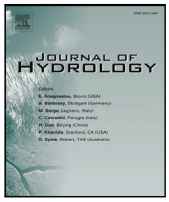
Journal of Hydrology 621 (2023) 129607

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Improving global hydrological simulations through bias-correction and multi-model blending

Amulya Chevuturi [a,*], Maliko Tanguy [a], Katie Facer-Childs [a], Alberto Martínez-de la Torre [a,b], Sunita Sarkar [a], Stephan Thober [c], Luis Samaniego [c,d], Oldrich Rakovec [c,e], Matthias Kelbling [c], Edwin H. Sutanudjaja [f], Niko Wanders [f], Eleanor Blyth [a]

[a] UK Centre for Ecology & Hydrology, Wallingford, UK
[b] Meteorological Surveillance and Forecasting Group, DT Galicia, Agencia Estatal de Meteorología (AEMET), A Coruña, Spain
[c] Department of Computational Hydrosystems, Helmholtz-Centre for Environmental Research (UFZ), Permoserstraße 15, 04318 Leipzig, Germany
[d] University of Potsdam, Institute of Environmental Science and Geography, Am Neuen Palais 10, 14469 Potsdam, Germany
[e] Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol 16500, Czech Republic
[f] Department of Physical Geography, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

This manuscript was handled by Marco Borga, Editor-in-Chief, with the assistance of Yadu Pokhrel, Associate Editor.

Keywords:
Global hydrological forecasts
Hydrological models
ULYSSES
HydroSOS
Bias-correction
Multi-model blending

ABSTRACT

There is an immediate need to develop accurate and reliable global hydrological forecasts in light of the future vulnerability to hydrological hazards and water scarcity under a changing climate. As a part of the World Meteorological Organization's (WMO) Global Hydrological Status and Outlook System (HydroSOS) initiative, we investigated different approaches to blending multi-model simulations for developing holistic operational global forecasts. The ULYSSES (mULti-model hYdrological SeaSonal prEdictionS system) dataset, to be published as "Global seasonal forecasts and reforecasts of river discharge and related hydrological variables ensemble from four state-of-the-art land surface and hydrological models" is used in this study. The first step for improving these forecasts is to investigate ways to improve the model simulations, as global models are not calibrated for local conditions. The analysis was performed over 119 different catchments worldwide for the baseline period of 1981–2019 for three variables: evapotranspiration, surface soil moisture and streamflow. This study evaluated blending approaches with a performance metric based (weighted) averaging of the multi-model simulations, using the catchment's Kling-Gupta Efficiency (KGE) for the variable to define the weight. Hydrological model simulations were also bias-corrected to improve the multi-model blending output. Weighted blending in conjunction with bias-correction provided the best improvement in performance for the catchments investigated. Applying modelled weights during blending original simulations improved performance over ungauged catchments. The results indicate that there is potential to successfully and easily implement the bias-corrected weighted blending approach to improve operational forecasts globally. This work can be used to improve water resources management and hydrological hazard mitigation, especially in data-sparse regions.

1. Introduction

Freshwater supply is fundamental to human existence; however droughts can be devastating (e.g., Wilhite et al., 2007), and floods very destructive (e.g., Bubeck et al., 2017). Considerable advances in rainfall predictions, at daily to seasonal timescales, in recent years do not always solve the problem of predicting floods (e.g., Kobold and Sušelj, 2005) and droughts (e.g., Ali et al., 2018) due to propagating uncertainties, associated with hydrology and land processes. Any land surface and hydrological model, dynamical or empirical or data-driven, needs information about the properties of the geology, vegetation, soil types etc., which may not be easily available over data-sparse regions. Thus, large-scale global hydrology and land surface models have uncertainties associated with model parameters derived from global datasets of soil properties, topography and vegetation cover (Sood and Smakhtin, 2015). Despite this, Schellekens et al. (2017) show that using the average of a suite of such models can perform as a reasonable proxy for a locally calibrated model in some circumstances. This gives us hope that we could provide reasonable regional forecasts of hydrological conditions in data-sparse regions.

* Corresponding author.
E-mail address: amuche@ceh.ac.uk (A. Chevuturi).

https://doi.org/10.1016/j.jhydrol.2023.129607
Received 2 August 2022; Received in revised form 6 March 2023; Accepted 27 April 2023
Available online 3 May 2023

The World Meteorological Organization's (WMO) Global Hydrological Status and Outlook System (HydroSOS) initiative, approved for implementation at WMO Congress (Resolution 25 Cg-18), aims to utilise global scale modelled products from multiple sources to provide current global hydrological status and sub-seasonal to seasonal outlooks (Jenkins et al., 2020). The HydroSOS initiative builds on existing knowledge and data to deliver monthly global hydrological forecasts that combine model output with local, national and regional scale data so that the final product is locally informed (WMO, 2021). This is achievable, as new sources of observations and state-of-the-art models are closing the knowledge gap in the understanding of the water cycle and its forecasting (Lahoz and De Lannoy, 2014; Lavers et al., 2020). Multi-model hydrological forecasts have been shown to provide reliable predictions for different regions (Velazquez et al., 2011; Wanders and Wood, 2016). As blending forecasts from multiple models has been successful regionally (e.g., Ajami et al., 2006), it can be applied globally to provide holistic hydrological predictions. However, it is critical that any new hydrological forecast's blending is tested in a scientifically justifiable and operationally implementable way.

Different multi-model blending techniques have long been used to improve forecasts by leveraging the skill of certain models while excluding the errors in others (e.g., Shamseldin et al., 1997; Roy et al., 2020). The simplest blending method uses an average of all model members, which can include model simulation standardisation, to remove forecast bias (Georgakakos et al., 2004; Ajami et al., 2006). However, the simple average blending does not exploit enhanced skill of certain models selectively, which can be implemented using weighted averaging methods by assigning weights to the model members (e.g., Diks and Vrugt, 2010). Weights can be estimated based on different methods e.g., multiple linear regression (Wanders and Wood, 2016), sub-ensemble selection methods (Thober and Samaniego, 2014; Thober et al., 2015), constrained least squares technique (Ajami et al., 2006), using machine learning methods (Zaherpour et al., 2019), or Bayesian model averaging methods (Darbandsari and Coulibaly, 2019), which reward the skillful models and penalise the less skillful ones (Arsenault et al., 2015). However, some of these methods can become computationally intensive and time consuming when applied at a global scale (Neuman, 2003; Jozaghi et al., 2021; Wang et al., 2023). Studies can also assign weights to the models directly based on their skill metrics such as in Arsenault et al. (2015), which can be less computationally intensive methods but may not be optimal over ungauged basins (Arsenault and Brissette, 2016). Further, although weighted averaging outperforms individual models most of the time (e.g., Abrahart and See, 2002; Duan et al., 2007), sometimes the best performing individual model performs better (e.g., Arsenault and Brissette, 2016). Thus, studies have shown that it is imperative to rigorously investigate any new modelling system's post-processed output performance for generalised global applications (Duan et al., 2007).

Our ultimate goal is to develop a methodology to improve global hydrological predictions which can be best achieved by using global multi-model forecast products. Such forecast products can help us deliver holistic (i.e. forecasts for different variables over the same catchments) and skillful predictions while sending a simplified message to the users through multi-model blending (Roy et al., 2020). Given the global nature of our end product, it is essential that the blending method used should be computationally non-intensive while being sophisticated enough to exploit the skill of models. Considering the computational cost has a three-fold objective: the economic cost, time-efficiency and carbon footprint, all are especially important if these methods are to be applied in an operational set-up. The latter (carbon footprint of intensive computational science) is a growing concern in the scientific community in an era where international efforts are focused in achieving Net Zero, as the energy needed to power intensive computing is the main source of green-house gas emissions from scientific activities (Eichhorn et al., 2022; Lannelongue and Inouye, 2023). Since the model blending is a post-processing step of an already computationally expensive product (ensemble of global distributed hydrological simulations), it is paramount that the additional carbon footprint of this step is kept as minimal as possible, while still providing added value to the end product. Further, bias-correcting the simulations using simple methods before blending the multi-model output has shown to significantly improve performance for the forecasts (e.g., Dion et al., 2021). Thus, in this study we analyse the possibility of using the performance metric of the model as a weight for blending a global multi-model output in a simple and computationally inexpensive way, tested using the baseline simulations, and also investigate the improvements by bias-correcting the simulations before blending. We test our blending methods on ungauged catchments by modelling the performance of the models over catchments without observations for three different hydrological variables: streamflow, soil moisture and evaporation. We compare our blending method, which uses model performance metrics as weights, against the arithmetic averaging method, which was used as the benchmark. These methods investigated in the study provide a promising future for global water resource management and flood forecasting, especially over data-sparse regions.

The article is organised as follows: Section 2 introduces the data; the methods used are provided in Section 3; the results are described in Section 4; the recommendations from our results (Section 5.1), limitations of our study (Section 5.2) and future avenues of research (Section 5.3) are discussed in Section 5; and summary and conclusions are given in Section 6.

2. Data

For this study, we used 119 sample catchments across the world (Fig. 1). All of the study catchments have Global Runoff Data Centre (GRDC; BfG, 2020) daily streamflow data available, along with catchment characteristics (Table S1). We wanted to evaluate the performance of more than 100 basins, with extensive spatial coverage across the world and large hydroclimatic heterogeneity to be able to investigate the performance of ULYSSES forecasts across the globe for different types of catchments. Thus, the 119 catchments were chosen on the following criteria: (i) catchments are larger than 5000 km^2; (ii) distributed in most hydroclimatic zones (based on the Budyko analysis); (iii) manual quality check of observations did not reveal significant errors (abrupt changes or constant values); (iv) at least 5 years of GRDC observed data were available on record between 1981–2019; (v) subjective selection of catchments to be spread across the globe (Samaniego et al., 2020). Setting a larger threshold for observation years for catchment selection would have led to fewer catchments available over Africa, Asia and South America. Our selection criteria left us with 7 catchments in Africa, 16 in Asia, 34 in South America, 35 in Northern America, 10 in Australia and 17 in Europe (Fig. 1).

We used output from the ULYSSES contract (mULti-model hYdrological SeaSonal prEdictionS system; Samaniego et al., 2020; UFZ, 2020) in this study, which will be available at Copernicus Climate Change Service (C3S) as global seasonal forecasts and reforecasts of river discharge and related hydrological variables ensemble from four state-of-the-art land surface and hydrological models. For brevity, we refer to this dataset as ULYSSES dataset. The four land surface and hydrological models are:

- *mesoscale Hydrologic Model* (mHM; Samaniego et al., 2010; Kumar et al., 2013) is a spatially explicit distributed hydrological model developed by Helmholtz Centre for Environmental Research.
- *PCR-GLOBWB* (PGB; Sutanudjaja et al., 2018) is a grid-based global hydrology and water resources model developed at Utrecht University.
- *Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land* (HTESSEL; Johnson et al., 2019) is the land-surface model of the coupled European Centre for Medium-Range Weather Forecasts system 5 seasonal forecasting model.

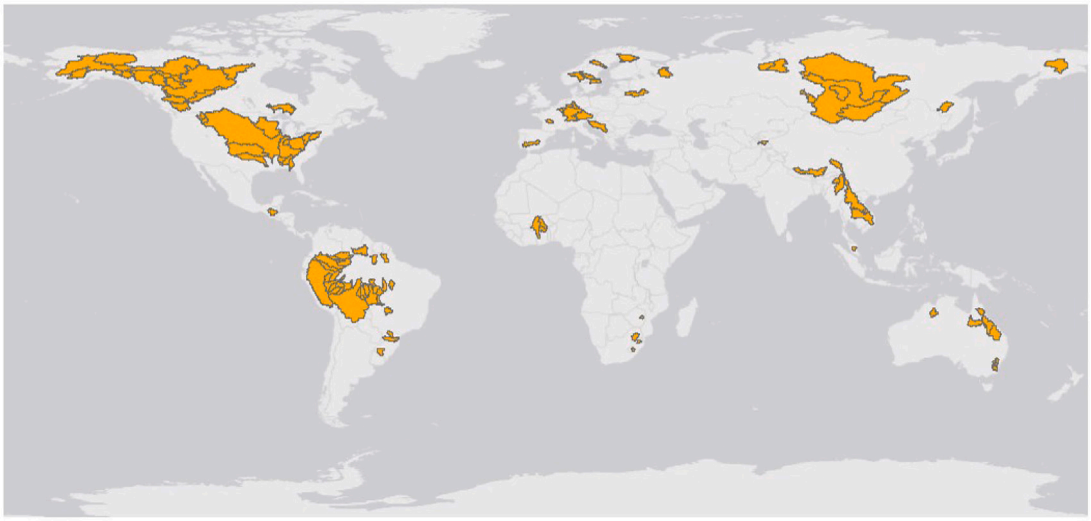


Fig. 1. 119 Catchments used in this study (shaded). Please see Table S1 for full details of each catchment.

