Water Resources Research

TECHNICAL REPORTS: DATA

10.1029/2020WR028787

G-RUN ENSEMBLE: A Multi-Forcing Observation-Based Global Runoff Reanalysis

G. Ghiggi[1,2], V. Humphrey[3,4], S. I. Seneviratne[2], and L. Gudmundsson[2]

[1]Environmental Remote Sensing Laboratory (LTE), EPFL, Lausanne, Switzerland, [2]Institute for Atmospheric and Climate Science, ETH, Zurich, Switzerland, [3]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA, [4]Department of Geography, University of Zurich, Zurich, Switzerland

Key Points:

- A multi-forcing global runoff reanalysis is created by means of machine learning and a global collection of river discharge observations
- G-RUN ENSEMBLE allows for an unprecedented view of global terrestrial water dynamics on time scales ranging from months to a full century
- Quantification of the uncertainty stemming from the atmospheric forcing data makes the G-RUN ENSEMBLE a good candidate for reliable and robust water resources assessments

Supporting Information:

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

Correspondence to:

G. Ghiggi,
gionata.ghiggi@epfl.ch

Citation:

Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2021). G-RUN ENSEMBLE: A multi-forcing observation-based global runoff reanalysis. *Water Resources Research*, *57*, e2020WR028787. https://doi.org/10.1029/2020WR028787

Received 9 SEP 2020
Accepted 23 APR 2021

Author Contributions:

Conceptualization: G. Ghiggi, L. Gudmundsson
Data curation: G. Ghiggi
Formal analysis: G. Ghiggi
Funding acquisition: S. I. Seneviratne, L. Gudmundsson
Investigation: G. Ghiggi
Methodology: G. Ghiggi, V. Humphrey, L. Gudmundsson

Abstract River discharge is an Essential Climate Variable (ECV) and is one of the best monitored components of the terrestrial water cycle. Nonetheless, gauging stations are distributed unevenly around the world, leaving many white spaces on global freshwater resources maps. Here, we use a machine learning algorithm and historical weather data to upscale sparse in situ river discharge measurements. We provide a global reanalysis of monthly runoff rates for periods covering decades to the past century at a resolution of 0.5° (about 55 km), and with up to 525 ensemble members based on 21 different atmospheric forcing data sets. This global runoff reconstruction, named Global RUNoff ENSEMBLE (G-RUN ENSEMBLE), is evaluated using independent observations from large river basins and benchmarked against other publicly available runoff data sets over the period 1981–2010. The accuracy of the data set is evaluated on observed river flow from basins not used for model calibration and is found to compare favorably against state-of-the-art global hydrological model simulations. The G-RUN ENSEMBLE estimates the global mean runoff volume to range between 3.2×10^4 and 3.8×10^4 km^3 yr^{-1}. This publicly available data set (https://doi.org/10.6084/m9.figshare.12794075) has a wide range of applications, including regional water resources assessments, climate change attribution studies, hydro-climatic process studies as well as the evaluation, calibration and refinement of global hydrological models.

Plain Language Summary River discharge is related to numerous impacts associated with floods and droughts. It is relatively well observed in North America and Europe, but no global data sets of monthly river discharge accounting for the uncertainty in historical hydrometeorological conditions have been available so far. Here we derive a new global 119-year-long monthly data set of river discharge at 0.5°-resolution (ca. 55 km) using a machine learning algorithm and historical weather data in combination with local discharge data. This new reconstruction, the G-RUN ENSEMBLE, is publicly available (https://doi.org/10.6084/m9.figshare.12794075). It is relevant for numerous water cycle applications.

