streamflow for 1,000 random parameter sets (VI in Figure 1) (cf. Seibert et al., 2018).

2.2. Streamflow Data Set

The CAMELS-BR data set (Chagas et al., 2020) contains the input data (precipitation, temperature, and monthly potential evapotranspiration [PET]) and streamflow data for 897 catchments across Brazil for the 1980–2018 time period. We restricted the analyses to the 807 catchments for which the consumptive water use and the regulation degree were both less than 50%. This 50% threshold is an arbitrary value and reflects a trade-off between excluding catchments with a large human influence on streamflow, while still having enough catchments for the analyses. For 20 of these 807 catchments, none of the 100,000 model runs with random parameters resulted in a volume error smaller than 30%. Therefore, these catchments were excluded from the analyses as well (see Section 2.3.2). This 30% threshold is also arbitrary but based on the assumption that we can estimate the annual streamflow for a catchment based on the hydro-climatological setting, streamflow data from nearby gauges, or satellite data on the evapotranspiration with a 30% error (see also Section 2.3.2). The 787 remaining catchments cover a range of sizes (11–4.7 million km^2; median: 2,097 km^2), climate (annual precipitation: 584–3,584 mm/year; median: 1,492 mm/year; annual PET/annual P: 0.3–2.0; median: 0.7), and mean annual streamflow (19–2,547 mm/year; median: 546 mm/year).

2.3. Model Application

2.3.1. HBV Model

For the model simulations, we used the HBV model (Bergström, 1976; Lindström et al., 1997) in the software implementation HBV-light (Seibert & Vis, 2012), version 4.0.0.23. The HBV model is a lumped conceptual (bucket-type) model with low data requirements, a short running time, and a relatively small number of parameters (eight, when snowmelt processes are not considered; Table S1 in Supporting Information S1). This allows the model to be calibrated multiple times to assess parameter uncertainty. The HBV model has previously been used to evaluate the value of data (e.g., Etter et al., 2020; Pool et al., 2017; Seibert & Beven, 2009; van Meerveld et al., 2017) and has been applied to a range of catchments, including large catchments (e.g., Graham, 1999; Seibert & Vis, 2016).

The HBV model has four main routines representing snow, soil moisture (SM), groundwater, and routing. The snow routine was not used in this study because of the absence of snow in the study catchments. The SM routine calculates the water balance in the soil, groundwater recharge, and evaporation. Evaporation is equal to the PET as long as SM divided by the maximum soil storage (FC) is higher than a certain threshold (LP) and decreases linearly with SM below this value. Groundwater recharge is calculated based on a relation between SM and the maximum soil storage (FC). The response (or groundwater) routine consists of two connected reservoirs (representing the shallow and deep groundwater). Flow out of these reservoirs depends non-linearly on the storage (via parameters alpha, K1 and K2). The routing routine simulates streamflow at the catchment outlet with a triangular weighting function (Bergström, 1976; Lindström et al., 1997).

2.3.2. Model Calibration and Data Sets

We used the period from 1 January 1997 to 31 August 1999 as a warm-up period and the period from 1 September 1999 to 31 August 2009 for calibration (hydrologic year consistent with CAMELS-BR). The model parameters for the different calibration experiments were optimized using the Genetic Algorithm and Powell optimization (Seibert, 2000) using 5,000 model runs for the genetic algorithm and 1,000 runs for local optimization. To account for parameter equifinality, the optimization was repeated 10 times. The model parameters and the boundaries used for the calibration are given in Table S1 in Supporting Information S1.

For each catchment, we calibrated the model using the different data sets (yellow boxes in Figure 1). For all data sets, the model was ran at a daily time step. For the synthetic experiments used to determine the effect of the lower temporal resolution of remotely sensed data, we pretended that stream width data were available and perfectly correlated to either the daily or the monthly mean streamflow for all the catchments in the CAMELS-BR data set. We calculated the monthly mean, median and maximum streamflow for each catchment from the daily streamflow data and compared these values to the maximum water extent (see Section 2.4 Water extent extraction via Google Earth Engine). The monthly mean and median streamflow data were better correlated to the water extent than the monthly maximum streamflow (Figure S1 in Supporting Information S1). Because there were no systematic differences between the mean and median values, we used monthly mean streamflow for the model calibration.

The CAMELS-BR data set does not contain stream width data and the HBV model does not simulate stream width. Streamflow was instead used as an indicator of stream width with the Spearman rank correlation (r_s) as the objective function in model calibration (green boxes in Figure 1). This assumes that streamflow and stream width are correlated, that is, that the stream is widest when the flow is highest. This approach assumes a strictly monotonic relationship between streamflow and stream width and does not work when the relation between streamflow and width is non-monotonic (i.e., there is considerable hysteresis). It has been successfully used to assess the value of water level data for 671 catchments in the US by Seibert and Vis (2016) and the value of water level class data for 21 catchments in Switzerland and Austria by Etter et al. (2020). The advantage of this approach is that no information on the (shape of the) rating curve is required, and that it does not require any additional parameters to relate streamflow to stream width or vice-versa. A disadvantage is that there is no information regarding the streamflow volume. Therefore, we incorporated a maximum 30% volume error constraint into the calibration, that is, the optimization process only considers simulations for which the volume error was less than 30%. This assumes that we can estimate the water balance of a catchment with a maximum error of 30% based on either knowledge of the hydroclimatic setting, remotely sensed evapotranspiration data, regionalization from gauged catchments in the region, or a few measurements in time covering the full range of streamflow magnitudes (Pool et al., 2017; Seibert & Beven, 2009).

More specifically, we calibrated the model for each catchment using six different data sets (Figure 1).

I. *Upper benchmark*: model calibration with daily streamflow data based on the non-parametric variant of the Kling Gupta efficiency (KGE) (E_u) as objective function. The non-parametric KGE metric (E) consists of three error terms: volume (β), variability (α_{NP}) and dynamics (r_s) (Pool et al., 2018).
II. *Synthetic daily stream width data*: model calibration with daily streamflow data using the Spearman rank correlation (r_s) as the objective function and the <30% volume error constraint. This approach pretends that daily stream width data are available and perfectly correlated with streamflow. The Spearman rank correlation (r_s) only considers the relative ranking between the values, regardless of the absolute values. The Spearman rank correlation (r_s) is the same as the dynamics term (r_s) in the non-parametric KGE metric (E) used for the upper benchmark.
III. *Synthetic monthly stream width data*: model calibration with monthly mean streamflow data using the Spearman rank correlation (r_s) as the objective function and the <30% volume error constraint. The temporal resolution of satellite imagery varies and daily stream width data is unlikely to be available. This approach pretends that stream width data are available only monthly but are perfectly correlated with streamflow.
IV. *Synthetic cloud-free monthly stream width data*: model calibration with monthly mean streamflow data from April–October using the Spearman rank correlation (r_s) as the objective function and the <30% volume error constraint. Some remote sensing approaches (e.g., optical remote sensing) cannot obtain data during wet periods due to frequent cloud cover (Allen et al., 2020). Therefore, we tested if a lack of data due to frequent cloud cover during the wet season affects model calibration. The exact period with frequent cloud cover varies across the country but generally falls between November and March (Figure S2 in Supporting Information S1). Therefore, for this data set we assumed that monthly stream width data are only available from April to Octobe.
V. *Actual remotely sensed water extent data*: Model calibration with actual remotely sensed water extent data based on the Global Surface Water data set (GSW, Pekel et al., 2016), which is based on Landsat data, using the Spearman rank correlation (r_s) as objective function and the <30% volume error constraint. See Section 2.4 for the details about the extraction of the GSW data.
VI. *Lower benchmark*: For the lower benchmark, we assumed that no streamflow or other data would be available (cf., Seibert et al., 2018). Instead, we ran the model with random parameter sets until the <30% volume error was fulfilled for 1,000 times. We then computed the ensemble mean streamflow, and calculated the non-parametric KGE for this ensemble mean streamflow (E_L; Pool et al., 2018).

The comparison of the model performance for the daily streamflow and synthetic daily stream width data sets (I vs. II) allowed us to assess the effect of a lack of information on the streamflow volume (β) on model calibration performance. The comparison of the model performance for the synthetic stream width data sets with a different temporal resolution (II, III, and IV) allowed us to assess the effects of the lower temporal resolution of remotely sensed data on model calibration performance. The comparison of monthly synthetic stream width and actual remotely sensed water extent data sets (III or IV vs. V) allowed us to assess the effects of uncertainties in the remotely sensed water extent data (e.g., due to the coarse spatial resolution of the data) and a non-uniform relation between streamflow and water extent on model calibration. Finally, the comparisons with the lower benchmark (VI) provide information about the value of the data set for model calibration, if no data would be available.

2.3.3. Model Evaluation

For data sets I–V, we obtained 10 calibrated parameter sets for each catchment. We used these parameter sets to simulate daily streamflow for the calibration period (1 September 1999 to 31 August 2009) and the validation period (1 September 1989 to 31 August 1999). For each catchment, we computed the mean of the simulated streamflow for each day for the 10 calibrated parameter sets to obtain the ensemble mean streamflow for each data scenario. We compared the ensemble mean streamflow for the calibration and validation periods to the observed daily streamflow. The agreement between the observed and the simulated (i.e., ensemble mean) streamflow was evaluated with the non-parametric KGE metric (E; Pool et al., 2018). Note that these non-parametric KGE values (E) are not directly comparable with the KGE values (Pool et al., 2018). The results for the calibration period are described in the text of the manuscript. Those for the validation period are similar and given in Supporting Information S1 (Figures S4 and S6 in Supporting Information S1).