- *Joint UK Land Environment Simulator* (JULES; Best et al., 2011; Clark et al., 2011) is a land-surface model developed by a wide community of researchers, coordinated by the UK Met Office and UK Centre for Ecology and Hydrology. It is used both as a standalone model and as the land surface component in the Met Office Unified Model.

ULYSSES uses the four hydrological models to output multiple hydrological variables as gridded output (e.g., runoff, evapotranspiration, soil moisture; Samaniego et al., 2020). The streamflow output of different rivers for all four models is derived through the multiscale routing model (mRM; Thober et al., 2019). In this study we evaluate blending of three output variables: evapotranspiration (ET), surface soil moisture (SM), and streamflow (SF) for the modelled baseline period of 1981–2019. We use monthly-mean values for the three variables extracted for the 119 catchments. SF is extracted at the catchment outflow point on the river network for each of the 119 catchments to compare against the observations (Figure S1). ET and SM are extracted as area-averaged values for the same 119 catchments from gridded output (Figures S2; S3). We derive the catchment extent from the digital elevation information from Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010; Danielson and Gesch, 2011) for each catchment.

We validate our blended output against observations for the three variables. For ET, we use Global Land Evaporation Amsterdam Model (GLEAM) version 3.5a actual evaporation data (Figure S2a; Martens et al., 2017). For SM, we use the European Space Agency (ESA) Climate Change Initiative (CCI) volumetric soil moisture product version 02.2 (Figure S3a; Liu et al., 2011; Dorigo et al., 2017). For SF, we use GRDC observed SF from The Global Runoff Data Centre, D-56002 Koblenz, Germany (Figure S1; GRDC-WMO, 2021). Using these observations leads to some missing observations for SF (Figure S1) and some catchments having no observed data for SM (Figure S3a), implications of which is discussed in Section 5.2. However, we evaluate all the three variables for the same 119 catchments, as we would like to analyse if the ULYSSES output can provide skillful yet holistic hydrological forecasts over different regions of the world. Further, as we are evaluating the performance of the models at catchment-scale, we need to have observations representing the whole catchment, which are usually generated through remote sensing or modelling network of in-situ observations for ET and SM, rather than using station (or point) observations, which may not be able to represent intra-catchment variability. Thus, this method of analysis allows us to investigate if ULYSSES can provide hydrologically consistent global forecasts.

3. Methodology

To provide a skillful blended hydrological products using the ULYSSES simulations, we evaluate different blended approaches applied on native (original) and bias-corrected model variables at a monthly time step for all the 119 catchments for the baseline period 1981–2019. We also evaluate the application of our methods on daily vs. monthly SF values. We further investigate the application of blending approaches on ungauged stations. Fig. 2 summarises the methods used in this study as a flow chart. The following subsections describe these approaches in detail.

3.1. Blending approaches

We use the following two blending approaches, to generate blended output for native and bias-corrected simulations.

1. **Arithmetic average:** The simplest form of blending multi-model output is an average of the four model variables without any weighting, as used in Shamseldin et al. (1997), Arsenault et al. (2015), and will be referred to as the "arithmetic blending" approach. Fig. 3 shows an example of the observed and native simulated variables along with the arithmetic blended versions, which is referred to as *native arithmetic* output. We also blend the bias-corrected variables (discussed in Section 3.2) and the blended output is referred to as *bias-corrected arithmetic* output (Figs. 2; 3). The performance of the arithmetic blended output is used as a benchmark to compare other blending approaches in this study.
2. **Weighted average:** The weighted average method uses weights based on model performance, and from here on is referred to as "weighted blending" approach. Model performance at each catchment was assessed for monthly-mean variables (ET, SM, SF) from all four models over the baseline period, using the Kling-Gupta Efficiency (KGE; Gupta et al., 2009). The KGE metric evaluates the model performance for contribution of mean, variance and correlation (see Section 3.4 for more detail). For this approach, the KGE metric for each model and variable is identified as their respective weights, and the blended output is calculated using weighted averaging (*native weighted* and *bias-corrected weighted*). Please note that each catchment has its own set of weights. Here we apply weighted arithmetic mean, in which sum of the catchment values multiplied with their respective weights is divided by the sum of weights. We only use

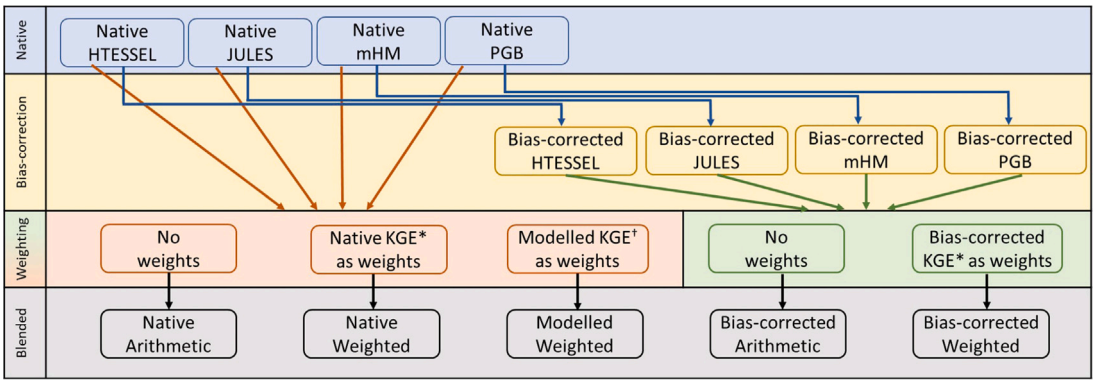


Fig. 2. Schematic showing the different blending approaches, based on different types of weighting, used for native (original) simulations from ULYSSES (blue –> red) and bias-corrected product (yellow –> green), to generate five different blended products (black). * denotes that we weighted the native and bias-corrected simulations using each variable's own KGE (native weighted and bias-corrected weighted) but also used SF KGE as weights for ET and SM blending (native SF-weighted and bias-corrected SF-weighted). † modelled KGE as weights are only applied for SF native simulations.

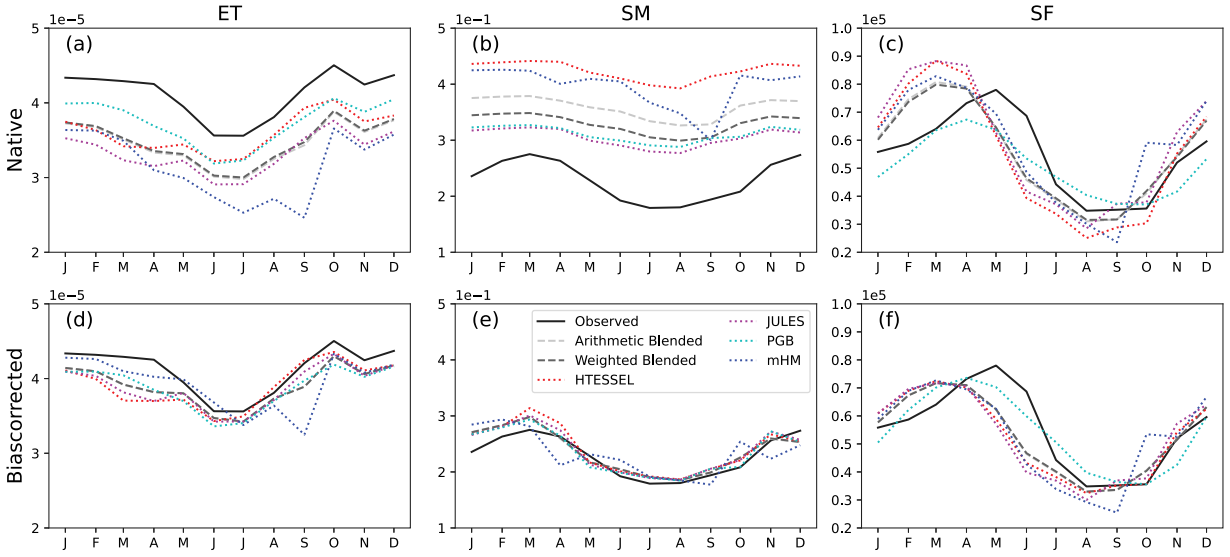


Fig. 3. Monthly time series for 1993 of observed (black solid line), individual models (coloured dotted lines) and blended output (grey dashed lines) for (a–c) native and (d–f) bias-corrected simulations of (a, d) ET ($\times 10^{-5}$ kg m^{-2} s^{-1}), (b, e) SM ($\times 10^{-1}$ m^3 m^{-3}), (c, f) SF ($\times 10^5$ m^3 s^{-1}) over an example catchment for Sao Paulo De Olivenca, Amazon River, Brazil (ID = 3623100).

positive KGE values as weights; for any model with a negative KGE metric the weight is set to zero, and thus ignored. For any catchment and variable with all models showing negative KGE metric, we use arithmetic blending approach instead i.e. all model weights are set to one. This only occurs for 25 catchments for SM, 1 catchment for SF and no catchments for ET for native simulations and only 8 catchments for SM bias-corrected simulations. Fig. 3 shows examples of the native and bias-corrected variables for each model along with their weighted blended counterparts, referred to as native weighted and bias-corrected weighted output (Fig. 2). We also test another version of the original "weighted blending" approach, in which we use the KGE metric calculated for SF, and apply it to the ET and SM simulations. We will refer to this method as "SF-weighted blending" (*native SF-weighted* and *bias-corrected SF-weighted* outputs for ET and SM). This approach is intended to derive blended hydrological products that still maintain the catchment hydrological balance, which for some applications can be more important than getting the "best" possible estimate for all three variables.

3.2. Bias-correction

The bias-correction methodology tested here is based on Farmer et al. (2018), and is applied by Sanchez Lozano et al. (2021) to operationally bias-correct GEO Global Water Sustainability (GEOGloWS) SF forecasts. Although Farmer et al. (2018) recommend 14 complete years of observed data to be able to accurately calculate the percentiles, the method is applied here to all stations, regardless of the length of record (shortest record is of 5 years). However, the median length observations is 29 years over all study catchments (compared to the full period of 39 years between 1981–2019) and out of 119 catchments, 98 have more than 14 years of observations. Please note we have not performed bias-correction for SM over catchments with fully missing data.