1. Introduction

River discharge is listed as an Essential Climate Variable (ECV) by the World Meteorological Organization (WMO) (Bojinski et al., 2014) and is one of the best monitored variables of the terrestrial water cycle. Nonetheless, in recent decades available observations have decreased significantly, often in relation to a lack of financial resources or political barriers to data access (Crochemore et al., 2020; Fekete et al., 2012; Fekete et al., 2015; Fekete & Vörösmarty, 2007; Hannah et al., 2011; Laudon et al., 2017; Shiklomanov, Lammers, & Vörösmarty, 2002; Viglione et al., 2010). Human-induced climate change affects the hydrological cycle and thus the availability of water resources (Gudmundsson et al., 2021; Gudmundsson, Seneviratne, & Zhang, 2017; Padrón et al., 2020). Water scarcity arises due to temporal mismatch between water demand and availability, and 80% of the world population is exposed to high levels of threat to water security (Mekonnen & Hoekstra, 2016; Vörösmarty et al., 2010). Consequently, reliable information on the present and past evolution of the world's freshwater resources is essential for assessing ongoing climate change and for putting emerging extreme events into context. Although global hydrological models (GHMs) provide gridded estimates of the various water balance components, several model evaluation studies did highlight large discrepancies between simulations, in situ observations and remote-sensing estimates of evapotranspiration (Miralles et al., 2016; Mueller et al., 2013; Wartenburger et al., 2018), terrestrial water storage (Humphrey & Gudmundsson, 2019; Humphrey, Gudmundsson, & Seneviratne, 2017; Scanlon et al., 2018) and runoff

Resources: S. I. Seneviratne, L. Gudmundsson
Software: G. Ghiggi
Visualization: G. Ghiggi
Writing – original draft: G. Ghiggi, L. Gudmundsson
Writing – review & editing: G. Ghiggi, V. Humphrey, S. I. Seneviratne, L. Gudmundsson

(Beck et al., 2017; Ghiggi et al., 2019; Gudmundsson et al., 2012). In recent years, a number of studies have further investigated the potential of optimizing global hydrology models for producing high-resolution discharge estimates, often with an operational focus (Alfieri et al., 2020; Harrigan et al., 2020; Lin et al., 2019). This study aims to further constrain previous data-driven estimates of monthly runoff rates derived from the G-RUN data set (Ghiggi et al., 2019). G-RUN is an observation-based runoff reconstruction which employed a machine learning (ML) algorithm to estimate global runoff rates on an observational basis. One drawback of G-RUN is that it is based on a single atmospheric forcing data set (GSWP3; Kim et al., 2017). As a result, the uncertainty related to the forcing data is not accounted for, even though it can be significant, in particular for long-term trends (Humphrey & Gudmundsson, 2019). In this study we aim to account for this uncertainty by using an ensemble of 21 gridded atmospheric forcing data sets, including a set of atmospheric reanalysis, post-processed reanalysis and interpolated-stations data. In the following, we reiterate the methodology detailed in Ghiggi et al. (2019) to produce this ensemble of runoff reconstructions referred to as Global-RUNoff ENSEMBLE (G-RUN ENSEMBLE). The G-RUN ENSEMBLE is evaluated using a database of river discharge observations from large river basins and benchmarked against a comprehensive collection of publicly available global monthly runoff reconstructions spanning the period 1981–2010. The manuscript concludes with examples of new applications enabled by the G-RUN ENSEMBLE.

2. Data

2.1. Monthly River Discharge Data

The Global Streamflow Indices and Metadata Archive (GSIM) (Do et al., 2018; Gudmundsson et al., 2018) provides a publicly available collection of streamflow indices at more than 35,000 stations that were obtained by merging existing international and national databases. Prior to production of the data product presented in this study, monthly river discharge data have been screened to remove timeseries with inhomogeneous behavior, unphysical values (i.e., negative values), mislabeled missing values, and uncertain catchment area (see Ghiggi et al., 2019 for details on the full procedure). Stations with catchment area between 10 and 2,500 km^2 have been selected to retrieve the grid cell runoff rates used as observations for model training, while stations of large river basins with catchment area larger than 10,000 km^2 have been used to benchmark the G-RUN ENSEMBLE reconstructions against the GHM runoff simulations.