To be able to compare the results for the different catchments for which the model efficiency values can vary greatly, and thus to obtain a clearer understanding of the value of the different data sets for model calibration, the

model efficiency (E) for the different calibration strategies was compared to that of the upper (E_U) and lower (E_L) benchmark for each catchment (Seibert et al., 2018), to obtain the relative model efficiency (E_{Rel}):

$$E_{\text{Rel}} = \frac{E - E_L}{E_U - E_L} \tag{1}$$

where E refers to the non-parametric KGE for a specific data set (II–V), E_L to the non-parametric KGE for the lower benchmark (i.e., the Monte Carlo simulations; data set VI) and E_U to the non-parametric KGE of the upper benchmark (i.e., the model calibrated with the daily streamflow data; data set I). A relative efficiency value E_{Rel} greater than 0 indicates that the data set is informative for model calibration, while a negative value indicates that the data are not informative because the simulated streamflow is not better than that of the lower benchmark. A value of E_{Rel} equal to 1 indicates that the data set leads to a streamflow simulation that is as good as the calibration with daily streamflow data. To indicate the effect of the data set on the optimized model parameters, we compared the median value (from the 10 parameter sets) to that for the upper benchmark. To do this, we first scaled all parameter values to a range of 0–1, where 0 is the lowest value of the parameter range and 1 is the highest value (Table S1 in Supporting Information S1).

2.4. Water Extent Extraction in Google Earth Engine

Monthly water extent data were extracted from the GSW data set (Pekel et al., 2016) for every month between 1984 and 2020 using Google Earth Engine and its application programming interface, with a code written in JavaScript (Gorelick et al., 2017). The GSW data set is based on Landsat data: Landsat 5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper-plus (ETM+) and Landsat 8 Operational Land Imager. The data set consists of monthly data for 30-m resolution pixels that are classified as water, not water, or no data. This classification required sophisticated techniques to merge different Landsat missions and was performed with big data techniques (expert systems, visual analytics and evidential reasoning) (Pekel et al., 2016). Note that the monthly water extent data set is limited by the 16–18 days Landsat revisit time. This means that the monthly image is representative of 1–2 day(s) per month, and not the mean nor the maximum water extent for that month.

We extracted the water extent for a circular area around each gauging station. We tested three buffer sizes (radius R of 2, 5, and 10 km around the gauging station) and converted the number of pixels classified as water to an "Equivalent Width" (W), which represents the width of the river if it was a line through the center of the circle (Equation 2):

$$W = \frac{n_{\text{water}}}{n_{\text{valid}}} \cdot n_{\text{total}} \cdot \frac{s}{2R} \tag{2}$$

where n_{water} is the number of pixels classified as water, n_{valid} is the total number of valid pixels (i.e., the difference between the total number of pixels [n_{total}] and the number of pixels with no data [n_{noData}]), R is the buffer radius (in meters) and s is the pixel size (30 m, for Landsat). There was no significant difference in the median Spearman rank correlation (r_s) between the Equivalent Width W and monthly mean streamflow for the three buffer sizes (Kruskall wallis, p-value: 0.932) (Figure S3 in Supporting Information S1). Therefore, the results are presented for the 5-km radius buffer only. Images for which the percentage of NoData pixels exceeded 10% were excluded from the analyses. Images for which the Equivalent Width W was less than the mean minus three times the standard deviation were also excluded as they represented images with very few water pixels.

We constrained the analyses of the value of remotely sensed water extent data for model calibration to rivers for which the minimum water extent (W_{min}) was larger than one (i.e., an equivalent straight line of pixels that is one pixel wide through the buffer area) because images with too few water pixels resulted in noisy data. For the 787 catchments in the database, 144 fulfilled this minimum water extent criteria. To ensure that the Equivalent Width W changed sufficiently throughout the study period, the ratio between the maximum and median water extent also had to be larger than 1.2. For the 787 catchments in the database, 689 fulfilled this variability criteria. Only 89 catchments fulfilled this criteria and the minimum water extent (W_{min}) criteria. The selected value for the minimum variation in water extent was arbitrary. A smaller value would have resulted in a lower signal-to-noise ratio. A larger value would have excluded even more catchments (e.g., for a value of 1.5, only 29 catchments remained in the database).

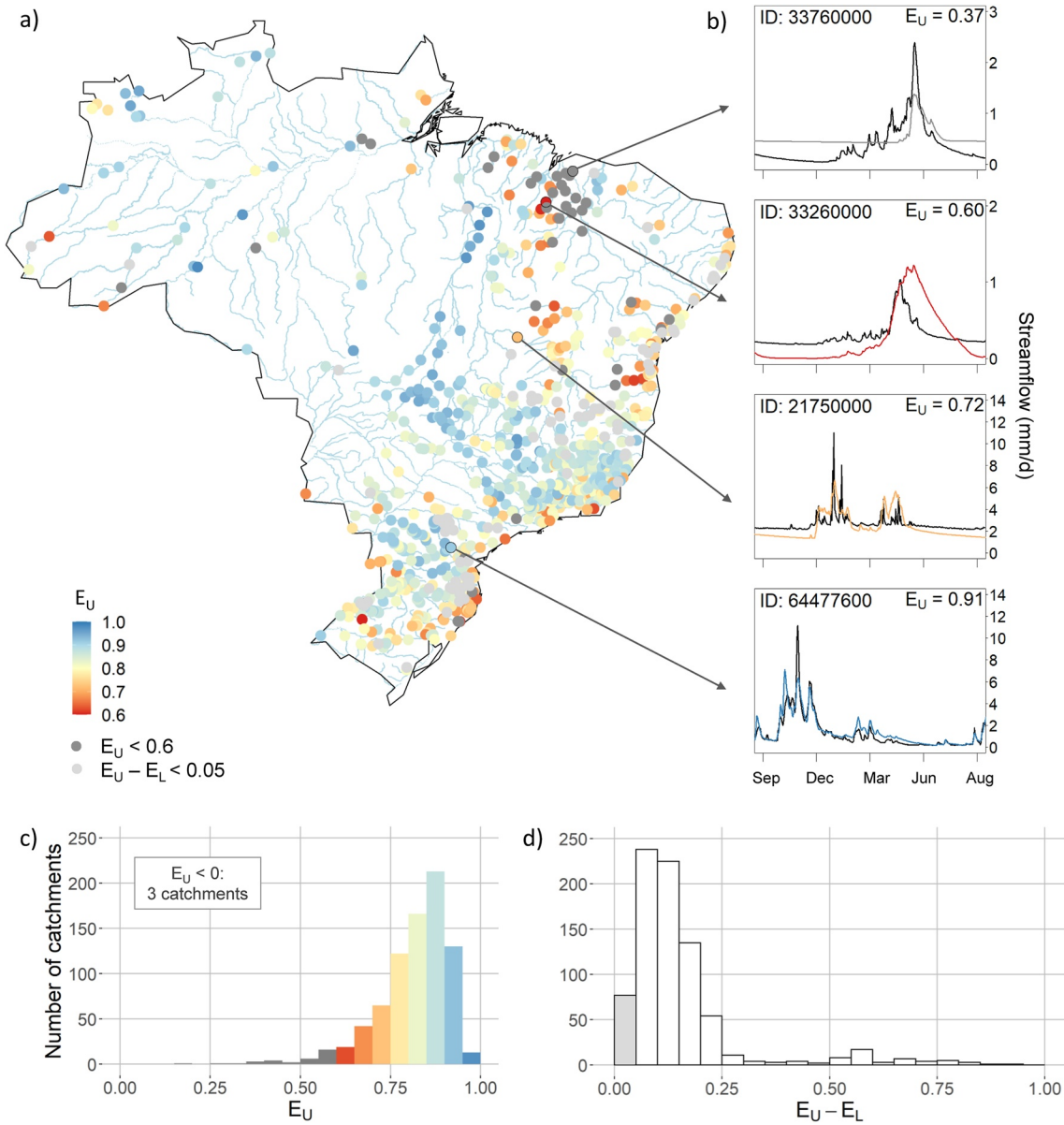

Figure 2. (a) Map showing the model performance (non-parametric Kling Gupta efficiency) for the 787 selected catchments from the CAMELS-BR data set for the calibration period when the HBV model was calibrated with daily streamflow data (upper benchmark; E_U); (b) observed (black line) and simulated hydrographs (colored line) for four catchments with different model performances for hydrologic year 2005/2006; (c) histogram of the upper benchmark values; (d) and the histogram of the difference between the upper and lower benchmark (E_U-E_L). The 116 catchments for which the upper benchmark was less than 0.60 (shown in dark gray in (a) and (c)) or the difference between the upper and lower benchmark was smaller than 0.05 (shown in light gray in (a) and (d)) were excluded from further analyses. Note that the three catchments for which the upper benchmark (E_U) was less than zero are not shown in (c). ID (in b) refers to gauging station ID.