For bias-correction of SF model output, we first calculate a Flow Duration Curve (FDC) for all data (for daily data this is done for each month separately) in observed and simulated time series. Fig. 4 shows an example of the FDCs for an example catchment for the model HTESSEL. Using the FDCs, the non-exceedance probability of every simulated value can be estimated. The observed SF value corresponding to that non-exceedance probability can be deduced. The simulated value is then converted by replacing it with the equivalent observed

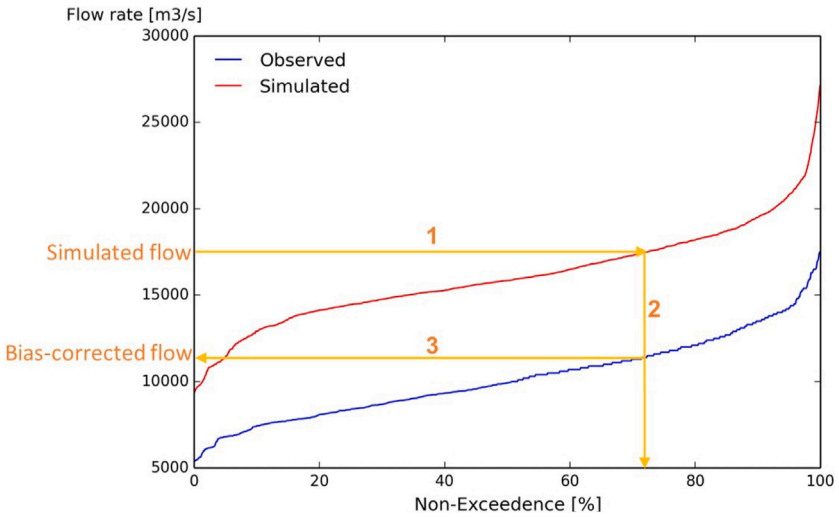


Fig. 4. Percentage of non-exceedance against the SF rate for observed (blue) and native simulated (red) SF for HTESSEL over an example catchment for Sao Paulo De Olivenca, Amazon River, Brazil (ID = 3623100). The yellow arrows show the steps involved in the bias-correction process: (1) For a given native simulated flow, the point on the simulated FDC (red line) is identified; (2) the non-exceedence corresponding to that simulated flow is determined, and (3) the observed flow for that same non-exceedence is determined from the observed FDC (blue line), and this value corresponds to the bias-corrected flow.

SF to the same non-exceedance probability (i.e. quantile mapping method). This will be referred to as the bias-corrected SF. KGE metrics are calculated for all the bias-corrected SF simulations from the four models, and arithmetic and weighted blending approaches are applied to get bias-corrected arithmetic and bias-corrected weighted blended SF time series. For bias-correction of ET and SM, the same approach is applied replacing SF in the FDC by the variable in question. Please note that for the daily SF, bias-correction was performed for each month separately, whereas for the monthly data (SF, ET and SM), the time series was not split into months, as the records were not long enough.

3.3. Modelling KGE weights at ungauged basins

To apply the weighted average blending technique (Section 3.1) at ungauged basins, the KGE metric needs to be modelled, which can then be used as weights for each model at the ungauged basins for blending. In this study we refer to the ungauged basins where the observed data is hidden for testing the method over regions with no observations. This method is only tested for SF, as other hydrological variables (ET and SM) have observations from remote measurements available all over the globe in the form of gridded data and thus do not have "ungauged" catchments.

Catchment characteristics extracted for all 119 study catchments from the HydroATLAS dataset (Linke et al., 2019) were used to build the KGE statistical model (KGE-stats-model) to estimate SF KGE at "ungauged" catchments using statistical approaches. From the full list of 56 catchments attributes available in HydroATLAS, 17 attributes most relevant to hydrology and land surface modelling were retained (Table 1). For a full description of the variables, please see the HydroATLAS RiverATLAS catalogue (Lehner, 2019).

The number of variables were then further reduced by analysing the cross-correlation between the catchment attributes and the hydrological models performance (KGE metric). Cross-correlation between the 17 variables allows for the identification of a subset of 9 variables with high correlations. For the modelling of KGE at ungauged catchments, we tested different machine learning methods: Multiple Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression. For each of these six methods, we conducted two combinations of experiments: (i) with all 17 variables, and with a subset of 9 variables (selected after cross-correlation analysis) as input predictors; (ii) with Principal Component Analysis (PCA) to reduce dimensionality, and without PCA.

In addition to the Regression methods, an Artificial Neural Network (ANN) with one hidden layer was also built. The ANN was tested with a range of different number of neurons in the hidden layer, different batch sizes and number of epochs. All the regression and ANN models were built separately for each of the four hydrological models, and assessed individually to identify the best KGE-stats-model. This means that each hydrological model has a different statistical model to estimate KGE. Therefore, the best KGE-stats-model for HTESSEL is not necessarily the same as the one for JULES, for example.

To assess the performance of each KGE-stats-model, their R2 score (coefficient of determination) was compared using the Leave-One-Out Cross-Validation (LOOCV) method (Sammut and Webb, 2010). LOOCV consists of leaving one observation out from the dataset to build the model, and using the observation left out to assess the model. This process is repeated for every observation available in the dataset, where each time a different observation is left out. The best KGE-stats-model based on this assessment was selected for each of the four hydrological models, and were used to produce a set of estimated KGE metric for all four models at all 119 case study catchments. We expand on the modelling of KGE metrics in Section 4.3.

The blended output is calculated after applying the new modelled KGE metrics as weights to the native SF simulations (referred to as *native modelled weighted*). This "modelled weighted blending" approach, identifies how well the weighted blended approach for SF performs over ungauged stations. KGE of the resulting modelled weighted blended product is calculated to quantify and compare its performance with the original weighted approach.

3.4. Validation

Using the methods described in Section 3.1–3.3 we develop different blended products (Fig. 2) for the three variables (ET, SM, SF) over the baseline period (1981–2019) for the 119 catchments:

1. Native arithmetic blended output
2. Native weighted blended output (along with native SF-weighted blended output for ET and SM)
3. Native modelled weighted blended output (only for native SF simulations)
4. Bias-corrected arithmetic blended output

Table 1
17 Catchment characteristics used in this study, extracted from the HydroATLAS dataset.

Variable full name	Abbreviation used	Source data	Reference
Natural Discharge	discharge	WaterGAP v2.2	Döll et al. (2003)
Land Surface Runoff	runoff	WaterGAP v2.2	Döll et al. (2003)
Lake Volume	lake	HydroLAKES	Messager et al. (2016)
Reservoir Volume	reservoir	GRanD v1.1	Lehner et al. (2011)
Degree of Regulation	degree_regulation	HydroSHEDS & GRanD	Lehner et al. (2011)
Catchment Area	area	HydroSHEDS & WaterGAP	Lehner and Grill (2013)
Terrain Slope	slope	EarthEnv-DEM90	Robinson et al. (2014)
Climate Zones	climate_zone	GEnS	Metzger et al. (2013)
Precipitation	precip	WorldClim v1.4	Hijmans et al. (2005)
Potential Evaporation	pet	Global-PET	Zomer et al. (2008)
Snow Cover Extent	snow	MODIS/Aqua	Hall and Riggs (2016)
Forest Cover Extent	forest	GLC2000	Bartholome and Belward (2005)
Cropland Extent	crop	EarthStat	Ramankutty et al. (2008)
Pasture Extent	pasture	EarthStat	Ramankutty et al. (2008)
Irrigated Area Extent (Equipped)	irrigation	HID v1.0	Siebert et al. (2015)
Soil Water Content	soil_water_content	Global Soil-Water Balance	Trabucco and Zomer (2010)
Human Development Index	human_dev_index	HDI v2	Kummu et al. (2018)

5. Bias-corrected weighted blended output (along with bias-corrected SF-weighted blended output for ET and SM)

The respective observed datasets are used to validate the native and bias-corrected model simulations and the blended outputs. Validation can only be carried out for the period over which the observed data is available, and thus, model simulations and blended output (for SM and SF) are masked over the period where the corresponding observed data is missing. Please note that some catchments are missing for SM due to missing observations for the whole study period (see Figure S3a; Section 2).

We verify the bias, mean and variance of the model simulations and blended output using the KGE and Nash–SutcliffeEfficiency (NSE; Nash and Sutcliffe, 1970) metrics calculated at monthly time step.

$$KGE = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{sim}}{\sigma_{obs}} - 1)^2 + (\frac{\mu_{sim}}{\mu_{obs}} - 1)^2} \quad (1)$$

where r is the linear correlation coefficient between observations and model simulations, σ_{sim} and σ_{obs} are standard deviations of the simulations and observations respectively, and μ_{sim} and μ_{obs} are means of the simulations and observations respectively. KGE equal to 1 indicates a perfect agreement between the simulations and observed; KGE less than 0 indicates poor model performance Gupta et al. (2009).

$$NSE = 1 - \frac{\sum_{t=1}^{t=T}(V_{sim}(t) - V_{obs}(t))^2}{\sum_{t=1}^{t=T}(V_{obs}(t) - \mu_{obs})^2} \quad (2)$$

where T is the total number of time steps, $V_{sim}(t)$ is the model simulation at time t, $V_{obs}(t)$ is the observations at time t, and μ_{obs} is the mean observed variable. NSE equal to 1 indicates perfect correspondence between model and observations; NSE equal to 0 indicates that the model simulations have similar performance as the mean of the observations; and NSE less than 0 indicates model simulation is worse than the observed mean.

The KGE metric is one of the most commonly used metric to evaluate model performance in the field of hydrology (Knoben et al., 2019) and has been extensively used for model calibration and evaluation (e.g., Siqueira et al., 2018; Sutanudjaja et al., 2018). For clarity, we would like to draw the reader's attention to the different uses made of the KGE metric throughout this study:

- KGE was used to define the *"weights"* assigned to each of the four models for the weighted blending approach. The KGE metric is used instead of NSE metric, because KGE improves upon some of the drawbacks of NSE, associated with underestimating runoff variability, by better representing the constitutive components: correlation, variability bias and mean bias (Gupta et al., 2009). For native and bias-corrected simulations respective KGE values have been calculated for each catchment and each model. As each variable will have its own set of weights, the water balance of the blended product will not be maintained. Therefore, as a separate experiment, KGE calculated for SF is also applied to ET and SM (SF-weighted blending), so that the same set of weights is used for all variables, which maintains catchment water balance within each model. KGE for SF was also modelled using statistical approaches, to be used as weights for the native simulations to produce the modelled weighted blending to be tested for ungauged catchments.
- KGE was used as one of the *"validation"* metrics, to compare the performance of the different individual models and blending approaches for each variable (ET, SM and SF) and for native and bias-corrected simulations.

This study also compares the baseline model simulations and blended output as categorical forecasts, divided into 5 categories: low (0%–10%), below normal (10%–33%), normal (33%–67%), above normal (67%–90%) and high (90%–100%). These categories can be combined to consider the outputs simply as equal terciles, but the five categories provide the likelihood of more extreme events better. The categorical verification of simulation allows for the evaluation of forecast skill for specific categories rather than for models identifying the exact value of the variable. Simulated, bias-corrected and blended outputs' categorical skill is measured using Accuracy (ACC; Wilks, 2011) and Heidke skill score (HSS; Heidke, 1926). These skill scores represent absolute categorical skill (ACC) of the output and skill relative to that of a random chance (HSS).

$$ACC = \frac{1}{N}\sum_{i=1}^{C} n(S_i, O_i) \quad (3)$$

where N is the total number of simulations, C is the number of categories (5 for this study) and $n(S_i, O_i)$ are the number of times that the model simulates correctly for the category i. ACC measures how many times the model hits the correct category, but it can be influenced by the most common category. ACC has range of 0 to 1 with 1 for a perfect simulation.

$$HSS = \frac{\frac{1}{N}\sum_{i=1}^{C} n(S_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{C} n(S_i)n(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{C} n(S_i)n(O_i)} \quad (4)$$

where N is the total number of simulations, C is the number of categories (5 for this study), $n(S_i, O_i)$ are the number of times the model simulates correctly for the category and $N(S_i)N(O_i)$ is the product of the number of simulations and observed for a particular category i. HSS measures only the correct model simulations which occur beyond that of a random chance. HSS ranges from $-\infty$ to 1, 1 is the perfect score and anything less than 0 shows that correct simulations are due to random chance.