2.2. Monthly Gridded Precipitation and Temperature Data

Gridded observations of precipitation and 2-m air temperature are obtained from 21 global data sets (Figure 1). The choice has been restricted to data sets with spatial resolution equal or higher than 0.5° and spanning a period of at least 40 years. All data were aggregated to monthly resolution and spatially resampled to a common 0.5-degree grid using conservative remapping (Jones, 1999). The considered atmospheric data are classified according to their primary mode of production into interpolated station observations (Harris et al., 2020), atmospheric reanalysis (Dee et al., 2011; Gelaro et al., 2017; Hersbach et al., 2020) and post-processed atmospheric reanalysis (Balsamo et al., 2015; Boogaard et al., 2020; Cucchi et al., 2020; Kim et al., 2017; Lange, 2019; Mengel et al., 2020; Muñoz-Sabater et al., 2021; Reichle et al., 2017; Sheffield, Goteti, & Wood, 2006; Weedon et al., 2014; Weedon et al., 2011). Atmospheric reanalyses assimilate ground and satellite observations to adjust the variables states within numerical weather prediction models. In this study, post-processed atmospheric reanalysis refers to data from an atmospheric reanalysis which have been further adjusted to better match selected observations, e.g., through means of bias correction against an observational reference such GPCC (Becker et al., 2013). Note that the considered atmospheric data sets have different temporal coverage because of design decisions that are often related to a tradeoff between availability and quality of observations. In order to include an additional precipitation data set, MSWEP v2.2 precipitation (Beck et al., 2019) has been combined with 2 m temperature from ERA5 (denoted as MSWEP v.2.2*) and ERA5-Land (denoted as MSWEP v.2.2**).

2.3. Other Runoff Estimates Used for Benchmarking

The accuracy of the G-RUN ENSEMBLE is benchmarked against a comprehensive set of runoff simulations from global hydrological models (GHM) covering the period 1981–2010. This includes the multi-model

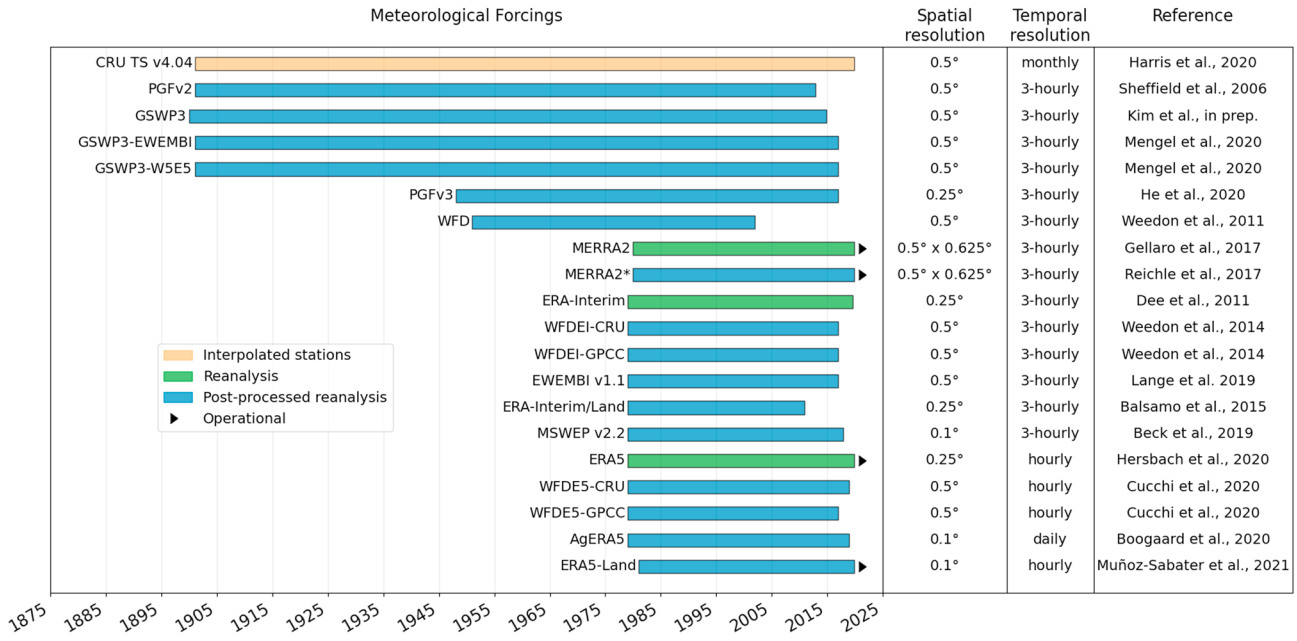

Figure 1. Characteristics of the atmospheric forcing data sets used in the study.