3. Results

3.1. Upper and Lower Benchmarks

The median model efficiency E for the calibration with daily streamflow data (i.e., the upper benchmark) for the 787 selected catchments was 0.84 (mean: 0.81; range: −0.35 to 0.98; Figures 2a and 2c). The model performance was generally better for the wetter catchments (i.e., high mean precipitation, mean streamflow, and runoff ratio) than the drier catchments (e.g., high mean evapotranspiration, aridity index, and frequency of dry days). It was also better for more responsive catchments (e.g., steeper slope of the flow duration curve, higher 95th percentile of specific discharge) and larger catchments (Table S2 in Supporting Information S1). For 37 of the catchments

the upper benchmark was poor ($E_U < 0.60$). A poor model performance can indicate an inadequate model structure or poor data quality (Beven, 2018). Because the focus of this study was on the value of different data sets for model calibration, not model performance itself, we excluded these catchments from further analyses. The excluded catchments are mainly located in the Northeast of Brazil, but also included some catchments in the Amazon, and catchments close to the Atlantic coast (Figure 2a).

For 79 of the remaining 750 catchments, the model performance of the lower benchmark was very similar to that of the upper benchmark (E_U-$E_L < 0.05$; Figure 2d), suggesting that model calibration is not needed for these catchments. These catchments were also excluded from further analysis because the value of alternative data for calibration cannot be assessed for catchments for which calibration does not improve model performance. They did not have any particular characteristics in common and were also not located in a specific region (Figure 2a).

For the remaining 671 catchments for which the influence of the data used for model calibration was tested, the median model performance for the calibration with daily streamflow data was 0.85 (mean: 0.83, range: 0.60–0.98) for the calibration period (Figure 2c) and 0.81 (mean: 0.78; range: −0.63–0.97) for the validation period (Figure S4 in Supporting Information S1).

3.2. Synthetic Experiments: Daily Stream Width Data

Calibration with synthetic daily stream width data (data set II; Figure 1) resulted in a median model performance E of 0.75 (mean: 0.75; range 0.35–0.95) for the calibration period (Figure S5 in Supporting Information S1). For the validation period, the median model performance was also 0.75 (mean: 0.72; range −1.66 to 0.97). The median decline in model performance for calibration with daily streamflow compared to the calibration with synthetic daily stream width data (E_U−E) was 0.08 (mean: 0.09; range: −0.02 to 0.36) for the calibration period and 0.05 (mean: 0.06; range: −0.22 to 1.04) for the validation period. The decline in model performance was larger for drier catchments (with a lower mean streamflow and runoff ratio) than for the wetter catchments (Table 1).

The median relative model performance (E_{Rel}) for the calibration with synthetic stream width data was 0.35 (mean: 0.21; range −3.34 to 1.14) for the calibration period (Figure 3) and 0.46 (mean: −5.34; range −2,218 to 357) for the validation period (Figure S6 in Supporting Information S1). The wide range in E_{Rel} for the validation period is caused by the 144 catchments for which the model performance (E) was very close to the lower benchmark (E−$E_{\text{lower}} < 0.05$). For 452 out of the 671 (67%) catchments, the model performance was better than the lower benchmark ($E_{\text{Rel}} > 0$) for the calibration period, suggesting that stream width data would be informative for the majority of the catchments if it were perfectly correlated with streamflow and available at a high temporal resolution. For the validation period, this was the case for 467 (70%) of the catchments.

For the 33% of the catchments for which the performance of the model was not better than the lower benchmark ($E_{\text{Rel}} < 0$), the median difference between the model performance and the lower benchmark (E−E_L) was only −0.04 (mean: −0.05; range: −0.18 to −0.0005) for the calibration period and −0.03 (mean: −0.04; range: −0.68 to 0.14) for the validation period. Even though the overall model performance varied little from the lower benchmark for these catchments, the parameter range was still constrained by the calibration with the synthetic daily stream width data. In particular, parameters FC, BETA, Alpha, K2 and MAXBAS were better constrained, but parameters LP, K1, and PERC were not (Figure 4).

3.3. Synthetic Experiments: Monthly Stream Width Data

The median change in model performance when using synthetic monthly stream width data instead of synthetic daily stream width data was −0.02 (mean: −0.02; range: −0.25 to 0.16) for the calibration period, and also −0.02 (mean: −0.02; range: −0.24 to 0.46) for the validation period. For only 9% of the catchments the decline in model performance was >0.10. For around a quarter of the catchments (23%) the calibration with monthly synthetic stream width data resulted in a better model performance than calibration with daily synthetic stream width data. The change in model performance due to the decrease in the temporal resolution of the synthetic stream width data was larger for catchments with a more seasonal precipitation pattern and for wetter catchments with a lower frequency of low-flow days (Tables S2 and S3 in Supporting Information S1).

The median relative model performance E_{Rel} for calibration with the synthetic monthly stream width data was 0.17 (mean: −0.01; range −4.11 to 1.38) for the calibration period and 0.22 (mean: −2.56; range −1,605 to

Table 1
Spearman Rank Correlation (r_s) Between the Difference in Model Performances for the Calibration With Different Data Sets (as Specified in the Header of the Table) and Catchment Characteristics

	Difference in non-parametric Kling Gupta efficiency				
	Upper benchmark (I)	Daily stream width (II)	Monthly stream width (III)	Stream width-cloud free (IV)	Water extent (V)
	Versus	Versus	Versus	Versus	Versus
	Daily stream width (II)	Monthly stream width (III)	Stream width-cloud free (IV)	Water extent (V)	Lower benchmark (VI)
Mean precipitation	**−0.22**	**−0.14**	0.05	−0.12	0.35
Mean potential evapotranspiration	**0.28**	0.05	−0.14	0.18	−0.08
Mean evapotranspiration	**−0.34**	**0.09**	**0.04**	−0.05	0.30
Aridity index[a]	**0.29**	**0.13**	−0.09	0.16	−0.30
Seasonality of precipitation	−0.06	**0.19**	0.00	0.17	−0.35
Asynchronicity[b]	**0.34**	**−0.08**	**−0.14**	0.03	0.07
Frequency of high precipitation days	−0.05	0.02	0.05	0.02	−0.19
Duration of high precipitation events	**0.19**	0.07	**0.00**	0.17	**−0.44**
Frequency of dry days	0.02	0.04	0.06	0.08	**−0.27**
Percentage of consumptive use	**0.25**	0.02	**−0.04**	−0.01	−0.24
Percentage of reservoir storage	−0.01	**−0.08**	0.02	−0.11	0.02
Mean streamflow	**−0.26**	**−0.12**	**0.11**	−0.11	0.25
Runoff ratio	**−0.26**	**−0.09**	**0.13**	−0.11	0.19
Stream elasticity	0.05	**−0.15**	0.05	−0.09	0.17
Slope of flow duration curve	**0.13**	**−0.19**	**0.03**	−0.16	**0.28**
Baseflow index	−0.07	**0.12**	−0.08	0.05	0.04
Mean half-flow date	**0.13**	**−0.15**	−0.06	−0.15	**0.42**
Q_5 (low flow)	**−0.29**	**0.10**	**0.02**	0.08	0.02
Q_{95} (high flow)	**−0.22**	**−0.12**	0.11	−0.08	0.20
Frequency of high streamflow days	**0.12**	**−0.13**	0.06	0.07	−0.20
Duration of high streamflow events	**0.12**	**−0.10**	**0.13**	0.13	−0.29
Frequency of low-flow days	**0.13**	**−0.24**	**0.07**	−0.12	**0.26**
Frequency of zero-flow days	**0.18**	**−0.13**	0.03	−0.03	0.13
Gauge elevation	−0.06	**0.12**	0.07	**0.33**	**−0.46**
Catchment mean elevation	**−0.17**	**0.13**	0.09	0.15	**−0.33**
Catchment mean slope	**−0.27**	**0.13**	0.10	0.00	−0.21
Catchment area	**0.09**	−0.07	**−0.08**	**−0.23**	**0.30**

Note. Correlations reported in bold are statistically significant ($p < 0.05$). A darker color shading indicates a stronger negative (red) or positive (blue) correlation. All catchments characteristics were obtained from the CAMELS-BR data set (Chagas et al., 2020).

[a]Aridity index, computed as the ratio of the mean annual potential evapotranspiration and mean annual precipitation. [b]Asynchronicity between the annual precipitation and potential evapotranspiration.

178) for the validation period. When considering only the cloud-free months, the median E_{Rel} was 0.19 (mean: 0.00; range −4.67 to 1.49) for the calibration period (Figure 3) and 0.22 (mean: −21; range −10,158 to 337) for the validation period (Figure S6 in Supporting Information S1). The performance of the model calibrated with synthetic monthly mean river width data was better than the lower benchmark for 388 out of the 671 (58%) catchments (59% for the validation period). This number increased slightly when using data for only the "cloud-free months" (April–October): 394 and 400 catchments for the calibration and validation periods, respectively (Figure 3).