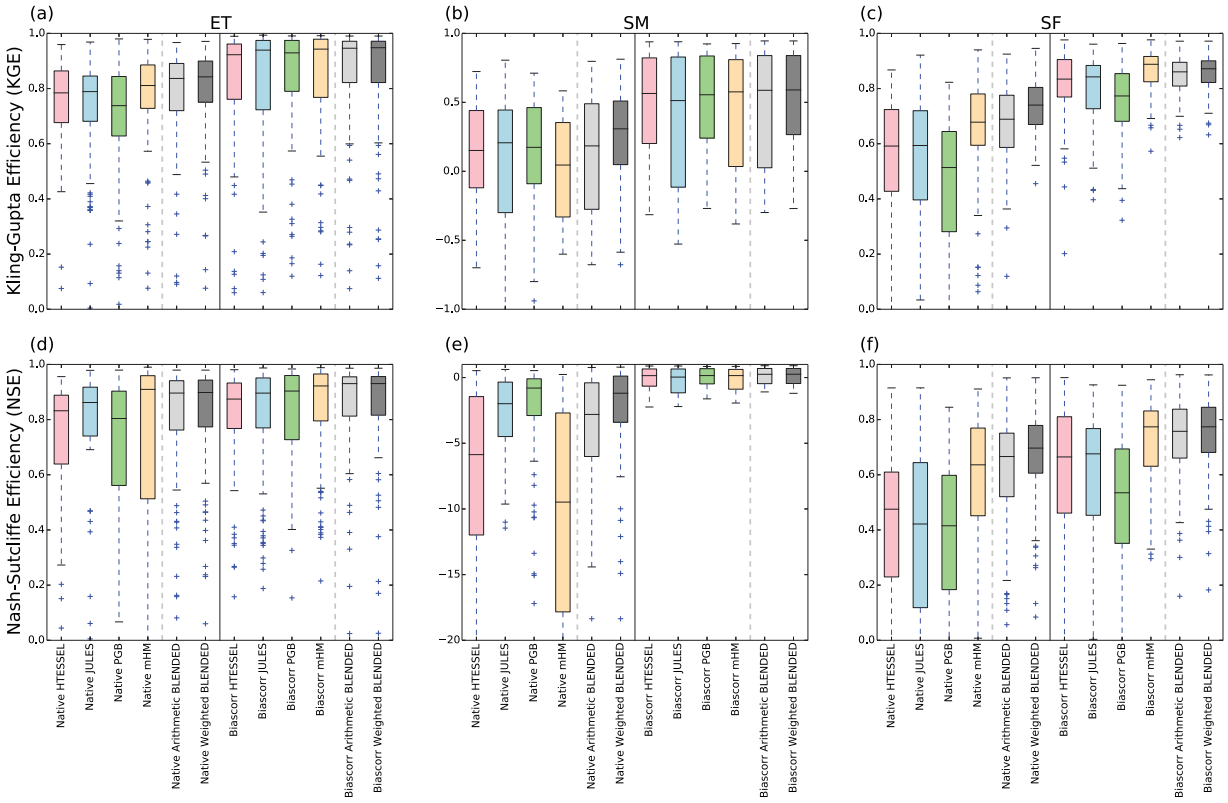


Fig. 5. KGE-monthly range over all 119 catchments (boxplots) for (a) ET, (b) SM and (c) SF shown for all 4 models simulations (the first four boxplots in colours), native arithmetic blended product (the fifth boxplot in light grey), native weighted blended product (the sixth boxplot in dark grey). The six boxplots on the right hand side are same as the left hand side ones showing the KGE for the bias-corrected model simulations (coloured boxplots), blended bias-corrected arithmetic blended product (in light grey), bias-corrected weighted blended product (in dark grey). Please note weights are derived from KGE metrics for each individual variable. (d–f) same as (a–c) but for NSE-monthly.

4. Results

4.1. Arithmetic and weighted blending

For each catchment, arithmetic average and weighted average blended output is calculated from all four model native and bias-corrected simulations for all three variables (Section 3.1). Fig. 3a–c shows an example catchment native model time series for all four models for the three variables and the native blended model outputs for 1993. For the Sao Paulo De olivenca River catchment all models underestimate ET (Fig. 3a), overestimate SM (Fig. 3b) and have their phase shifted for peak SF (Fig. 3c). Native PGB simulations perform best over this catchment. Blending native simulations reduces the errors, but the weighted blending only shows improvement over arithmetic blending for SM in this catchment. We note that different catchments and different years have very different errors and improvements with blending and bias-correction (not shown). Thus, we summarise our results using KGE metric (Figure S4; Fig. 5).

Comparison between the native simulations and weighted blended SF shows that the blended product out-performs the individual models for 75% of the catchments (Figure S4c; Fig. 5c), but there is only moderate improvement for ET and SM (Figure S4a, b; Fig. 5a, b). The native arithmetic blended has lower performance than the native weighted blended output for all three variables (Fig. 5). For ET and SF, mHM native simulations have similar performance to the native arithmetic blended output. mHM may be performing better as it uses multiscale parameter regionalisation (MPR) technique which can parameterise across basins and scales using nonlinear transfer functions and is shown to perform better than standard regionalisation (Samaniego et al., 2010).

SM has the largest spread in KGE metric compared to ET and SF, despite lower number of catchments verified (Figure S4b; Section 2). For SM, 15 catchments (Figure S4b) have no observations for the whole baseline period (Figure S3a). For all four models, KGE metric for SM is, in general, lower than for the other two variables suggesting that SM is particularly challenging to estimate from global models or from satellite sensors for observations. Please see Section 5.2 for more detailed discussion.

The same arithmetic and weighted blending methods are also used to blend the bias-corrected simulations using the KGE metric from the bias-corrected simulations (Fig. 3d–f). Modelled KGE metric is used for weighted averaging of native simulations to evaluate performance of the method over ungauged catchments for SF. These two points will be discussed in the subsequent subsections.

4.2. Bias-correction of variables

Each model simulation is bias-corrected using the method described in Section 3.2. As the ULYSSES global product cannot be calibrated to every location, the bias-correction further aligns the estimated streamflow to local conditions, and thus provides an added value for users applying the forecasts at local-scale. The new KGE metric from the bias-corrected model simulations is used as weights for blending the bias-corrected model simulations, to calculate the bias-corrected weighted blended output. Bias-corrected simulations are also simply averaged to get the bias-corrected arithmetic blended product. Fig. 3d–f shows an example of bias-corrected and blended output. Bias-correction (Fig. 3d, e) improves the native ET and SM simulations (Fig. 3a, b). For SF, bias-correction (Fig. 3f) improves the magnitude but does not correct the timing i.e. the phase shift seen in the native SF simulations (Fig. 3c).

As expected, the bias-corrected simulations for individual models outperform the native simulations for all three variables (Figure S4), showing that bias-correction improves model performance. Other studies also show the importance of bias-correction of hydrological

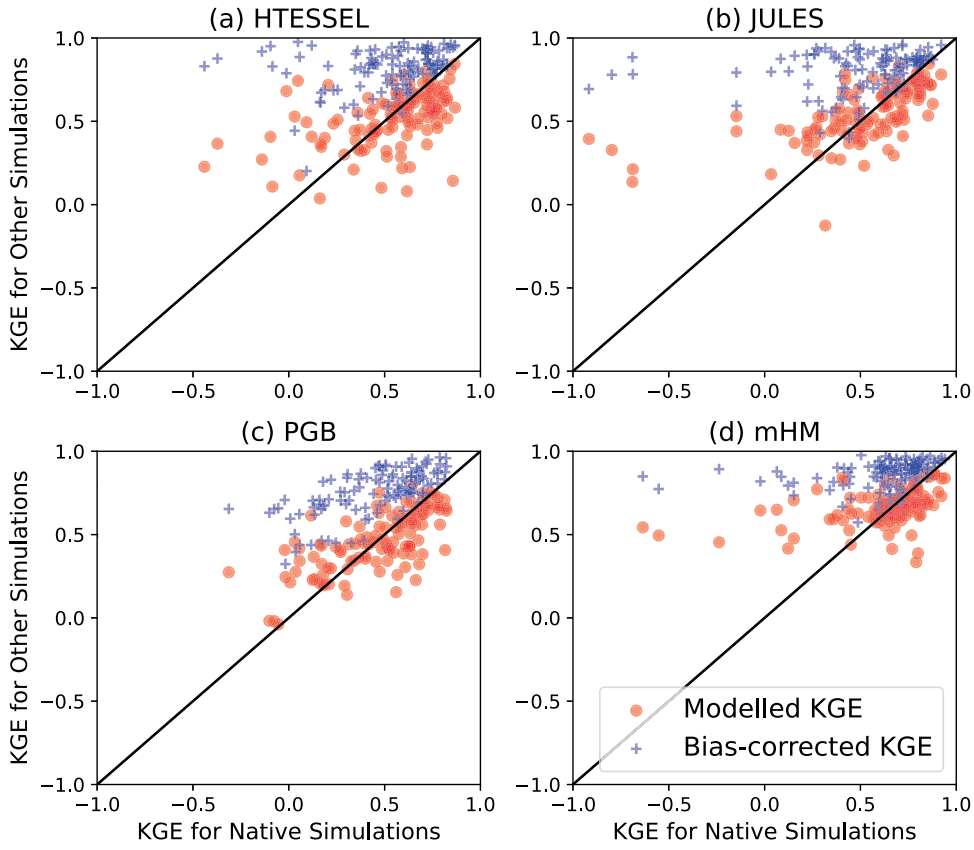


Fig. 6. Scatter plot of KGE metric between native SF simulations vs. bias-corrected SF simulations (blue crosses) and modelled SF KGE (red circles) for the model (a) HTESSEL, (b) JULES, (c) PGB and (d) mHM.

forecasts to improve prediction skill (e.g., Zalachori et al., 2012). The improvement is smaller with ET, where the native simulations already perform well (Fig. 5a), but the improvement is larger for SM and SF (Fig. 5b, c). Bias-correction almost always improves the performance of any model output, as shown for the SF simulations in Fig. 6. The bias-corrected (arithmetic or weighted) blended output outperforms native (arithmetic or weighted) blended output for almost all the catchments for the three variables (Figure S4; Fig. 5).

It should be noted that this bias-correction method for SF can only be performed over gauged catchments which have observed data available. Bias-correction of ungauged catchments is beyond the scope of this study, but bias-correcting simulations before producing a blended product is important for any operational implementation.

4.3. KGE estimation at ungauged stations

At ungauged catchments, where observed data is not available to calculate performance metrics, KGE-weights for SF need to be estimated in order to implement the weighted blending approach. We model SF KGE using the method described in Section 3.3. Cross-correlation analysis between the different catchment characteristics (Fig. 7a), and correlation of each of the catchment characteristics (Table 1) with the KGE of each of the four hydrological model variables (Fig. 7b) shows that some characteristics are highly correlated amongst themselves. Where two catchment characteristics were highly correlated, the one that was most correlated to the KGE of the hydrological models was retained. Thus, we select a subset of 9 catchment characteristics: discharge, lake, area, slope, precip, pet, snow, forest and soil_water_content (see Table 1 for the list of full names of catchment characteristics).

From (Fig. 7b), we can observe that in general, the correlations between the models' KGE and the various catchment characteristics are weak. One notable exception is the high correlation observed between JULES' KGE and Soil Water Content (positive correlation) and Pasture cover (negative correlation). These two catchment characteristics are highly correlated between them (negatively). This observation indicates that JULES struggles to simulate river flows for catchments where pastures are the dominating land cover. All models' performances are negatively correlated with the degree of regulation. PGB's performance increases with snow cover extent, which suggests the snow module of this model is more efficient than the other models. With the exception of HTESSEL, models are generally worse at predicting flow in regions with high evaporative demand (PET). HTESSEL is worse than the other models at simulating flows where lakes are present.

From the different KGE-stats-models tested (Section 3.3), the best performing model was selected based on the R2-score (coefficient of determination) for each of the four hydrological models (Table 2). PCA did not improve the R2-score for any of the KGE-stats-model tested and the ANN model did not outperform the "best" regression model in any of the cases. The sample size is believed to be too small to effectively train an ANN model. Note that the R2-score even with the best of the KGE-stat-models is relatively low, especially for mHM, which means that KGE is difficult to predict accurately from catchment characteristics alone.

Using the KGE-stats-model shown in Table 2, the modelled KGE metric for each of the 119 study catchments for all four hydrological models was derived (Fig. 6). For example, the Support Vector Regression method is applied to estimate KGE values (which we called the modelled KGE) of the PGB model for all catchments, as if the real KGE values for these catchments cannot be calculated due to the observations being unavailable. One modelled KGE value is derived for each catchment and each hydrological model. These modelled KGE values are then used for the weighted averaging of the four model simulations to derive the "native modelled weighted" blended output (Section 3.3).