ensemble median of the GHMs intercomparison projects ISIMIP2a "nosoc" (Warszawski et al., 2014), eartH2Observe Tier-1 WRR1 (Schellekens et al., 2017), eartH2Observe Tier-2 WRR2 (Dutra et al., 2017); the LORA v1.0 product (Hobeichi et al., 2019) resulting from the statistical post-processing of the eartH2Observe WRR1 and WRR2 GHM runoff simulations; the runoff reconstruction used in the Global Drought and Flood Catalog (GDFC) (He et al., 2020); as well as runoff simulations from ERA-Interim (Dee et al., 2011), ERA-Interim/Land (Balsamo et al., 2015), ERA5 (Hersbach et al., 2020) and ERA5-Land reanalysis (Muñoz-Sabater et al., 2021). All runoff simulations are aggregated to the monthly resolution and interpolated to a common 0.5-degree grid using conservative remapping.

3. Methods

3.1. Model Setup

Runoff is defined here as the amount of water that is draining from a given land unit (i.e., grid cell) eventually entering the river system, including surface and sub-surface runoff as well as snowmelt. Runoff is difficult to measure over an extended area, but at a monthly timescale, the monthly river discharge measured at the outlet of small catchments divided by the catchment's area can be used as a proxy of the average catchment runoff, provided storage of river water (e.g., in dams, reservoirs) and/or river water losses (e.g., river channel and lake evaporation, irrigation) are minimal. Here observational grid cell runoff is estimated using the procedure described in Ghiggi et al. (2019). To this end, river flow observations from small catchments with an area between 10 and 2,500 km^2 are first assigned to 0.5° grid cells. Following Gudmundsson and Seneviratne (2015, 2016), a random forest (RF) algorithm (Breiman, 2001) is then used to learn the runoff generation process without the explicit description of the involved hydrological processes. The monthly runoff rate (R) is modeled as a function of antecedent monthly precipitation (P) and monthly near-surface temperature (T) such that

$$R_{s,t} = f\left(\tau_n\left(P_{s,t}\right), \tau_n\left(T_{s,t}\right)\right), \quad (1)$$

where f corresponds to the RF model (RFM) characterized here by 300 trees with a maximum depth of 60 splits and no minimum leaf size; s represents the identifier of the grid cell, t is the time step, and $\tau_n\left(X_{s,t}\right) = \left[X_{s,t}, X_{s,t-1}, \cdots, X_{s,t-n} \right]$ is a time lag operator that provides information about meteorological

conditions of the past n months (here $n = 6$) allowing the RFM to approximate memory effects that influence the runoff generation process. The decision to only consider precipitation and temperature as explanatory variables is motivated by Gudmundsson and Seneviratne (2015), who found that the inclusion of other atmospheric variables as well as selected land parameters (topography and soil texture) did not significantly improve the overall accuracy of the estimate. Furthermore, reducing the number of predictor variables also helped to reduce computational costs significantly. While a more extensive screening of other land parameters is beyond the scope of this study, this could be the subject of future research.

Gridded precipitation and temperature data are then used to reconstruct runoff rates globally, also at ungauged grid cells. Since machine learning algorithm predictions are conditioned by the data used for model training, the following strategy is adopted to characterize the sampling uncertainty. For each forcing data set, 25 runoff reconstructions are generated using a Monte Carlo approach in which each RFM is trained using a random subset of only 60% of all grid cells containing runoff observations across the whole time period spanned by the forcing data set. Each of these 25 reconstructions thus represents the result of an RFM calibrated with slightly different observational data, thus providing a proxy for the effects of sampling uncertainty on the runoff estimate.