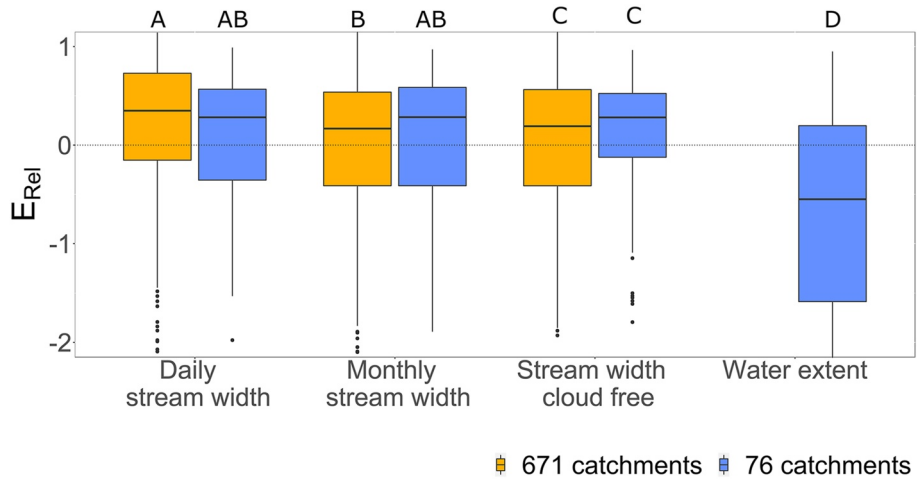

Figure 3. Boxplots of the relative model performance E_{Rel} for the calibration period when the model was calibrated with synthetic daily stream width data, synthetic monthly stream width data, synthetic monthly stream width data for the cloud-free months (April–October), and the actual remotely sensed water extent data. Results are shown for all 671 catchments included in the synthetic study (orange) and all 76 catchments that fulfilled the water extent criteria (blue). The box represents the 25th and 75th percentiles, the line the median, and the whiskers extend to 1.5 times the inter-quartile range. The dots are outliers. The *y* axis is limited between −2 and 1 for better visualization. Groups that share a similar capital letter (plotted above the boxplot) are not significantly different (Kruskal-Wallis, $\alpha > 0.05$). Values of $E_{Rel} > 0$ indicate an improvement in model performance compared to the lower benchmark, and thus that the data are informative for model calibration. The number of catchments with $E_{Rel} > 0$ (i.e., better than the lower benchmark; above the dotted line) is: 452, 388, 394 (out of 671) and 50, 47, 54 (out of 76) for calibration with the synthetic daily, monthly, and cloud-free monthly stream width data, and 24 (out of 76) for the calibration with actual remotely sensed water extent data. Boxplots of absolute values of *E* are presented in Figure S5 in Supporting Information S1.

3.4. Correlation Between Streamflow and Remotely Sensed Water Extent

Of the 89 gauging stations that satisfied the criteria for the minimum water extent and variability in water extent, 76 were included in the data set of the 671 catchments that fulfilled the requirements for the upper and lower benchmark (see Section 3.1). The median Spearman rank correlation (r_s) between the remotely sensed water extent (Equivalent Width *W*) and monthly mean streamflow for these 76 gauging stations was 0.52 (mean: 0.50; range: −0.18–0.94; Figures 5c and 5f, Figures S7 and S8 in Supporting Information S1). For 6 of these 76 gauging stations, there was no significant correlation between the remotely sensed water extent and monthly mean streamflow ($p > 0.05$). The correlation was better for larger catchments with a larger minimum water extent W_{min} (Figure 6). There was no clear sign of hysteresis in the relation between streamflow and water extent or difference in water extent for the rising and falling limbs for the 76 gauging station locations (Figures 5c and 5f and Figure S8 in Supporting Information S1).

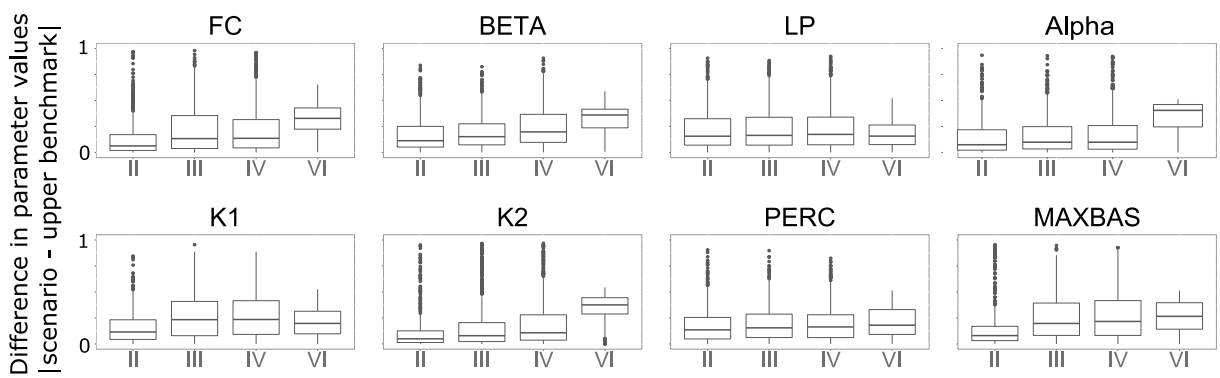

Figure 4. Boxplots of the difference between the median value of the calibrated model parameters for the different scenarios and those for the upper benchmark for all 671 catchments included in this study: II, synthetic daily stream width data; III, synthetic monthly stream width data; IV, synthetic monthly cloud-free stream width data; VI, lower benchmark. The parameter values were re-scaled to values between 0 and 1. See Table S1 in Supporting Information S1 for a description of the parameters and the actual ranges of parameter values used in model calibration. For the results for data set V (remotely sensed water extent), see Figure 9.

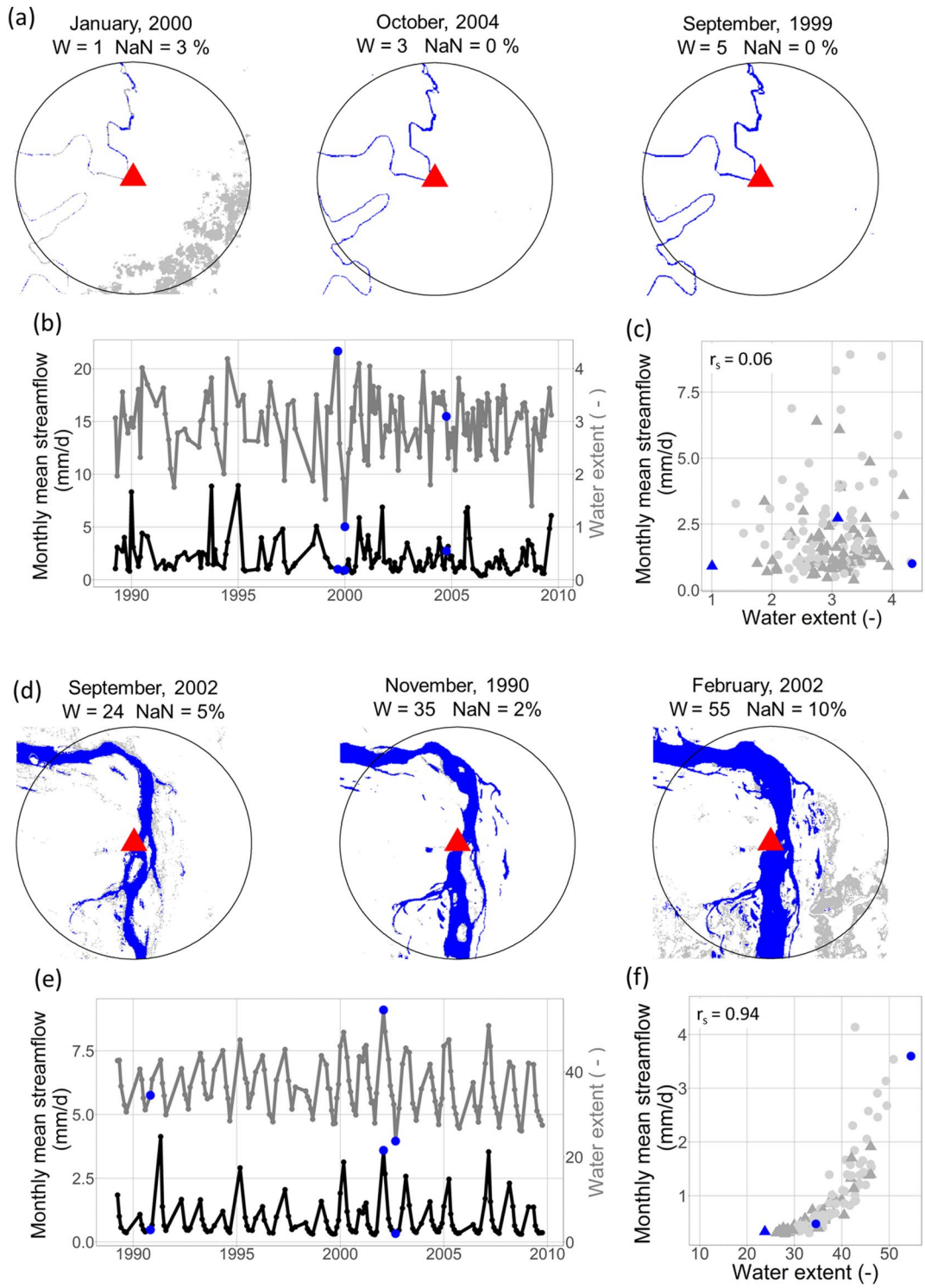

Figure 5. (a, d) Images showing the remotely sensed water extent within a 5 km radius from the gauging station (red triangle) for the day with the minimum, median, and maximum water extent; (b, e) time series of the remotely sensed water extent and monthly mean streamflow; and (c, f) the correlation between the remotely sensed water extent and monthly streamflow for two different catchments: (a–c) gauge ID: 64453000, catchment area 1,040 km^2; (d–f) gauge ID: 26350000, catchment area 194,000 km^2. In (a) and (d), *W* is the remotely sensed water extent, expressed in terms of Equivalent Width and NaN is the percentage of invalid pixels (NoData). In (c) and (f), the circles in light gray represent data points on the rising limb and the dark gray triangles data points on the falling limb. The blue symbols in (b), (c) and (e), (f) represent the streamflow and water extent for the three images shown in (a) and (d), respectively.