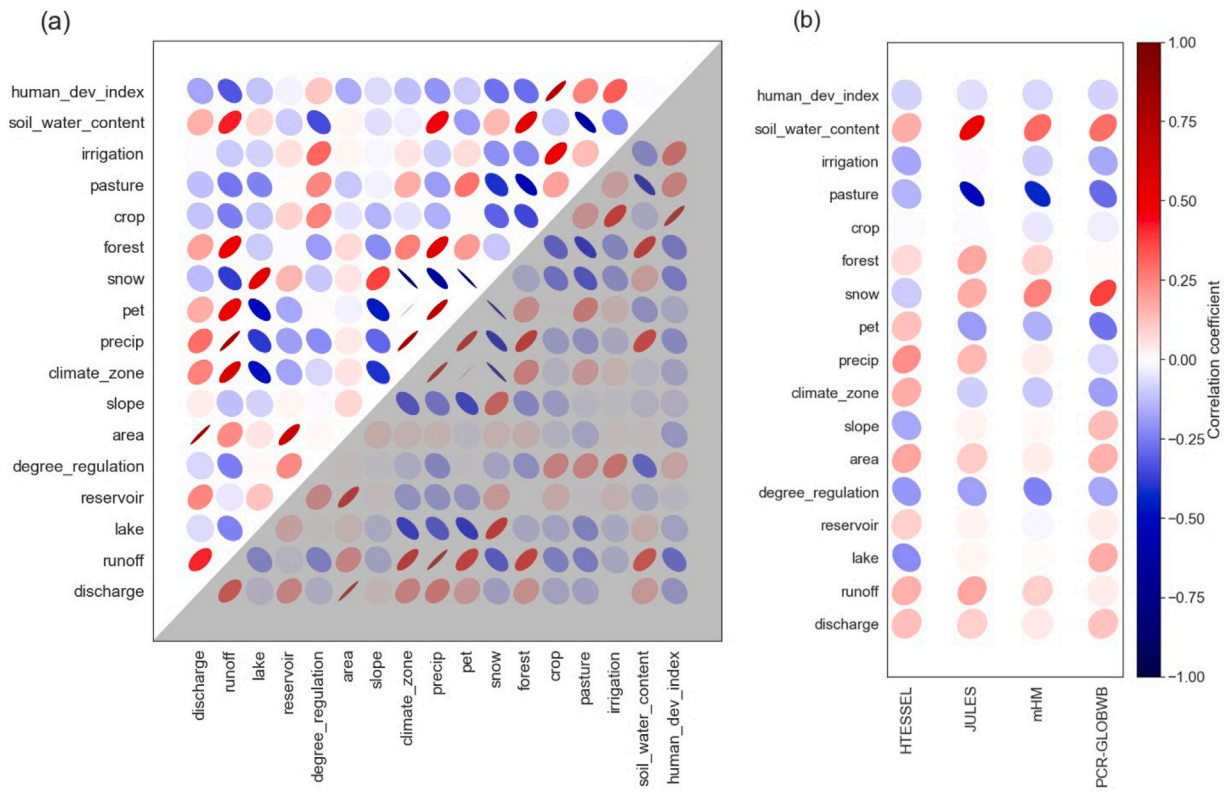


Fig. 7. (a) Correlation matrix of all 17 variables (discharge, runoff, lake, reservoir, degree_regulation, area, slope, climate_zone, precip, pet, snow, forest, crop, pasture, irrigation, soil_water_content and human_dev_index). (b) Correlation between the same 17 variables and SF KGE of the four hydrological models (HTESSEL, JULES, mHM, PGB). The higher the correlation, the "flatter" and darker the ellipse. Blue ellipses and tilted to the left indicate negative correlations, whereas red ellipses and tilted to the right indicate positive correlations.

Table 2
KGE-stats-model for SF selected for the four hydrological models, with their respective LOOCV R2-score (i.e. Leave-One-Out Cross-Validation method coefficient of determination).

Model	Selected KGE-stats-model	Variables used	LOOCV R2-score
HTESSEL	Random Forest Regression	Subset of 9 variables	0.184
JULES	Support Vector Regression	All variables	0.218
mHM	Support Vector Regression	All variables	0.148
PGB	Support Vector Regression	All variables	0.207

Over most catchments, the modelled KGE closely matches the native KGE for all four models, except for catchments with very low KGE values (red dots in Fig. 6). The statistical models estimate "improved" KGE in such catchments, which is not an accurate representation of hydrological model performance and thus not useful for blending methods over ungauged catchments. The errors in modelled KGE metric over such catchments may stem from failure in dynamically modelling such catchments (e.g., very dry basins are challenging to model), and thus it may be beyond the capability of these statistical models to estimate the KGE metric based on catchment characteristics alone. Further in-depth research may be required to develop these methods further.

These new modelled KGE metrics were used as weights for the native SF simulations to derive the modelled weighted blended product (see Section 3.3, 3.4). The modelled weighted blended product has similar performance as native weighted blended output over catchments where the KGE-stats-model estimates the KGE metric well, as expected, and different (poorer or better in some cases) performance over catchments where the KGE-stats-model estimates the KGE metric poorly (not shown). Modelled weighted simulations outperforms arithmetic blended in most cases, but not as much as weighted blending using observed KGE-weights (Fig. 8a).

4.4. Validation

From the methods discussed in above three subsections, a combination of blended products are derived: native arithmetic average, native weighted average (also alternatively native SF-weighted average for ET and SM), modelled weighted average (only for SF), bias-corrected arithmetic average and bias-corrected weighted average (also alternatively bias-corrected SF-weighted average for ET and SM). Fig. 5 shows that blending of any type improves KGE achieved compared with the native simulations from the four individual models, except for SM arithmetic blending. As discussed before, bias-correction improves individual model simulation (Fig. 6), although the degree of improvement is catchment and model specific (Figure S4). Blending bias-corrected simulations shows the highest improvement compared to all the other approaches. For example, only 6 out of the 119 catchments show poorer performance for the bias-corrected weighted blended SF compared to the native weighted blended SF (Figure S4). The comparison between bias-corrected weighted blended product and bias-corrected arithmetic blended product shows improvement for all three variables, albeit only marginal, except for SM (Figs. 5; 8a, c). For SM, the bias-corrected weighted blending performs much better than bias-corrected arithmetic blending, especially for stations where the catchments have a negative

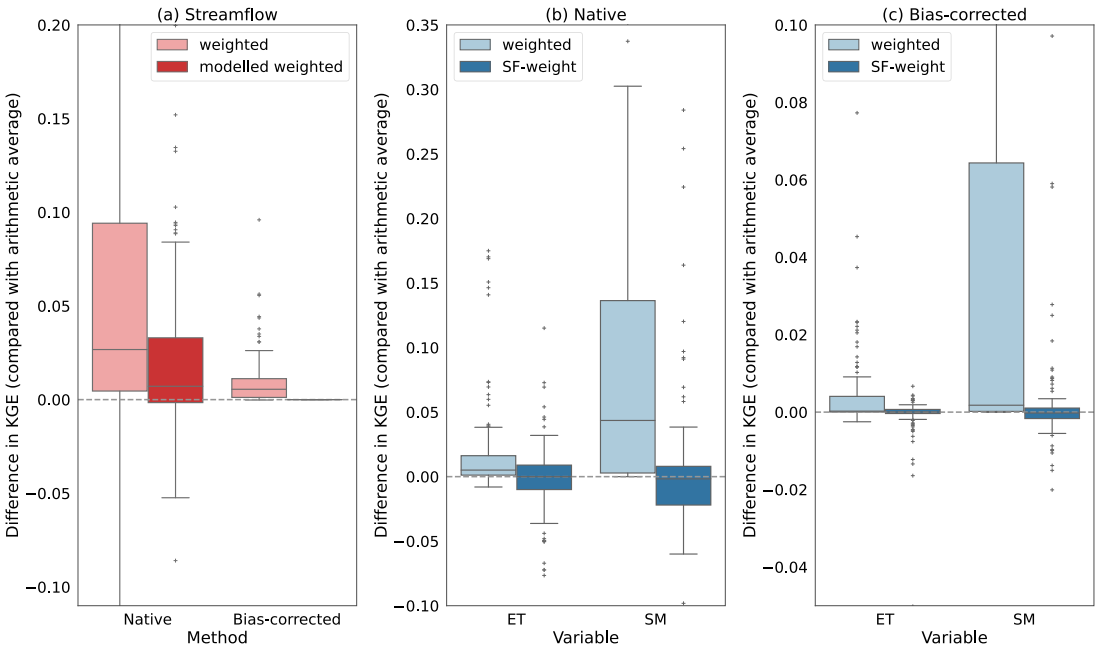


Fig. 8. (a) The difference between the KGE for the native and bias-corrected weighted blended output (light red boxplot) and the modelled weighted blended output (dark red boxplot) against the corresponding arithmetic blended output for the SF simulations. Please note for SF, there is no boxplot showing the difference between bias-corrected modelled weighted output and arithmetic weighted bias-corrected output. Difference between the KGE for the weighted and arithmetic blended product (light blue boxplot) and between the KGE of SF-weighted and arithmetic blended product (dark blue boxplot) for the ET and SM variables with (b) native and (c) bias-corrected simulations.

KGE for native simulations (not shown). Further, for the blending of the native simulations, performance of the blended output is improved by the use of KGE weights (Figs. 5; 8b). Thus, weighted blending is useful with both native and bias corrected simulations.

As mentioned in Section 3.1, we also test the original weighted blending method against the SF-weighted for ET and SM, and we show the differences of these two weighted blending approaches against arithmetic blending of native and bias-corrected simulations in Fig. 8b, c. For both native and bias-corrected simulations, weighted blended outperforms the SF-weighted blended for both ET and SM. In fact, the SF-weighted blending mostly performs similarly or slightly worse than the arithmetic blending. The performance difference is larger in SM than in ET (Fig. 8b, c), as the KGE spread in SM simulations is larger (Fig. 5b).

We also test the weighted blending approaches over "ungauged" stations for SF by modelling KGE using catchment characteristics (modelled weighted blending; Section 4.3). For the native simulations, modelled weighted performs slightly worse than weighted blending, but is still overall better than arithmetic blending (Fig. 8a). Thus, overall the KGE-stats-model was capable of estimating the SF KGE for "ungauged" catchments to a certain degree, but as mentioned in Section 3.3, KGE is not solely attributable to catchment characteristics. Please note that we have not compared weighted and modelled weighted blending for bias-corrected SF simulations (Fig. 8a), as this was beyond the scope of this study.

The performance of models' native and blended simulations is also measured with another performance metric, NSE (Fig. 5d–f). Performance with NSE shows similar results to performance with KGE: blended products show higher performance compared to individual models (except for mHM in ET and SF); bias-correction improves model performance (more in SM and SF than ET); and KGE based weights improve performance (more when applied to native simulations than bias-corrected simulations). We also compare the application of our methods on daily vs. monthly SF simulations using KGE and NSE performance metrics (Figure S5). Applying our methods to daily simulations shows similar improvement in performance as compared to methods applied to monthly simulations, and the overall conclusions for blending and bias-correction methods' performance remain the same.

Categorical forecast skill, calculated for five categories (see Section 3.4), is measured using ACC and HSS skill scores. Compared to the performance metrics (KGE and NSE), using skill scores (ACC and HSS) show smaller improvements after blending (Fig. 9). HSS for ET and SF are positive, which indicates that the accuracy of getting each category right is better than that from a random chance (Fig. 9d, f), but SM has some negative HSS values (Fig. 9e). There is no improvement in performance with bias-correction, which is to be expected as the bias-correction will not modify the relative values of the variables. There is improvement in the native (arithmetic and weighted) blended products as compared to individual models in SF, but this improvement is not very large in ET and SM. PGB performs the worst amongst the individual models with both skill scores for the native SF simulations; mHM outperforms the other models for SF and has similar skill to the blended products (Fig. 9c, f). This may be due to the fact that mHM uses the MPR technique to parameterise across basins and scales (see Section 4.1). Further, it should be noted that there are large gaps in the observed data for some stations with SF (Section 2), which shortens the length of the time period of evaluation, and may lead to poorer categorical skill for SF with blended products for certain stations.