The G-RUN ENSEMBLE data repository (https://doi.org/10.6084/m9.figshare.12794075) provides individual ensemble members ($25 \times 21 = 525$ realizations), as well as the ensemble mean of these realizations for each forcing data set. The multi-model median of the G-RUN ENSEMBLE members (combining the estimates obtained with all atmospheric forcing data sets) is referred to as G-RUN ENSEMBLE MMM.

3.2. Model Evaluation

A selection of 1,205 river discharge observations from large basins (with areas bigger than 10,000 km^2) that are included in GSIM is used to evaluate the accuracy of the G-RUN ENSEMBLE reconstructions and for benchmarking against GHM simulations. Since the RFM is calibrated using only runoff observations from small catchments (with areas smaller than 2,500 km^2), the verification of the G-RUN ENSEMBLE can be considered as "out-of-sample," although it must be noticed that some large river basins used for model validation include small sub-watersheds used for model training. Selection criteria of these stations follow the methodology described in Ghiggi et al. (2019). To obtain a common temporal coverage across all forcing data sets and the GHM simulations (except for G-RUN based on WFD which stops in 2001), only river discharge observations of the period 1981–2010 have been selected. Note that in contrast to the G-RUN ENSEMBLE, GHM simulations are rarely provided alongside detailed information on model calibration and tuning. Consequently, it cannot be excluded that observations from large river basins that are used here for model evaluation might have been also used for the GHMs calibration (Alcamo et al., 2003; Alfieri et al., 2020; Döll et al., 2003; Hirpa et al., 2018; Hobeichi et al., 2019; Hunger & Döll, 2008; Nijssen et al., 2001).

In order to compare gridded runoff estimates to observed river discharge from large basins we adopt the approach of Gudmundsson and Seneviratne (2015, 2016). To this end, river discharge is estimated by spatially averaging the grid cell runoff times series within the basin and multiplying it by the drainage area. At a monthly timescale, the effect of river routing is considered negligible, except for a few very large basins (Allen et al., 2018). As in Ghiggi et al. (2019), six performance metrics have been used to assess the accuracy of the reconstructions in reproducing different aspects of the river discharge time series. The terms p_t and o_t refer to the predicted and observed time series respectively.

The relative bias (relBIAS) has an optimal value of zero and allows to investigate the presence of systematic errors. A positive (negative) value indicates a general overestimation (underestimation). It is defined as:

$$\text{relBIAS} = \frac{\text{mean}\left(p_t - o_t\right)}{\text{mean}\left(o_t\right)} \tag{2}$$

The ratio of standard deviations (rSD) has an optimal value of one. Values lower than one indicate underestimation, while values higher than one indicate overestimation of the observed variability. It is defined as:

$$\mathrm{rSD} = \frac{\mathrm{sd}\left(p_t\right)}{\mathrm{sd}\left(o_t\right)} \tag{3}$$

The squared correlation coefficient, R2, ranges between zero and one. It measures the degree of the linear association between the predicted time series and the observed one. It is insensitive to the bias. The optimal value is one.

The Nash–Sutcliffe efficiency (NSE), also called model efficiency (Nash & Sutcliffe, 1970), is a measure of the overall predictive skill of the model relative to the long-term mean of the time series. An NSE value of one corresponds to a perfect match between predicted and observed data, while a value lower than zero indicates that model predictions are on average less accurate than those obtained by using the long-term mean of the observed time series (mean(o_t)) as predictor. It is defined as:

$$\mathrm{NSE} = \frac{\sum_t \left(p_t - o_t\right)^2}{\sum_t \left(o_t - \mathrm{mean}\left(o_t\right)\right)^2} \tag{4}$$

The squared correlation coefficient between the observed and predicted monthly standardized anomalies (i.e., monthly time series with the monthly climatology removed, divided by the long-term standard deviation of each month) is $R2_{anom}$. It ranges from zero to one (best value).

The squared correlation coefficient between the observed and predicted monthly climatology is $R2_{clim}$. It ranges between zero and one (best value).