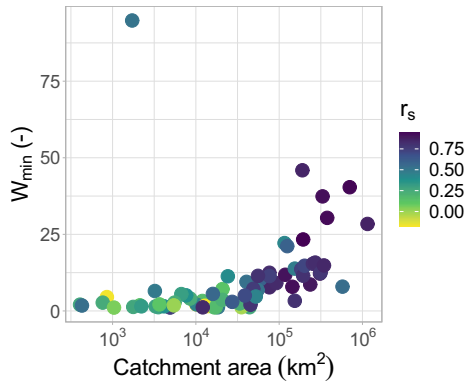

Figure 6. Relation between the catchment area and the minimum remotely sensed water extent, color-coded by the Spearman rank correlation (r_s) between the water extent and monthly mean streamflow. The Spearman rank correlation between the minimum water extent W_{min} and catchment area is 0.74.

3.5. Calibration With Remotely Sensed Water Extent Data

For the calibration with the actual remotely sensed water extent data, the median model performance E for the calibration period was 0.69 (mean: 0.64, range: 0.16–0.96). The median relative model performance E_{Rel} for the calibration period was −0.55 (mean: −1.13; range: −10.4 to 0.95; Figure 3). For 24 out of the 76 (31%) catchments E_{Rel} was larger than zero and the calibration was thus better than the lower benchmark. For 37 catchments (49%), E_{Rel} was larger than −0.5. For the synthetic experiments with monthly data and monthly data for the cloud free period, E_{Rel} was larger than zero for 47–54 out of the same group of 76 catchments (62%–71%) (Figure 3). The results for the validation period were similar, with E_{Rel} being larger than zero for 38% of the catchments, compared to 67% for the synthetic monthly data (Figure S6 in Supporting Information S1).

The performance of the model calibrated with the actual remotely sensed water extent data was better for bigger, lower elevation, or more responsive catchments (e.g., a steeper slope of the flow duration curve) than for smaller, higher elevation, or less responsive catchments (Figure 7; Table 1 and Table S2 in Supporting Information S1). The minimum remotely sensed water extent was also an essential factor for model performance (Figure 8b and Figure S9 in Supporting Information S1): the Spearman rank correlation coefficient for the relation between the difference in model performance for the model calibrated with the actual remotely sensed water extent data (E) and the lower benchmark (E_L) and minimum water extent W was 0.38. The variability in water extent alone affected model performance less (Figure 8c).

The model parameters were overall better constrained when they were calibrated with the remotely sensed water extent data than for the lower benchmark (Figure 9, Figure S10 in Supporting Information S1). In particular, parameters FC, Alpha, K1 and K2 were better constrained. However, other parameters were less sensitive to calibration (BETA, LP, MAXBAS) and for one parameter (PERC) the calibration with water extent data was disinformative (i.e., the median calibrated parameter value was further away from the calibrated value for the upper benchmark than the uncalibrated median parameter value for the lower benchmark).

4. Discussion

4.1. HBV Model Performance for Brazilian Catchments

The HBV model was able to represent the streamflow dynamics for 75% of the study catchments in Brazil well (i.e., $E > 0.60$) when it was calibrated with daily streamflow data (Figure 2). This is a relevant finding because the HBV model had not yet been widely applied to Brazilian catchments (Seibert & Bergström, 2022). The HBV is a lumped model and, therefore, the spatial variation in the hydrological processes is not represented in the model. This can be a problem for large catchments, but the results for the upper benchmark show that streamflow can be simulated adequately for many of the largest catchments in Brazil as well (e.g., $E_U = 0.85$ for the 61,950 km^2 watershed within

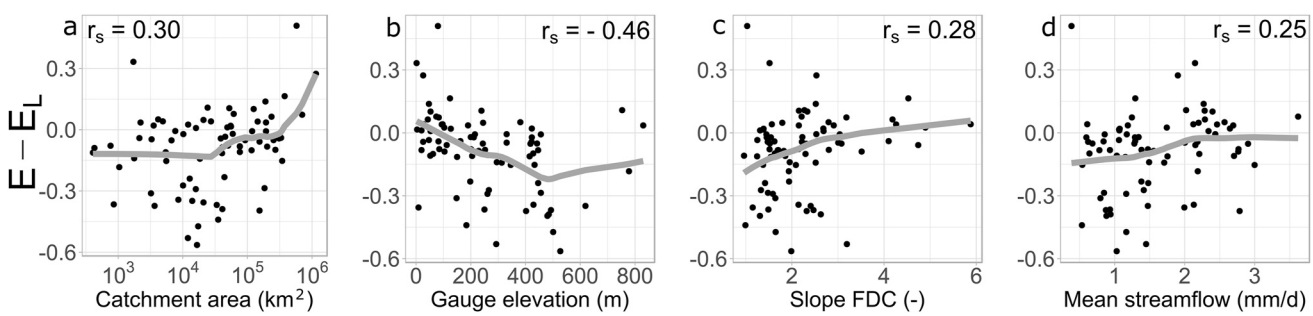

Figure 7. Correlation between the difference in model performance for the model calibrated with actual remotely sensed water extent data (E) and the lower benchmark (E_L) and (a) catchment area, (b) gauge elevation, (c) slope of the flow duration curve, (d) mean streamflow. Each dot represents one catchment; the gray lines show the Lowess regression. The value printed in the upper corner of each subplot is the Spearman rank correlation coefficient.

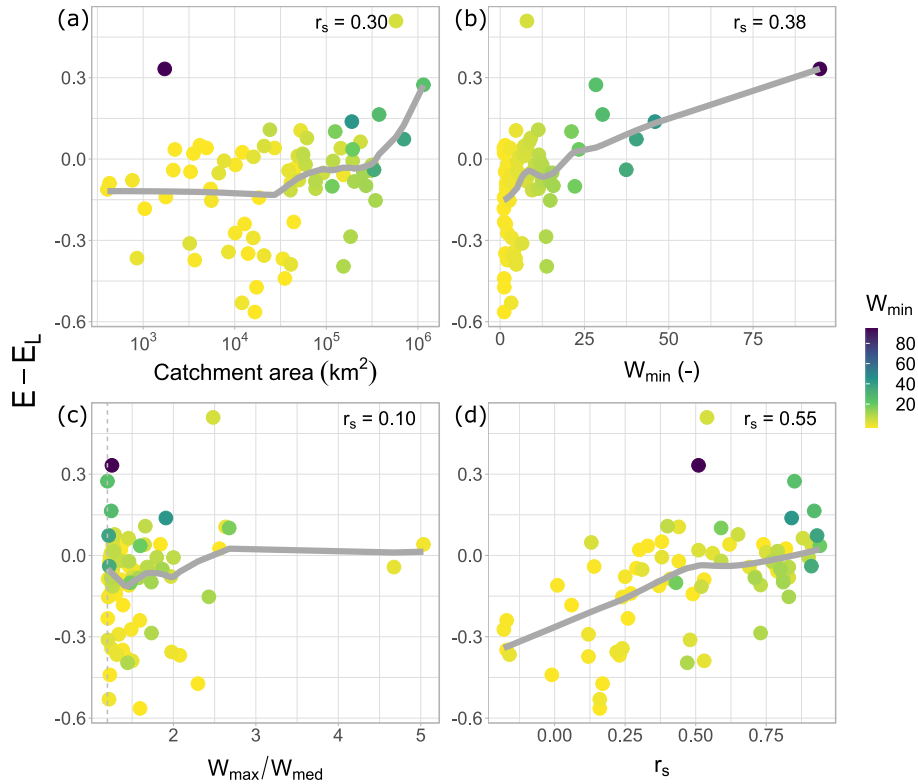

Figure 8. Correlation between the difference in model performance for the model calibrated with the actual remotely sensed water extent data (E) and the lower benchmark (E_L) and (a) catchment area, (b) minimum water extent (W_{min}), (c) variability in water extent, expressed as the ratio between the maximum water extent (W_{max}) and median water extent (W_{med}), and (d) the Spearman rank correlation between the surface water extent and monthly mean streamflow (r_s) for the 76 catchments for which the water extent was large and variable enough to be used in this study. Each dot represents one catchment and is color coded by the minimum remotely sensed water extent W_{min}. The gray line shows the Lowess regression. The value printed inside the graph shows the Spearman rank correlation for the shown relation.