5. Discussion

In this study we aimed to identify simple blending approaches and thus their feasibility for implementation in operational forecasts. However, this is a pilot study and has only been conducted for baseline simulations for just three variables and 119 catchments. Thus, more future work is required before implementation of these methods operationally. In this section, we discuss the practical implications of our study that need to be carefully considered before implementation.

5.1. Recommendations

Our results show that arithmetic blending rarely improved performance over that of the best performing individual model, which has

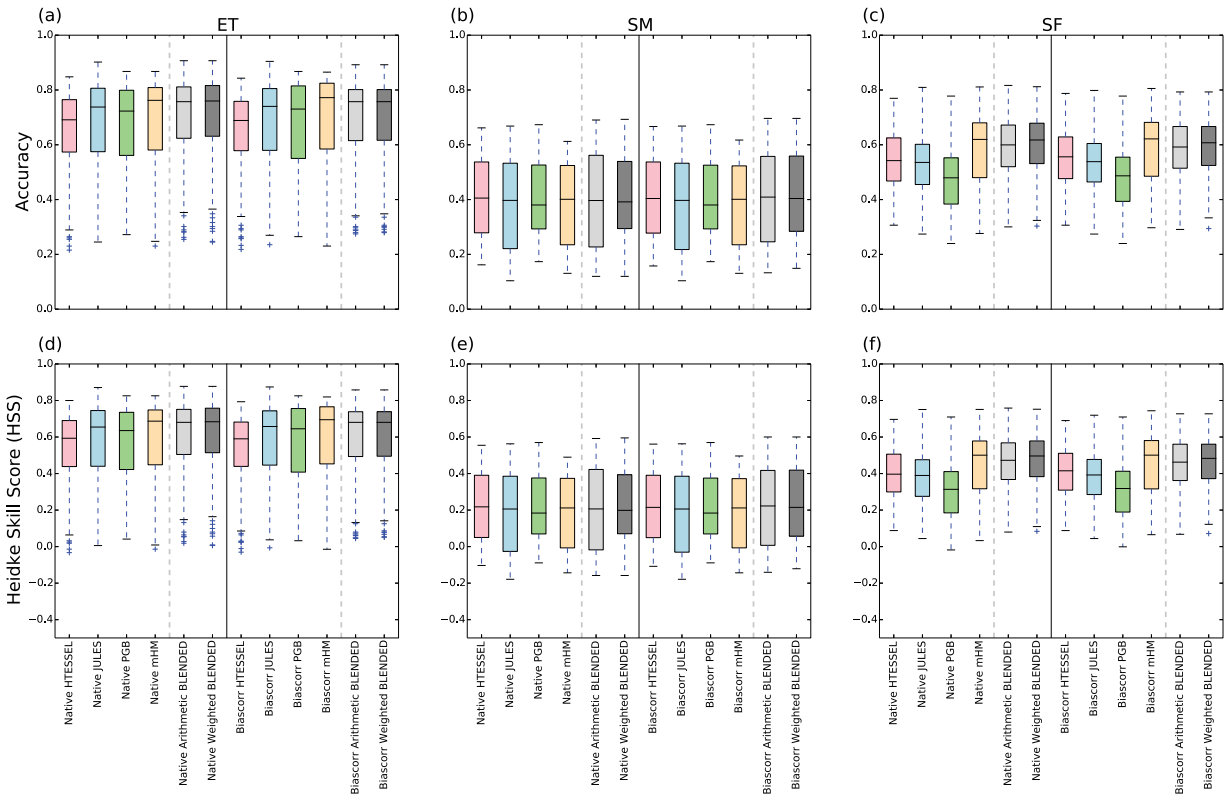


Fig. 9. Same as Fig. 5 but for (a–c) Accuracy and (d–f) for Heidke Skill Score (HSS).

also been shown by other studies (e.g., Mo and Lettenmaier, 2014). Weighted blending adds value to native simulations, but bias-correction leads to much higher improvement. Other studies have also found that bias-correction of forecasts, as a post-processing method, is very effective for multi-model ensemble hydrological forecasts (e.g., Dion et al., 2021). Bohn et al. (2010) also show that individual best bias-corrected model outperforms the multi-model arithmetic average; and blending after bias-correction only marginally improves forecast performance over individual bias-corrected models. Thus, bias-correction is recommended wherever possible. When providing categorical forecasts, bias-correction is not necessary, as the improvement in performance in bias-corrected weighted blending is modest, thus native weighted blended product can be used instead. However, multi-model blending for categorical forecasts could still be beneficial in order to simplify the message to the end users.

Where the application of bias-correction is difficult (e.g., ungauged catchments), weighted blending, using modelled KGE, is a good alternative, and performs better than arithmetic averaging. Applications of the multi-model merging improves the quality of forecasts by maintaining the ensemble dispersion with lead time (e.g., Thiboult et al., 2016; Xu et al., 2019) and enhances the reliability of any forecasting system by incorporating the model uncertainty (e.g., Ahmed et al., 2019). Merging multi-model streamflow forecasts, using methods like Bayesian model averaging, leads to an improvement in skill (e.g., Duan et al., 2007; Luo and Wood, 2008); however, these are approaches that have been investigated on the multi-model probabilistic forecasts rather than baseline simulations as in this study. Probabilistic forecasts include hydrological model uncertainty, as in baseline simulations, but also have uncertainty related to the atmospheric ensemble forecasts used to drive the models (e.g., Krzysztofowicz, 2001). As the skill of these forecasts vary in space and time, the effect of the uncertainties from the atmospheric forcing on the hydrological forecasts could also vary with lead-time and catchment characteristics. At shorter lead-times, forecasts will be influenced more by the initial conditions, whereas for longer lead-times, the climate forcing will have a stronger effect (e.g., Li et al., 2009). For slow-responding catchments, the uncertainties in the climate forcing will have less importance than for flashy catchments (e.g., Sutanto and Van Lanen, 2022). This means that the benefits of our blending and bias-correction method found in this study for the baseline might not transfer directly or uniformly when applied to forecasts. Thus, there is further need to investigate the uncertainty before implementing our method to probabilistic forecasts.

Using consistent weights for all variables (i.e. SF-KGE weighted blended product) has poorer performance as compared to using KGE from respective variables. Thus, it is expected that we use the better performing blending approach (i.e. the weighted blended product) for operational forecasts. Using weights, based on skill of the individual models, to merge multi-model ensemble simulations has been shown to reduce the bias in climate simulations (Thober and Samaniego, 2014). However, weighting each variable with different KGE weights provides us with forecasts without local hydrological balance, and will not be suited for certain user requirements. For applications where water balance closure is necessary (e.g., drought forecasting systems predicting water availability in both rivers and soils), blended product using SF weights gives an alternative to arithmetic averaging, which optimises SF blended product, with ET and SM blended product similar or slightly worse than arithmetic blended product. Multi-model blending at daily timesteps has similar performance to blending at monthly timesteps, and we therefore recommend the use of simulations at either timestep as per the user requirement. This result is different in snowmelt-dominated catchments, where the performance generally improves on aggregating the SF merged simulations from daily to monthly time scales (e.g., Bohn et al., 2010).

5.2. Limitations

Our study is limited by the number of catchments evaluated and the missing data for the variables (SM and SF). Although, we attempt

to make our results as robust as possible, we could not ascertain the exact influence nested catchments have on the results. Uncertainty analysis performed using bootstrapping different sample sizes of catchments (not shown) suggests that our recommendations from Section 5.1 remain valid even with a smaller sample catchments.

Another limitation of our study is that we consider the validation data to be "the truth", but in reality it has its own errors. The errors in SM observations play an important role in poor performance of SM simulations over the catchments studied (e.g., Rakovec et al., 2016). The SM "observations" are derived from a remote sensing product that has some errors and the ET "observations" are a blended observed-modelled product. However, these validation datasets are the best estimate of "observations" available to us for such a global analysis investigating the performance of all three variables over the same catchment sample, but we do recognise the observational uncertainty inherent in our methodology. Further, many previous studies have used these same ET and SM observations to calibrate hydrological models to improve SF forecasts (e.g., López López et al., 2017; Dembélé et al., 2020; Ding and Zhu, 2022). GLEAM ET dataset performs well at a regional (e.g., Yang et al., 2017) or global (e.g., Miralles et al., 2011) scale when validated against in-situ observations in different studies (e.g., Martens et al., 2018). Similarly, studies have also shown that CCI SM product performs well against in-situ observations regionally (e.g., González-Zamora et al., 2019) and globally (e.g., Dorigo et al., 2015). We ourselves have validated the CCI SM product against international soil moisture network (ISMN) stations using the QA4SM platform (https://qa4sm.eu/ui/home) using multiple metrics. The results from mean squared error (MSE; Figure S6a), unbiased root mean squared difference (ubRMSD; Figure S6a) and other metrics (not shown) clearly show that CCI-SM has high performance and skill when validated against observations. These findings have given us confidence that we can use these datasets as proxy for observations to validate the ULYSSES model simulations, especially over data sparse regions.

For SM, an additional source of error stems from the fact that SM depth from models is different to the measurements from satellites, as satellites only estimate moisture in the top few centimetres of soil whereas land surface models simulate moisture at greater depth. Some studies (e.g., Beale et al., 2021) have investigated the possibility to correct for the mismatch in depth using soil moisture depth profile modelling when comparing soil moisture from different sources, and have shown that this can account for large differences. However, these methods require a range of input parameters (soil hydraulic properties, vegetation cover and type, etc.) which are not readily available for global studies such as ours. Therefore, we have followed the approach taken in other studies (e.g., Peng et al., 2021; Schellekens et al., 2017), and have directly compared the modelled SM with the SM observations, although we do acknowledge that this may lead to some uncertainties, and our calculated model performance may not reflect the true model performance.

5.3. Future work

Blended forecast products derived from the ULYSSES project are not limited to the three variables evaluated in this study. Thus, there is a need to assess the skill of the blended products for the whole set of forecast variables (e.g., runoff, snow cover, terrestrial water storage). Implementing our blending approaches over more catchments will provide better estimates of uncertainty. Further research may also evaluate blending approaches on a grid-point level rather than just catchment-level. Modelling KGE for ungauged catchments SF have shown somewhat promising results. However, as noted in the results, catchment characteristics do not completely represent the KGE for ungauged catchments. Improved modelling of KGE in future research can lead to improved weighted blending approaches.

Multi-model streamflow forecast show an improvement in skill upon merging the individual ensembles of all models (e.g., Duan et al., 2007; Luo and Wood, 2008). However, the blending and bias-correction approaches, used in this study, were tested on the model simulations from baseline period (1981–2019) which have a single simulated output per model. As a next step, ULYSSES data should be evaluated for all of its initialisations and the full ensemble of hindcasts and forecasts, to assess its skill and ascertain the uncertainty of the forecasts at different lead times. ULYSSES provides global hindcasts (1993–2019) and operational forecasts (2020–2021) for each model at monthly lead times with 6 months extent each. All model forecasts have a five member ensemble for each initialisation, except for February, May, August and November initialisations, which have 25 members each. For operational blended forecasts, performance metrics for weighting derived from the hindcast period might yield forecasts with better skill. However, additional uncertainty coming from errors in atmospheric forecasts may mean that the improvements found in our research do not directly translate into improvements in hydrological forecasts. Thus, further assessment on the hindcasts are needed. Further, understanding of the performance using the hindcasts will allow identification of forecast and ensemble uncertainty. Skill for blended products should also be assessed for anomalies or categories of variables rather than just for absolute values, to identify the best forecast product that can be provided to the users.

6. Summary and conclusions

This study analyses different approaches for blending multi-model simulations derived from the ULYSSES project, focusing on computationally inexpensive methods in an effort to limit the carbon footprint and cost of this post-processing step. The verification of the blending methods has been performed for 119 sample catchments for the baseline period of 1981–2019 using ET, SM and SF variables. The blending approach tested here is weighted averaging of the native multi-model simulations, using catchment performance metric (KGE) for each variable as the weight. The arithmetic multi-model averaging method is used to identify the added value of the weighted blended approach. The analysis also includes bias-correction of model simulations before applying the two blending approaches. Further, the blending approaches are also tested for SF over ungauged catchments by modelling the KGE metric based on catchment characteristics. We also apply SF KGE weights for weighted blending of ET and SM to have a hydrologically balanced forecast.