The skill metrics of G-RUN ENSEMBLE members at each GSIM station are available in Data Set S1 as basis for future model benchmarking.

4. Model Benchmarking

4.1. Accuracy of the Monthly Reconstructions

Figure 2a shows the distribution of the skill metrics of the G-RUN ENSEMBLE members and GHM simulations. A selected subset of observed river discharge time series is compared to the G-RUN ENSEMBLE and the GHM reconstructions in Figure 2b (the full set is provided in Data Set S3). Most runoff estimates tend to overestimate monthly river discharge rates in large basins, but in general the relative bias does not exceed 20%. A similar behavior is observed for the reproduction of the magnitude of monthly variability as illustrated by the rSD metric, even though the G-RUN ENSEMBLE members tend to overestimate the monthly variability less than the GHM simulations. To complement the analysis of the time series magnitude and amplitude, Figure S1 illustrates the average difference in annual runoff between GHM runoff simulations and the G-RUN ENSEMBLE MMM. The GHMs show a complex spatial pattern of lower and higher runoff estimates. Overall, the G-RUN ENSEMBLE MMM does not seems to show a systematic bias with respect to the full ensemble of the GHM simulations (Figure S1). Additionally, the spatial pattern of average difference in annual runoff between the G-RUN ENSEMBLE members and the G-RUN ENSEMBLE MMM (Figure S2) display much less variability compared to that of the GHMs. This may be explained by the fact that RFM is by construction less sensitive to bias in precipitation and temperature compared to the GHM parametrizations.

Overall, monthly river discharge dynamics (R2) tends to be better captured by the G-RUN ENSEMBLE members than by the GHMs. The large variations in R2 skill observed for the GHMs suggests a higher sensitivity of the GHM parametrizations to the input meteorological forcing. To highlight regions in which the GHMs struggle in reproducing runoff monthly oscillations, Figure S3 illustrates the grid cell wise R2 of each GHM runoff simulations using the G-RUN ENSEMBLE MMM as a reference. The GHMs diverge in the reproduction of monthly dynamics from the G-RUN ENSEMBLE MMM in areas characterized by cold-hydrology processes. On the other hand, most G-RUN ENSEMBLE members agree well with the