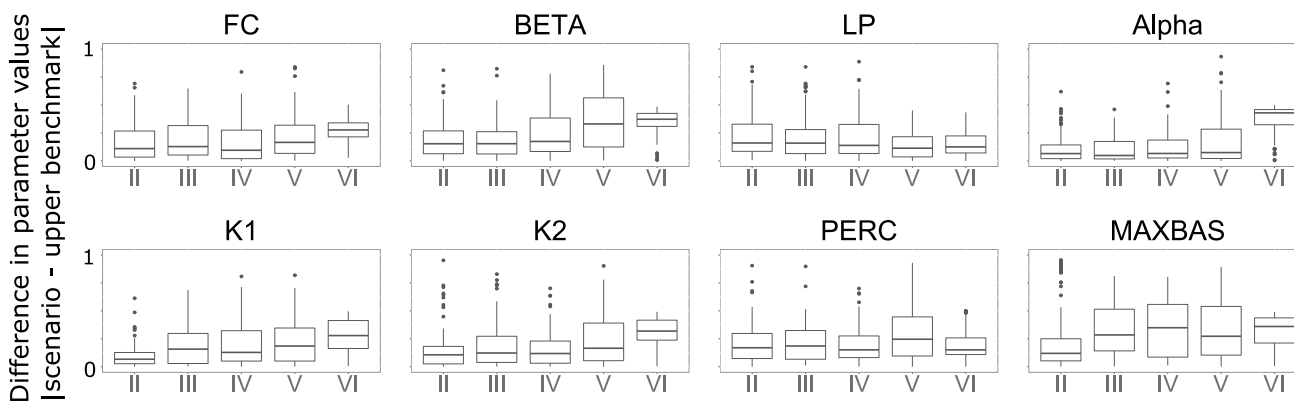

Figure 9. Boxplots of the differences between the median model parameter value obtained by calibration for the different scenarios (data sets II–V) and the median parameter value for the upper benchmark (data set I) for the 76 catchments for which the water extent data was used in model calibration (data set V). Parameter values were re-scaled between 0 and 1 before calculating the difference. For comparison, the results of the lower benchmark (VI) are shown as well, even though these parameters were not calibrated, but still had to result in a volume error <30%. Scenario II, synthetic daily stream width data; III, Synthetic monthly stream width data; IV, Synthetic monthly cloud-free stream width data; V, Actual remotely sensed water extent; VI, Lower benchmark. For the results for all 671 catchments for data sets II–IV and VI, see Figure 4. For the description of the parameters and parameter ranges used in calibration, see Table S1 in Supporting Information S1.

the Uruguai river and $E_U = 0.93$ for the 4.7 million km^2 Amazon river watershed). The performance was generally better for wetter and larger catchments than for drier and smaller catchments (Table S2 in Supporting Information S1). The influence of aridity and catchment size on model performance has been reported by other studies as well (McMillan et al., 2016; Newman et al., 2015; Pechlivanidis & Arheimer, 2015). The catchments for which the model performance was poor were mainly located in Northeastern Brazil, which is a semi-arid region, where channel transmission losses are considerable (Costa et al., 2013). This process is not represented in most hydrological models, leading to a poor performance for most models when they are applied to this region (Siqueira et al., 2018).

For around 9% of the catchments, the upper and the lower benchmark were similar implying that the mean streamflow for the 1,000 uncalibrated runs that satisfied the <30% error in the mean annual streamflow criterium, was very similar to that of the model calibrated with daily streamflow data. This is probably because the volume error constraint of 30% is already highly informative for these catchments. We could not find any clear commonalities between these catchments, but they are overall more responsive, arid catchments. Although the performance of the uncalibrated model is good for these catchments, the values of the parameters that yield the ensemble mean streamflow are unknown and vary widely (e.g., data set VI in Figure S10 in Supporting Information S1). The calibrated models have the advantage of having a set of optimized parameters (e.g., data set I in Figure 4 and Figure S10 in Supporting Information S1) that can be used to simulate streamflow for different scenarios.

4.2. Usefulness of Stream Width Data for Model Calibration

The synthetic daily stream width data successfully informed model calibration for 452 out of 671 catchments. This indicates that stream width data that is perfectly correlated with streamflow are informative for 67% of the catchments in Brazil (Figure 3). It also means that for 33% of the catchments, the use of the Spearman rank correlation instead of the non-parametric efficiency E in the calibration leads to such a deterioration of the model performance that it is no longer better than the lower benchmark ($E_{\text{Rel}} < 0$). However, for 97% of these catchments, the decline in E was less than 0.1, so that the large drop in E_{Rel} can largely be attributed to the good performance of the lower benchmark. As mentioned before, the good performance of the lower benchmark for some catchments is probably due to the 30% volume error constraint. Nonetheless, the calibration with perfect daily stream width data constrained the model parameters considerably (Figure 4).

The wetter catchments were less impacted by the lack of information on stream volume in model calibration (i.e., the use of synthetic stream width data instead of streamflow data) than the drier catchments. This was also found by Seibert and Vis (2016) for catchments in the US. They suggested that additional information on the water balance may be needed for the drier catchments. We included the 30% volume error constraint for all our model calibrations. Although this constrained most model parameters (Figure 4), it was not sufficient to avoid the reduction in model performance when using the synthetic stream width data set instead of the daily streamflow for the dry catchments.

The decrease in the temporal resolution of the synthetic stream width data (from daily to monthly values) mainly impacted the wetter catchments with a lower frequency of low-flow days (Table 1). The reduction in model performance can be related to short floods (time scales less than a month) that may not have been captured well by the monthly average streamflow. In contrast, for 23% of the catchments the model performance was higher when the model was calibrated with less data (synthetic monthly mean stream width vs. synthetic daily stream width). This may be related to overfitting to the objective function (in this case, the Spearman rank correlation) leading to a decrease of the overall model performance.

The median performance for the model calibrated with the synthetic monthly-cloud free stream width data was not very different from the synthetic monthly stream width data (Figure 3). The number of catchments for which $E_{\text{Rel}} > 0$ was even higher when the model was calibrated only with the synthetic data from the cloud-free months. This suggests that the cloud-free data set was more informative for the representation of mainly the dry periods. Several other studies have shown that streamflow (e.g., Pool et al., 2017; Seibert & Beven, 2009) and stream level (Etter et al., 2020) data are highly redundant and that a limited number of measurements can be almost as informative as a large number of measurements. Overall, these results show that the lower temporal resolution of remotely sensed stream observations does not considerably hamper their value for hydrological model calibration. Even if 5 months of data need to be excluded per year due to cloud cover, this does not limit its value for the calibration of hydrological models for most catchments.

4.3. Usefulness of Landsat-Based Water Extent Data for Model Calibration

For 24 out of the 76 (31%) catchments, Landsat-based water extent data were informative for model calibration, that is, it resulted in a better streamflow simulation than the lower benchmark. The experiments with the synthetic

data show that the temporal resolution of stream width or water extent did not impair model calibration considerably. This allows us to attribute the poor performance for the model calibration with actual remotely sensed water extent data to the poor correlation with the monthly streamflow (Figure 8), for example, due to the noise in the water extent data, rather than the low temporal resolution of the data. The main assumption for our approach is that there is a strict monotonic relation between water extent and streamflow. This was indeed the case for many catchments, particularly the larger ones (Figure 8), but not for all catchments (Figure S8 in Supporting Information S1). For other catchments it may have been the already good performance of the lower benchmark with the 30% volume error constrained that caused the additional information on the water extent to not be informative. Note that we also tested the use of a 20% or 40% volume error constraint, but these results were similar (Figure S11 in Supporting Information S1). Even though only one-third of the catchments (24 out of 76) benefitted from the remotely sensed water extent data in terms of model performance, calibration with water extent led to parameter values that were more similar to those of the upper benchmark (Figure 9). However, in some cases, remotely sensed water extent data were disinformative for model calibration (Kauffeldt et al., 2013) due to inaccurate estimates of water extent (see Section 4.4).

The value of the remotely sensed water extent data for model calibration depended on the correlation between the remotely sensed water extent and streamflow (r_s = 0.55; Figure 8d). For the 24 catchments for which the calibration with water extent data led to a better model performance than the lower benchmark, the Spearman rank correlation ranged from 0.13 to 0.94 (median = 0.63). These catchments are large (>1,500 km^2, median: 53,770 km^2; median streamflow 685 m^3/s) and the rivers are wide (median W_{min} = 8; Figure 8). This indicates that the coarse spatial resolution of Landsat imagery was a main factor that impaired model calibration (see also Section 4.4). However, the limited revisit time of Landsat (16–18 days) may result in a less accurate estimate of the mean monthly surface water extent for quickly responding (small) rivers as well, and thus a lower correlation between the remotely sensed water extent and mean monthly streamflow for these rivers.

The coarse resolution of the water extent data has a particularly large effect on the temporal dynamics of the water extent when there are only few pixels with water (i.e., low signal-to-noise ratio). Even though the spatial resolution of the water extent data set is 30 m, Allen and Pavelsky (2018) reported that river width data are only sufficiently accurate for rivers wider than 90 m (i.e., W_{min} = 3). If only the catchments for which W_{min} > 3 are considered, there would be 49 catchments left for the analysis. For 18 of these 49 catchments (37%), E_{Rel} was larger than zero. For the nine catchments with W_{min} > 20, seven had E_{Rel} larger than zero (78%).