Our results show that the weighted blending approach has added value over the arithmetic blending when applied to the native simulation over most catchments. Weighted blending also adds value when applied to the bias-corrected simulations, but the improvement seen is lower than with native simulations. Bias-correction improves individual model simulations for all three variables absolute values but not the categorical forecasts. Results indicate that greater improvement is achieved through bias-correction than through weighted blending. From all the approaches tested, the best performing method is the bias-correction followed by weighted blending. Therefore, bias-correction of simulations before blending using weights is the recommended method, unless only categorical forecasts are considered, in which case native weighted blending is the recommended method.

Modelling of KGE for SF has proven to be a viable option to apply weighted blending at ungauged catchments. Applying consistent weights for blending across all the three variables (based on SF KGE metric) for hydrologically consistent forecasts, shows poorer performance than using variables' respective weights. However, using consistent weights for blending for all variables is a possible option if the local water balance needs to be maintained. The SF-weighted blending shows similar performance to arithmetic blending for ET, slightly poorer for SM, but shows vastly improved performance for SF. Thus, this method might be the most suited for certain applications where the water balance closure is more important (e.g., drought forecasting systems predicting water availability in both rivers and soils) than the accuracy of each individual variable.

The findings of this study show promising results for the application of blending approaches, along with bias-correction, for global hydrological forecasts. More detailed investigations are required for application of this method for providing operational quality hydrological forecasts to the end users.

CRediT authorship contribution statement

Amulya Chevuturi: Investigation, Software, Data curation, Visualisation, Writing – original draft, Review & editing. **Maliko Tanguy:** Investigation, Software, Data curation, Visualisation, Writing – original draft, Review & editing. **Katie Facer-Childs:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Alberto Martínez-de la Torre:** Conceptualization, Data development, Data curation, Writing – review & editing. **Sunita Sarkar:** Visualization, Writing – review & editing. **Stephan Thober:** Data development, Data curation, Visualisation, Writing – review & editing. **Luis Samaniego:** Data development, Data curation, Visualisation, Writing – review & editing. **Oldrich Rakovec:** Data development, Data curation, Visualisation, Writing – review & editing. **Matthias Kelbling:** Data development, Writing – review & editing. **Edwin H. Sutanudjaja:** Data development, Visualisation, Writing – review & editing. **Niko Wanders:** Data development, Visualisation, Writing – review & editing. **Eleanor Blyth:** Conceptualization, Investigation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgements

This research was funded by the Natural Environment Research Council, UK through UKCEH's SUNRISE (award number NE/R000131/1) and NC-International (award number NE/X006247/1) programmes delivering National Capability, as well as the Net Zero Capacity Building Project and UKRI-NERC STF Global Water Tools Project. NW acknowledges funding from NWO, Netherlands 016.Veni.181.049. The Global seasonal forecasts and reforecasts of river discharge and related hydrological variables from a multi-model ensemble, also referred to the ULYSSES dataset, is produced by the Copernicus Climate Change Service (C3S) contract C3S_432_Lot3 and will be hosted by Climate Data Store (CDS). Catchment characteristics information is available at Global Runoff Data Centre website https://www.bafg.de/GRDC/EN/Home/homepage_node.html. Observed data for digital elevation (https://doi.org/10.5066/F7J38R2N), evaporation (https://www.gleam.eu/), surface soil moisture (https://www.esa-soilmoisture-cci.org/node/202) and streamflow (https://www.bafg.de/GRDC/EN/02_srvcs/21_tmsrs/210_prtl/prtl_node.html) are freely available.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jhydrol.2023.129607.

References

Abrahart, R.J., See, L., 2002. Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrol. Earth Syst. Sci. 6 (4), 655–670. http://dx.doi.org/10.5194/hess-6-655-2002.

Ahmed, K., Sachindra, D.A., Shahid, S., Demirel, M.C., Chung, E.-S., 2019. Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. Hydrol. Earth Syst. Sci. 23 (11), 4803–4824. http://dx.doi.org/10.5194/hess-23-4803-2019.

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. J. Hydrometeorol. 7 (4), 755–768. http://dx.doi.org/10.1175/JHM519.1.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res. 207, 155–180. http://dx.doi.org/10.1016/j.atmosres.2018.02.024.

Arsenault, R., Brissette, F., 2016. Multi-model averaging for continuous streamflow prediction in ungauged basins. Hydrol. Sci. J. 61 (13), 2443–2454. http://dx.doi.org/10.1080/02626667.2015.1117088.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.-L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. http://dx.doi.org/10.1016/j.jhydrol.2015.09.001.

Bartholome, E., Belward, A.S., 2005. GLC2000: a new approach to global land cover mapping from Earth observation data. Int. J. Remote Sens. 26 (9), 1959–1977. http://dx.doi.org/10.1080/01431160412331291297.

Beale, J., Waine, T., Evans, J., Corstanje, R., 2021. A method to assess the performance of SAR-derived surface soil moisture products. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4504–4516. http://dx.doi.org/10.1109/JSTARS.2021.3071380.

Best, M., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C., Edwards, J., Hendry, M., Porson, A., Gedney, N., et al., 2011. The Joint UK Land Environment Simulator (JULES), model description–Part 1: energy and water fluxes. Geosci. Model Dev. 4 (3), 677–699. http://dx.doi.org/10.5194/gmd-4-677-2011.

BfG, 2020. The Global Runoff Data Centre. URL https://www.bafg.de/GRDC/EN/Home/homepage_node.html.

Bohn, T.J., Sonessa, M.Y., Lettenmaier, D.P., 2010. Seasonal hydrologic forecasting: Do multimodel ensemble averages always yield improvements in forecast skill? J. Hydrometeorol. 11 (6), 1358–1372. http://dx.doi.org/10.1175/2010JHM1267.1.

Bubeck, P., Otto, A., Weichselgartner, J., 2017. Societal impacts of flood hazards. In: Oxford Research Encyclopedia of Natural Hazard Science. http://dx.doi.org/10.1093/acrefore/9780199389407.013.281.

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M., Pryor, M., Rooney, G., Essery, R., Blyth, E., et al., 2011. The Joint UK Land Environment Simulator (JULES), model description–Part 2: carbon fluxes and vegetation dynamics. Geosci. Model Dev. 4 (3), 701–722. http://dx.doi.org/10.5194/gmd-4-701-2011.

Danielson, J.J., Gesch, D.B., 2011. Global multi-resolution terrain elevation data 2010 (GMTED2010). http://dx.doi.org/10.3133/ofr20111073,

Darbandsari, P., Coulibaly, P., 2019. Inter-comparison of different Bayesian model averaging modifications in streamflow simulation. Water 11 (8), 1707. http://dx.doi.org/10.3390/w11081707.

Dembélé, M., Ceperley, N., Zwart, S.J., Salvadore, E., Mariethoz, G., Schaefli, B., 2020. Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. Adv. Water Resour. 143, 103667. http://dx.doi.org/10.1016/j.advwatres.2020.103667.

Diks, C.G., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. 24, 809–820. http://dx.doi.org/10.1007/s00477-010-0378-z.

Ding, J., Zhu, Q., 2022. The accuracy of multisource evapotranspiration products and their applicability in streamflow simulation over a large catchment of Southern China. J. Hydrol.: Reg. Stud. 41, 101092. http://dx.doi.org/10.1016/j.ejrh.2022.101092.

Dion, P., Martel, J.-L., Arsenault, R., 2021. Hydrological ensemble forecasting using a multi-model framework. J. Hydrol. 600, 126537. http://dx.doi.org/10.1016/j.jhydrol.2021.126537.

Döll, P., Kaspar, F., Lehner, B., 2003. A global hydrological model for deriving water availability indicators: model tuning and validation. J. Hydrol. 270 (1–2), 105–134. http://dx.doi.org/10.1016/S0022-1694(02)00283-4.

Dorigo, W., Gruber, A., De Jeu, R., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R., et al., 2015. Evaluation of the ESA CCI soil moisture product using ground-based observations. Remote Sens. Environ. 162, 380–395. http://dx.doi.org/10.1016/j.rse.2014.07.023.

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., et al., 2017. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. Remote Sens. Environ. 203, 185–215. http://dx.doi.org/10.1016/j.rse.2017.07.001.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30 (5), 1371–1386. http://dx.doi.org/10.1016/j.advwatres.2006.11.014.

Eichhorn, A., Breitholtz, M., Domcke, V., Hladky, J., Hopkins, D., Kreil, A., Sörlin, S., Torney, D., 2022. Towards Climate Sustainability of the Academic System in Europe and Beyond. URL https://allea.org/wp-content/uploads/2022/05/ALLEA-Report-Towards-Climate-Sustainability-of-the-Academic-System.pdf,

Farmer, W.H., Over, T.M., Kiang, J.E., 2018. Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. Hydrol. Earth Syst. Sci. 22 (11), 5741–5758. http://dx.doi.org/10.5194/hess-22-5741-2018.

Georgakakos, K.P., Seo, D.-J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J. Hydrol. 298 (1–4), 222–241. http://dx.doi.org/10.1016/j.jhydrol.2004.03.037.

González-Zamora, Á., Sánchez, N., Pablos, M., Martínez-Fernández, J., 2019. CCI soil moisture assessment with SMOS soil moisture and in situ data under different environmental conditions and spatial scales in Spain. Remote Sens. Environ. 225, 469–482. http://dx.doi.org/10.1016/j.rse.2018.02.010.

GRDC-WMO, 2021. Global Runoff Data Centre (GRDC - WMO). URL https://www.un-spider.org/links-and-resources/data-sources/global-runoff-data-centre-grdc-wmo.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91. http://dx.doi.org/10.1016/j.jhydrol.2009.08.003.

Hall, D., Riggs, G., 2016. MODIS/Aqua Snow Cover Daily L3 Global 500 m SIN Grid, Version 6 [2002–2015]. http://dx.doi.org/10.5067/MODIS/MYD10A1.006,

Heidke, P., 1926. Berechnung des erfolges und der güte der windstärkevorhersagen im Sturmwarnungsdienst. Geogr. Ann. 8 (4), 301–349. http://dx.doi.org/10.1080/20014422.1926.11881138.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol.: J. R. Meteorol. Soc. 25 (15), 1965–1978. http://dx.doi.org/10.1002/joc.1276.

Jenkins, A., Dixon, H., Barlow, V., Smith, K., Cullmann, J., Berod, D., Kim, H., Schwab, M., Silva Vara, L.R., 2020. HydroSOS – The hydrological status and outlook system. In: WMO Bulletin. URL https://public.wmo.int/en/resources/bulletin/hydrosos-%E2%80%93-hydrological-status-and-outlook-system.

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., et al., 2019. SEAS5: the new ECMWF seasonal forecast system. Geosci. Model Dev. 12 (3), 1087–1117. http://dx.doi.org/10.5194/gmd-12-1087-2019.

Jozaghi, A., Shen, H., Ghazvinian, M., Seo, D.-J., Zhang, Y., Welles, E., Reed, S., 2021. Multi-model streamflow prediction using conditional bias-penalized multiple linear regression. Stoch. Environ. Res. Risk Assess. 35 (11), 2355–2373. http://dx.doi.org/10.1007/s00477-021-02048-3.

Knoben, W.J., Freer, J.E., Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331. http://dx.doi.org/10.5194/hess-23-4323-2019.

Kobold, M., Sušelj, K., 2005. Precipitation forecasts and their uncertainty as input into hydrological models. Hydrol. Earth Syst. Sci. 9 (4), 322–332. http://dx.doi.org/10.5194/hess-9-322-2005.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. J. Hydrol. 249 (1–4), 2–9. http://dx.doi.org/10.1016/S0022-1694(01)00420-6.