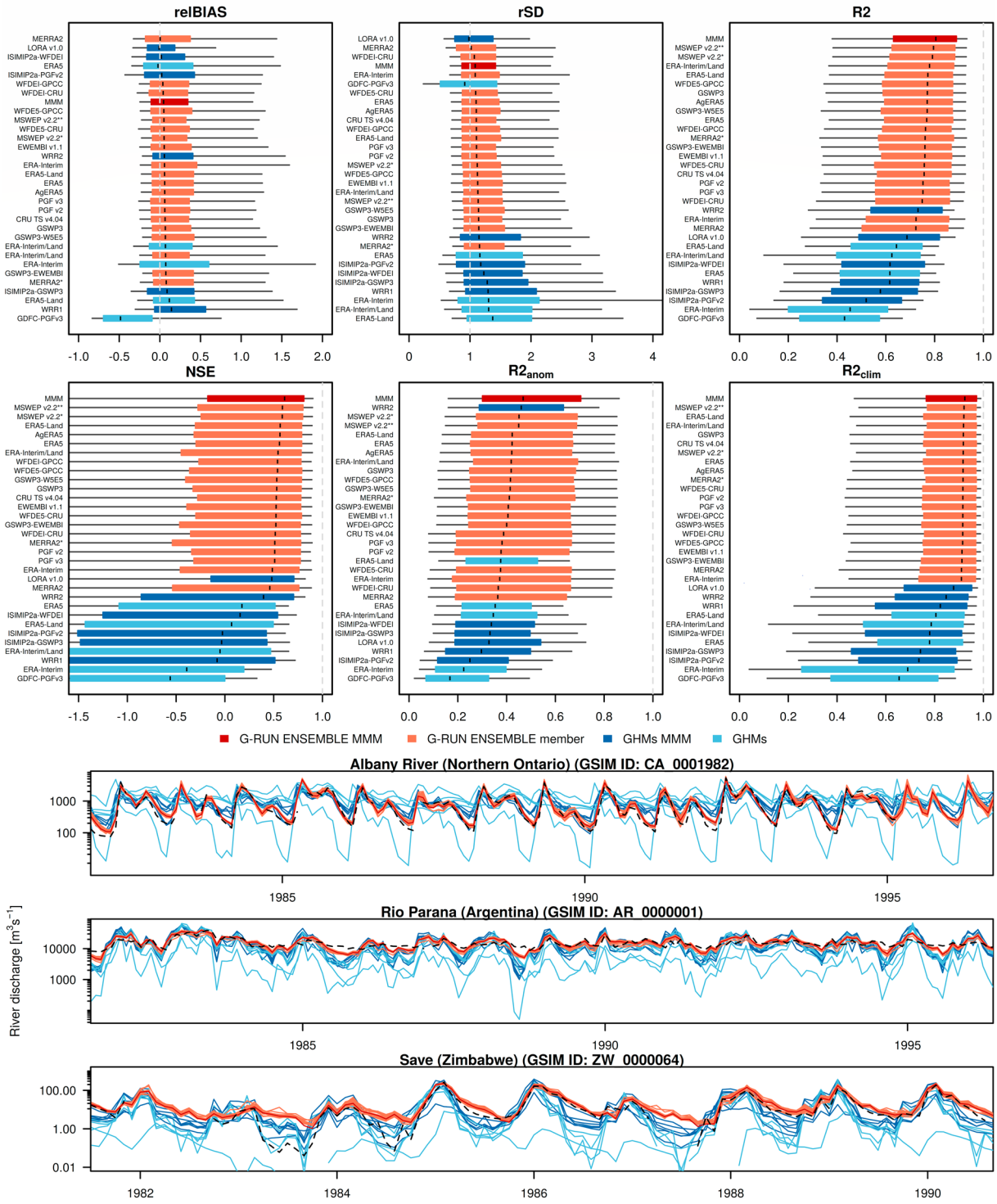

Figure 2. Benchmarking the G-RUN ENSEMBLE against GHM runoff simulations using river discharge observed from 1,205 large river basins. Boxplots: whiskers cover the 0.1 to 0.9 quantiles of the skill metric across all considered basins. The dashed vertical lines indicate the optimal value for the skill metric. The models are ranked by the median skill value. Figure S14 provides the box plots unranked. The *x*-axis of relBIAS is left and right truncated, for rSD it is right truncated and for NSE it is left truncated. Time series: Observed (dashed black line) and predicted (colored) river discharge time series at selected stations. The full set of time series comparison is provided in Data Set S3.