4.4. Remotely Sensed Water Extent as a Proxy for Streamflow

The correlation between the remotely sensed water extent and monthly streamflow was the main factor affecting the value of remotely sensed water extent data for model calibration (Figure 8d). The Spearman rank correlation between the remotely sensed water extent and streamflow for the 76 catchments ranged from −0.18 to 0.94 (median = 0.52; Figure 6 and Figure S7 in Supporting Information S1), and depended on the catchment size (Figure 6; r_s = 0.74). Previous studies that used remote sensing data with a higher spatial resolution reported better correlations between water extent and streamflow, but were usually restricted to one or a few catchments. For example, Pavelsky (2014) found that the coefficient of determination between streamflow and RapidEye water extent imagery with a 5-m spatial resolution ranged from r^2 = 0.19–0.94 for a river in Alaska. Junqueira et al. (2021) used Planet CubeSat data with a near daily revisit time at a 3-m spatial resolution, to estimate streamflow at one gauging station (ID = 26350000) in Araguaia river, in Brazil. They reported a coefficient of determination r^2 of 0.96 for the relation between water extent and water level. The Spearman rank correlation between the remotely sensed water extent and streamflow for this gauging station is 0.94 (Figure 5f). Revilla-Romero et al. (2014) analyzed 322 sites and reported a correlation r > 0.3 for 169 sites, and a correlation r > 0.5 for 42 sites. The spatial resolution of their satellite imagery was 10 km. The sites with a higher correlation had a mean streamflow larger than 500 m^3/s, a river width wider than 1 km, and were generally located in floodplain areas.

The method for water extent extraction adopted in this study has the advantage of being simple and can easily be applied via Google Earth Engine. However, it has the disadvantage that it may capture the extent of a larger river if the gauging station is located near the mouth of the tributary. This happened for catchment 87317060 (outlier in Figure 6), for which the gauging station is located close to a lagoon, resulting in high values of W_{min}, even though the river itself is small. For catchment 56992000, the correlation between streamflow and water extent was low because dam construction on the main river caused a higher W_{min} for the tributary, even though the flow out of

this tributary was not or only minimally affected by the reservoir. More robust methods exist for water extent extraction (Allen & Pavelsky, 2015; Hou et al., 2022; Pôssa et al., 2020), especially for geomorphological investigations. Investigating them goes beyond the scope of this study but we can conclude that the method adopted here can capture water extent dynamics, particularly for large and wide rivers with seasonally flooded floodplains (e.g., Figure 5 and Figure S8 in Supporting Information S1).

For some incised rivers, the river width does not change considerably when the stream level and flow increase or decrease (e.g., in canyons and deeply incised rivers) and the water extent data would not be useful as a proxy for streamflow. We removed these rivers from the analyses, by not considering sites for which the ratio between the maximum and median water extent was smaller than 1.2. Still, gauging stations are preferably located at confined cross-sections (Di Baldassarre & Montanari, 2009). Thus, we expect a similar or better correlation between water extent and streamflow for ungauged locations, where the stream width may vary more. This implies that our results are an underestimation of the performance of remotely sensed water extent data as a proxy of streamflow and, thus, the ability of water extent data to inform hydrological models.

Other rivers may flow overbank with extended flooding remaining after the water level in the main river and flow have receded. This would lead to a hysteretic relation between water extent and flow. We did not see any clear indication of hysteresis in the data for the 89 gauging stations for which the water extent was large and variable enough (see example in Figures 5c and 5f and Figure S8 in Supporting Information S1). For many gauges, there are, however, far more water extent data points for the falling limb than on the rising limbs because this period is longer and there is more cloud cover during the rising stage (Hou et al., 2022).

The correlation between water extent and streamflow was especially low for the smaller catchments (Figure 8 and Figure S9 in Supporting Information S1), suggesting that the data for these catchments is influenced by the extraction of the water extent and especially the resolution of the Landsat data. Newer satellites with a finer spatial resolution are likely to provide more accurate water extent estimates for these catchments. One main disadvantage is that the data are not freely available (e.g., SPOT, RapidEye). Other missions have been launched recently, thus having a limited temporal coverage (e.g., Sentinel-2) and are unlikely to contain many large flood events. Our analyses suggest that once these satellite products become more affordable and have longer time series, these data could be useful for model calibration. They will be especially informative for the streams in our data set for which the Landsat-derived water extent was too small and varied too little to be used in the model calibration. However, for the larger rivers, satellites with a spatial resolution of around 1-m may be unsuitable due to the small spatial coverage of each image (e.g., IKONOS, QuickBird) (Huang et al., 2018). The SWOT mission will provide streamflow estimates based on water extent and water surface heights (Biancamaria et al., 2016). This additional variable may be helpful for streamflow retrieval, especially for incised rivers. Still, the spatial resolution of the SWOT mission will be limited to rivers wider than 100 m, so that its usefulness for model calibration may also be limited to the largest rivers.

5. Conclusions

We systematically analyzed whether a lumped conceptual hydrological model (HBV model) could be calibrated with remotely sensed water extent data for 671 catchments in the CAMELS-BR data set. Overall, model performance was better for larger, wetter catchments than for smaller, drier ones. If water extent data were perfectly correlated with streamflow and available at a daily resolution, water extent observations would be useful for model calibration for around two thirds of the catchments. For most of the other catchments, the river width data would not improve the streamflow simulations compared to the lower benchmark (i.e., model runs with randomly generated sets of parameters and a water balance constraint) because the lower benchmark already performed well. In these cases, the river width data would still help to constrain most of the model parameters. Reducing the data to a monthly resolution or using only monthly data from the cloud-free months (here April–October) did not considerably change the model results, suggesting that the limited temporal resolution of the remote sensing data does not considerably influence its usefulness for model calibration.

For only 12% of the gauging stations in the CAMELS-BR data set the water extent was large and variable enough to be observable with Landsat data. The median correlation between streamflow and water extent for these 76 catchments was 0.52 (range: −0.18–0.94). A poor correlation between remotely sensed water extent and streamflow can be due to the low spatial resolution or accuracy of the remote sensing data, or hysteresis in the relation

between water extent and streamflow due to backwater effects or overbank flooding. The latter was not observed for the 76 gauging station sites for which the water extent was large and variable enough to be used in model calibration. The correlation between the remotely sensed water extent and streamflow was much better for rivers with a larger minimum water extent, draining larger catchments than for smaller rivers draining smaller catchments.

Model calibration with the remotely sensed water extent data led to a better model performance than the lower benchmark for only 24 of the 76 catchments. These were large catchments (>1,500 km^2) with wide rivers, and a large minimum water extent. Even when the calibration with the remotely sensed water extent data did not lead to a better streamflow simulation than the lower benchmark, the model parameters were more constrained and closer to those obtained from the calibration with daily streamflow data (i.e., the upper benchmark). We expect that remotely sensed water extent data will be more valuable than indicated by these results because gauging stations are often located in incised channels where river width changes little and extensive overbank flooding is limited. In ungauged catchments, less incised river sections where water extent varies more, should be selected for water extent extraction. Commercial satellite data with a higher spatial resolution than the Landsat data is expected to be useful for model calibration for more locations, especially when these time series have become longer and include more flood events.

Data Availability Statement

The CAMELS-BR dataset is available from Chagas et al. (2020): https://zenodo.org/record/3964745#.Y-taVHaZO5c.

The global monthly surface water extent is available from Pekel et al. (2016): https://global-surface-water.appspot.com/download. The JavaScript code to extract these data via Google Earth Engine is available at: https://zenodo.org/record/8200185 (last accessed on 31 July 2023).

Acknowledgments
We thank the anonymous reviewers for their constructive comments. This research is part of the WatForFun (Water level regimes and tropical forest functioning in floodplains and headwater springs) project funded by the Swiss National Science Foundation (project Grant: 186303).

References

Allen, G. H., & Pavelsky, T. M. (2018). Global extent of rivers and streams. *Science*, *361*(6402), 585–588. https://doi.org/10.1126/science.aat0636

Allen, G. H., & Pavelsky, T. M. (2015). Patterns of river width and surface area revealed by the satellite-derived North American River Width data set. *Geophysical Research Letters*, *42*(2), 395–402. https://doi.org/10.1002/2014GL062764

Allen, G. H., Yang, X., Gardner, J., Holliman, J., David, C. H., & Ross, M. (2020). Timing of Landsat overpasses effectively captures flow conditions of large rivers. *Remote Sensing*, *12*(9), 1510. https://doi.org/10.3390/RS12091510

Bergström, S. (1976). Development and application of a conceptual runoff model for Scandinavian catchments (Vol. *7*).

Bergström, S. (2006). Applications of the HBV hydrological model in prediction in ungauged basins. In *Large sample basin experiments for hydrological model parameterization: Results of the model parameter experiment MOPEX* (Vol. 307, pp. 97–107). IAHS Publ.