Kumar, R., Samaniego, L., Attinger, S., 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resour. Res. 49 (1), 360–379. http://dx.doi.org/10.1029/2012WR012195.

Kummu, M., Taka, M., Guillaume, J.H., 2018. Gridded global datasets for gross domestic product and Human Development Index over 1990–2015. Sci. Data 5 (1), 1–15. http://dx.doi.org/10.1038/sdata.2018.4.

Lahoz, W.A., De Lannoy, G.J., 2014. Closing the gaps in our knowledge of the hydrological cycle over land: Conceptual problems. Surv. Geophys. 35, 623–660. http://dx.doi.org/10.1007/s10712-013-9221-7.

Lannelongue, L., Inouye, M., 2023. Carbon footprint estimation for computational research. Nat. Rev. Methods Primers http://dx.doi.org/10.1038/s43586-023-00202-5.

Lavers, D.A., Ramos, M.-H., Magnusson, L., Pechlivanidis, I., Klein, B., Prudhomme, C., Arnal, L., Crochemore, L., Van Den Hurk, B., Weerts, A.H., Harrigan, S., Cloke, H.L., Richardson, D.S., Pappenberger, F., 2020. A vision for hydrological prediction. Atmosphere 11 (3), 237. http://dx.doi.org/10.3390/atmos11030237.

Lehner, B., 2019. RiverATLAS Attributes (version 1.0). URL https://hydrosheds.org/images/inpages/RiverATLAS_Catalog_v10.pdf.

Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27 (15), 2171–2186. http://dx.doi.org/10.1002/hyp.9740.

Lehner, B., Liermann, C.R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J.C., Rödel, R., Sindorf, N., Wisser, D., 2011. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. Front. Ecol. Environ. 9 (9), 494–502. http://dx.doi.org/10.1890/100125.

Li, H., Luo, L., Wood, E.F., Schaake, J., 2009. The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. J. Geophys. Res.: Atmos. 114 (D4), http://dx.doi.org/10.1029/2008JD010969.

Linke, S., Lehner, B., Dallaire, C.O., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., et al., 2019. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Sci. Data 6 (1), 1–15. http://dx.doi.org/10.1038/s41597-019-0300-6.

Liu, Y.Y., Parinussa, R., Dorigo, W.A., De Jeu, R.A., Wagner, W., Van Dijk, A., McCabe, M.F., Evans, J., 2011. Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. Hydrol. Earth Syst. Sci. 15 (2), 425–436. http://dx.doi.org/10.5194/hess-15-425-2011.

López López, P., Sutanudjaja, E.H., Schellekens, J., Sterk, G., Bierkens, M.F., 2017. Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. Hydrol. Earth Syst. Sci. 21 (6), 3125–3144. http://dx.doi.org/10.5194/hess-21-3125-2017.

Luo, L., Wood, E.F., 2008. Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. J. Hydrometeorol. 9 (5), 866–884. http://dx.doi.org/10.1175/2008JHM980.1.

Martens, B., De Jeu, R.A., Verhoest, N.E., Schuurmans, H., Kleijer, J., Miralles, D.G., 2018. Towards estimating land evaporation at field scales using GLEAM. Remote Sens. 10 (11), 1720. http://dx.doi.org/10.3390/rs10111720.

Martens, B., Miralles, D.G., Lievens, H., Van Der Schalie, R., De Jeu, R.A., Fernández-Prieto, D., Beck, H.E., Dorigo, W.A., Verhoest, N.E., 2017. GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. Geosci. Model Dev. 10 (5), 1903–1925. http://dx.doi.org/10.5194/gmd-10-1903-2017.

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nature Commun. 7 (1), 1–11. http://dx.doi.org/10.1038/ncomms13603.

Metzger, M.J., Bunce, R.G., Jongman, R.H., Sayre, R., Trabucco, A., Zomer, R., 2013. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. Global Ecol. Biogeogr. 22 (5), 630–638. http://dx.doi.org/10.1111/geb.12022.

Miralles, D., De Jeu, R., Gash, J., Holmes, T., Dolman, A., 2011. Magnitude and variability of land evaporation and its components at the global scale. Hydrol. Earth Syst. Sci. 15 (3), 967–981. http://dx.doi.org/10.5194/hess-15-967-2011.

Mo, K.C., Lettenmaier, D.P., 2014. Hydrologic prediction over the conterminous United States using the national multi-model ensemble. J. Hydrometeorol. 15 (4), 1457–1472. http://dx.doi.org/10.1175/JHM-D-13-0197.1.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Environ. Res. Risk Assess. 17 (5), 291–305. http://dx.doi.org/10.1007/s00477-003-0151-7.

Peng, J., Tanguy, M., Robinson, E.L., Pinnington, E., Evans, J., Ellis, R., Cooper, E., Hannaford, J., Blyth, E., Dadson, S., 2021. Estimation and evaluation of high-resolution soil moisture from merged model and Earth observation data in the Great Britain. Remote Sens. Environ. 264, 112610. http://dx.doi.org/10.1016/j.rse.2021.112610.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., Samaniego, L., 2016. Multiscale and multivariate evaluation of water fluxes and states over European river basins. J. Hydrometeorol. 17 (1), 287–307. http://dx.doi.org/10.1175/JHM-D-15-0054.1.

Ramankutty, N., Evan, A.T., Monfreda, C., Foley, J.A., 2008. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. Glob. Biogeochem. Cycles 22 (1), http://dx.doi.org/10.1029/2007GB002952.

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

Roy, T., Valdés, J.B., Serrat-Capdevila, A., Durcik, M., Demaria, E.M., Valdés-Pineda, R., Gupta, H.V., 2020. Detailed Overview of the multimodel multiproduct streamflow forecasting platform. J. Appl. Water Eng. Res. 8 (4), 277–289. http://dx.doi.org/10.1080/23249676.2020.1799442.

Samaniego, L., Kumar, R., Attinger, S., 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res. 46 (5), http://dx.doi.org/10.1029/2008WR007327.

Samaniego, L., Thober, S., Rakovec, O., Wanders, N., Sutanudjaja, E., Martínez-de la Torre, A., Blyth, E., 2020. Updated document describing the design of the production chain: C3S-432-Lot3-UFZ - Global Multi-model hydrological Seasonal predictions. URL https://www.ufz.de/export/data/2/242466_D121_v2.pdf. (Accessed: 19/04/2021).

Sammut, C., Webb, G.I., 2010. Leave-one-out cross-validation. Encycl. Mach. Learn. 600–601. http://dx.doi.org/10.1007/978-0-387-30164-8_469.

Sanchez Lozano, J., Romero Bustamante, G., Hales, R., Nelson, E.J., Williams, G.P., Ames, D.P., Jones, N.L., 2021. A streamflow bias correction and performance evaluation web application for GEOGloWS ECMWF streamflow services. Hydrology 8 (2), 71. http://dx.doi.org/10.3390/hydrology8020071.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., et al., 2017. A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset. Earth Syst. Sci. Data 9 (2), 389–413. http://dx.doi.org/10.5194/essd-9-389-2017.

Shamseldin, A.Y., O'Connor, K.M., Liang, G., 1997. Methods for combining the outputs of different rainfall–runoff models. J. Hydrol. 197 (1–4), 203–229. http://dx.doi.org/10.1016/S0022-1694(96)03259-3.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., Scanlon, B.R., 2015. A global data set of the extent of irrigated land from 1900 to 2005. Hydrol. Earth Syst. Sci. 19 (3), 1521–1545. http://dx.doi.org/10.5194/hess-19-1521-2015.

Siqueira, V.A., Paiva, R.C., Fleischmann, A.S., Fan, F.M., Ruhoff, A.L., Pontes, P.R., Paris, A., Calmant, S., Collischonn, W., 2018. Toward continental hydrologic–hydrodynamic modeling in South America. Hydrol. Earth Syst. Sci. 22 (9), 4815–4842. http://dx.doi.org/10.5194/hess-22-4815-2018.

Sood, A., Smakhtin, V., 2015. Global hydrological models: a review. Hydrol. Sci. J. 60 (4), 549–565. http://dx.doi.org/10.1080/02626667.2014.950580.

Sutanto, S.J., Van Lanen, H.A., 2022. Catchment memory explains hydrological drought forecast performance. Sci. Rep. 12 (1), 2689. http://dx.doi.org/10.1038/s41598-022-06553-5.

Sutanudjaja, E.H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J.H., Drost, N., Van Der Ent, R.J., De Graaf, I.E., Hoch, J.M., De Jong, K., et al., 2018. PCR-GLOBWB 2: A 5 arcmin global hydrological and water resources model. Geosci. Model Dev. 11 (6), 2429–2453. http://dx.doi.org/10.5194/gmd-11-2429-2018.

Thiboult, A., Anctil, F., Boucher, M.-A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrol. Earth Syst. Sci. 20 (5), 1809–1825. http://dx.doi.org/10.5194/hess-20-1809-2016.

Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., Samaniego, L., 2019. The multiscale routing model mRM v1.0: Simple river routing at resolutions from 1 to 50 km. Geosci. Model Dev. 12 (6), 2501–2521. http://dx.doi.org/10.5194/gmd-12-2501-2019.

Thober, S., Kumar, R., Sheffield, J., Mai, J., Schäfer, D., Samaniego, L., 2015. Seasonal soil moisture drought prediction over Europe using the North American Multi-Model Ensemble (NMME). J. Hydrometeorol. 16 (6), 2329–2344. http://dx.doi.org/10.1175/JHM-D-15-0053.1.

Thober, S., Samaniego, L., 2014. Robust ensemble selection by multivariate evaluation of extreme precipitation and temperature characteristics. J. Geophys. Res.: Atmos. 119 (2), 594–613. http://dx.doi.org/10.1002/2013JD020505.

Trabucco, A., Zomer, R., 2010. Global soil water balance geospatial database. URL https://cgiarcsi.community/. Available from the CGIAR-CSI GeoPortal.

UFZ, 2020. ULYSSES. URL https://www.ufz.de/index.php?en=47367. (Accessed: 2021-09-29).

Velazquez, J.A., Anctil, F., Ramos, M., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci. 29, 33–42. http://dx.doi.org/10.5194/adgeo-29-33-2011.

Wanders, N., Wood, E.F., 2016. Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environ. Res. Lett. 11 (9), 094007. http://dx.doi.org/10.1088/1748-9326/11/9/094007.

Wang, J., Wang, X., Khu, S.T., 2023. A decomposition-based multi-model and multi-parameter ensemble forecast framework for monthly streamflow forecasting. J. Hydrol. 129083. http://dx.doi.org/10.1016/j.jhydrol.2023.129083.

Wilhite, D.A., Svoboda, M.D., Hayes, M.J., 2007. Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. Water Resour. Manag. 763–774. http://dx.doi.org/10.1007/s11269-006-9076-5.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, third ed. In: International geophysics series, vol. 100, Academic Press, URL https://www.sciencedirect.com/bookseries/international-geophysics/vol/100/.

WMO, 2021. HydroSOS. URL https://public.wmo.int/en/our-mandate/what-we-do/application-services/hydrosos.

Xu, J., Anctil, F., Boucher, M.-A., 2019. Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. J. Hydrol. 578, 124002. http://dx.doi.org/10.1016/j.jhydrol.2019.124002.

Yang, X., Yong, B., Ren, L., Zhang, Y., Long, D., 2017. Multi-scale validation of GLEAM evapotranspiration products over China via ChinaFLUX ET measurements. Int. J. Remote Sens. 38 (20), 5688–5709. http://dx.doi.org/10.1080/01431161.2017.1346400.

Zaherpour, J., Mount, N., Gosling, S.N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Schmied, H.M., Tang, Q., et al., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environ. Model. Softw. 114, 112–128. http://dx.doi.org/10.1016/j.envsoft.2019.01.003.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Adv. Sci. Res. 8 (1), 135–141. http://dx.doi.org/10.5194/asr-8-135-2012.

Zomer, R.J., Trabucco, A., Bossio, D.A., Verchot, L.V., 2008. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agricult. Ecosys. Environ. 126 (1–2), 67–80. http://dx.doi.org/10.1016/j.agee.2008.01.014.