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *WIREs Water*, *5*(3), 1–8. https://doi.org/10.1002/wat2.1278

Biancamaria, S., Lettenmaier, D. P., & Pavelsky, T. M. (2016). The SWOT mission and its capabilities for land hydrology. *Surveys in Geophysics*, *37*(2), 307–337. https://doi.org/10.1007/s10712-015-9346-y

Bjerklie, D. M., Dingman, S. L., Vorosmarty, C. J., Bolster, C. H., & Congalton, R. G. (2003). Evaluating the potential for measuring river discharge from space. *Journal of Hydrology*, *278*(1–4), 17–38. https://doi.org/10.1016/S0022-1694(03)00129-X

Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., & Siqueira, V. A. (2020). CAMELS-BR: Hydrometeorological time series and landscape attributes for 897 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075–2096. https://doi.org/10.5194/essd-12-2075-2020

Costa, A. C., Foerster, S., de Araújo, J. C., & Bronstert, A. (2013). Analysis of channel transmission losses in a dryland river reach in North-Eastern Brazil using streamflow series, groundwater level series and multi-temporal satellite data. *Hydrological Processes*, *27*(7), 1046–1060. https://doi.org/10.1002/hyp.9243

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, *13*(6), 913–921. https://doi.org/10.5194/hess-13-913-2009

Driessen, T. L. A., Hurkmans, R. T. W. L., Terink, W., Hazenberg, P., Torfs, P. J. J. F., & Uijlenhoet, R. (2010). The hydrological response of the Ourthe catchment to climate change as modelled by the HBV model. *Hydrology and Earth System Sciences*, *14*(4), 651–665. https://doi.org/10.5194/hess-14-651-2010

Etter, S., Strobl, B., Seibert, J., & van Meerveld, H. J. I. (2020). Value of crowd-based water level class observations for hydrological model calibration. *Water Resources Research*, *56*(2), e2019WR02610. https://doi.org/10.1029/2019wr026108

Frasson, R. P. D. M., Pavelsky, T. M., Fonstad, M. A., Durand, M. T., Allen, G. H., Schumann, G., et al. (2019). Global relationships between river width, slope, catchment area, meander wavelength, sinuosity, and discharge. *Geophysical Research Letters*, *46*(6), 3252–3262. https://doi.org/10.1029/2019GL082027

Gleason, C. J., & Durand, M. T. (2020). Remote sensing of river discharge: A review and a framing for the discipline. *Remote Sensing*, *12*(7), 1–28. https://doi.org/10.3390/rs12071107

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Graham, L. P. (1999). Modeling runoff to the Baltic Sea. *Ambio*, *28*, 328–334.

Hou, J., Van Dijk, A. I. J. M., & Renzullo, L. J. (2022). Merging Landsat and airborne LiDAR observations for continuous monitoring of floodplain water extent, depth and volume. *Journal of Hydrology*, *609*(March), 127684. https://doi.org/10.1016/j.jhydrol.2022.127684

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)-a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Huang, C., Chen, Y., Zhang, S., & Wu, J. (2018). Detecting, extracting, and monitoring surface water from space using optical sensors: A review. *Reviews of Geophysics*, *56*(2), 333–360. https://doi.org/10.1029/2018RG000598

Junqueira, A. M., Mao, F., Mendes, T. S. G., Simões, S. J. C., Balestieri, J. A. P., & Hannah, D. M. (2021). Estimation of river flow using CubeSats remote sensing. *Science of the Total Environment*, *788*, 147762. https://doi.org/10.1016/j.scitotenv.2021.147762

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C. Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *17*(7), 2845–2857. https://doi.org/10.5194/hess-17-2845-2013

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., & Wood, E. F. (2015). Inroads of remote sensing into hydrologic science during the WRR era. *Water Resources Research*, *51*(9), 7309–7342. https://doi.org/10.1002/2015WR017616

Lin, P., Feng, D., Gleason, C. J., Pan, M., Brinkerhoff, C. B., Yang, X., et al. (2023). Remote Sensing of Environment Inversion of river discharge from remotely sensed river widths: A critical assessment at three-thousand global river gauges. *Remote Sensing of Environment*, *287*(January), 113489. https://doi.org/10.1016/j.rse.2023.113489

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Serbström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Liu, G., Schwartz, F. W., Tseng, K. H., & Shum, C. K. (2015). Discharge and water-depth estimates for ungauged rivers: Combining hydrologic, hydraulic, and inverse modeling with stage and water-area measurements from satellites. *Water Resources Research*, *51*(8), 6017–6035. https://doi.org/10.1002/2015WR016971

McMillan, H. K., Booker, D. J., & Cattoën, C. (2016). Validation of a national hydrological model. *Journal of Hydrology*, *541*, 800–815. https://doi.org/10.1016/j.jhydrol.2016.07.043

Meyer Oliveira, A., Fleischmann, A. S., & Paiva, R. C. D. (2021). On the contribution of remote sensing-based calibration to model hydrological and hydraulic processes in tropical regions. *Journal of Hydrology*, *597*, 126184. https://doi.org/10.1016/j.jhydrol.2021.126184

Montenegro, A., & Ragab, R. (2010). Hydrological response of a Brazilian semi-arid catchment to different land use and climate change scenarios: A modelling study. *Hydrological Processes*, *24*(19), 2705–2723. https://doi.org/10.1002/hyp.7825

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). 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. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Pavelsky, T. M. (2014). Using width-based rating curves from spatially discontinuous satellite imagery to monitor river discharge. *Hydrological Processes*, *28*, 3035–3040. https://doi.org/10.1002/hyp.10157

Pechlivanidis, I. G., & Arheimer, B. (2015). Large-scale hydrological modelling by using modified PUB recommendations: The India-HYPE case. *Hydrology and Earth System Sciences*, *19*(11), 4559–4579. https://doi.org/10.5194/hess-19-4559-2015

Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, *554*, 613–622. https://doi.org/10.1016/j.jhydrol.2017.09.037

Pôssa, É. M., Maillard, P., & de Oliveira, L. M. (2020). Discharge estimation for medium-sized river using multi-temporal remote sensing data: A case study in Brazil. *Hydrological Sciences Journal*, *65*(14), 2402–2418. https://doi.org/10.1080/02626667.2020.1808220

Qin, H., Cao, G., Kristensen, M., Refsgaard, J. C., Rasmussen, M. O., He, X., et al. (2013). Integrated hydrological modeling of the North China Plain and implications for sustainable water management. *Hydrology and Earth System Sciences*, *17*(10), 3759–3778. https://doi.org/10.5194/hess-17-3759-2013

Revilla-Romero, B., Beck, H. E., Burek, P., Salamon, P., de Roo, A., & Thielen, J. (2015). Filling the gaps: Calibrating a rainfall-runoff model using satellite-derived surface water extent. *Remote Sensing of Environment*, *171*, 118–131. https://doi.org/10.1016/j.rse.2015.10.022

Revilla-Romero, B., Thielen, J., Salamon, P., De Groeve, T., & Brakenridge, G. R. (2014). Evaluation of the satellite-based global flood detection system for measuring river discharge: Influence of local factors. *Hydrology and Earth System Sciences*, *18*(11), 4467–4484. https://doi.org/10.5194/hess-18-4467-2014

Ruhi, A., Messager, M. L., & Olden, J. D. (2018). Tracking the pulse of the Earth's fresh waters. *Nature Sustainability*, *1*(4), 198–203. https://doi.org/10.1038/s41893-018-0047-7

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J., & Bergström, S. (2022). A retrospective on hydrological catchment modelling based on half a century with the HBV model. *Hydrology and Earth System Sciences*, *26*(5), 1371–1388. https://doi.org/10.5194/hess-26-1371-2022

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: How many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892. https://doi.org/10.5194/hess-13-883-2009

Seibert, J., & Vis, M. J. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., & Vis, M. J. P. (2016). How informative are stream level observations in different geographic regions? *Hydrological Processes*, *30*(14), 2498–2508. https://doi.org/10.1002/hyp.10887

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., et al. (2018). Toward continental hydrologic-hydrodynamic modeling in South America. *Hydrology and Earth System Sciences*, *22*(9), 4815–4842. https://doi.org/10.5194/hess-22-4815-2018

Sorribas, M. V., Paiva, R. C., Melack, J. M., Bravo, J. M., Jones, C., Carvalho, L., et al. (2016). Projections of climate change effects on discharge and inundation in the Amazon basin. *Climatic Change*, *136*(3), 555–570. https://doi.org/10.1007/s10584-016-1640-2

Sun, W., Fan, J., Wang, G., Ishidaira, H., Bastola, S., Yu, J., et al. (2018). Calibrating a hydrological model in a regional river of the Qinghai–Tibet plateau using river water width determined from high spatial resolution satellite images. *Remote Sensing of Environment*, *214*, 100–114. https://doi.org/10.1016/j.rse.2018.05.020

Sun, W., Ishidaira, H., Bastola, S., & Yu, J. (2015). Estimating daily time series of streamflow using hydrological model calibrated based on satellite observations of river water surface width: Toward real world applications. *Environmental Research*, *139*(2015), 36–45. https://doi.org/10.1016/j.envres.2015.01.002

Sun, W. C., Ishidaira, H., & Bastola, S. (2010). Towards improving river discharge estimation in ungauged basins: Calibration of rainfall-runoff models based on satellite observations of river flow width at basin outlet. *Hydrology and Earth System Sciences*, *14*(10), 2011–2022. https://doi.org/10.5194/hess-14-2011-2010

van Meerveld, H. J. I., Vis, M. J. P., & Seibert, J. (2017). Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences*, *21*(9), 4895–4905. https://doi.org/10.5194/hess-21-4895-2017