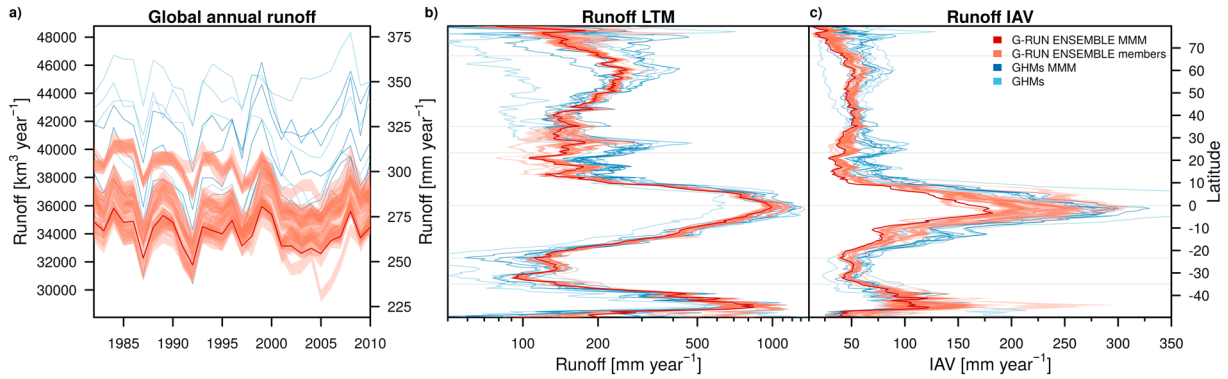

Figure 3. Uncertainty of each G-RUN ENSEMBLE member compared to the spread of the GHM simulations for the period 1981–2010. The orange shaded area around the G-RUN ENSEMBLE member lines shows the distribution range of the 25 realizations for each corresponding forcing data set. (a) Global annual runoff. (b) Latitudinal average of long-term mean (LTM) runoff. (c) Latitudinal average of the interannual variability (IAV) in runoff.

G-RUN ENSEMBLE MMM dynamics (Figure S4), except for the G-RUN ENSEMBLE members forced with MERRA2 and ERA-Interim which disagree in a general manner in the Southern Hemisphere.

The general impression of higher accuracy (Figure 2a) of the G-RUN ENSEMBLE members is confirmed by the NSE skill metrics. Among the GHM reconstructions, LORA v1.0 shows the highest skill, highlighting the benefit of post-processing the GHM runoff simulations. The importance of accurate atmospheric forcing for reproducing the monthly dynamics is highlighted by comparing the best GHM simulations and the best G-RUN ENSEMBLE members: reconstructions forced with MSWEP v2.2, ERA5-Land and ERA5 rank at the top of their respective model category. When considering post-processed atmospheric reanalysis, it also appears that bias correction of precipitation with data from the Global Precipitation Climatology Center (GPCC) (WFDEI-GPCC and WFDE5-GPCC) produces higher skill than bias correcting toward CRU TS (WFDEI-CRU and WFDE5-CRU). In general, the G-RUN ENSEMBLE members show higher accuracy than the GHMs (except the multi-model median of WRR2) in reproducing monthly discharge anomalies ($R2_{anom}$). The GHMs tend to disagree quite strongly from anomalies of the G-RUN ENSEMBLE MMM (Figures S5), while several G-RUN ENSEMBLE members show disagreement from the G-RUN ENSEMBLE MMM in Africa (Figure S6). Concerning the ability to reproduce the seasonal cycle of river discharge ($R2_{clim}$), the performance of the G-RUN ENSEMBLE reconstructions stands out compared to the GHMs. Previous studies already showed that some GHMs struggle in reproducing the seasonality of runoff (Ghiggi et al., 2019; Gudmundsson, Boulange, et al., 2012; Gudmundsson & Seneviratne, 2015).

To disentangle the impact of the meteorological forcing from model uncertainty, Ghiggi et al. (2019) considered a large set of GHMs and RFM estimates that were forced with the same meteorological data. The results highlighted that the lower accuracy of GHMs compared to RFM-based estimates is likely related to issues with GHM structure and parameters. We also note that extensive RFM cross-validation experiments performed in Ghiggi et al. (2019) highlighted that even when removing all observations from large sub-continental regions during training, the RFM predictions remained on average more accurate than the runoff simulations from a large set of GHMs forced with the same meteorological data. Furthermore, it was shown that at the grid cell level, the accuracy of G-RUN is even higher than for large basins. This is related to the nature of the employed approach because RFM is trained at the grid cell level.

4.2. Global Runoff Characteristics at Annual Time Scales

Figure 3 compares global annual runoff characteristics of the G-RUN ENSEMBLE members against the GHM simulations. The G-RUN ENSEMBLE reconstructions estimate the global runoff volume to lie between 32,000 and 39,000 km^3 yr^{-1} (Figure 3a). The impact of sampling uncertainty in the RFM-based estimates (characterized by the 25 realizations for each G-RUN ENSEMBLE member and depicted with an orange shaded area) is smaller than the uncertainty introduced by the atmospheric forcings. This confirms the importance of accounting for uncertainty in the atmospheric forcing data sets which was the motivation for this work. In terms of the latitudinal profile of long-term mean (LTM) runoff rates (Figure 3b), the

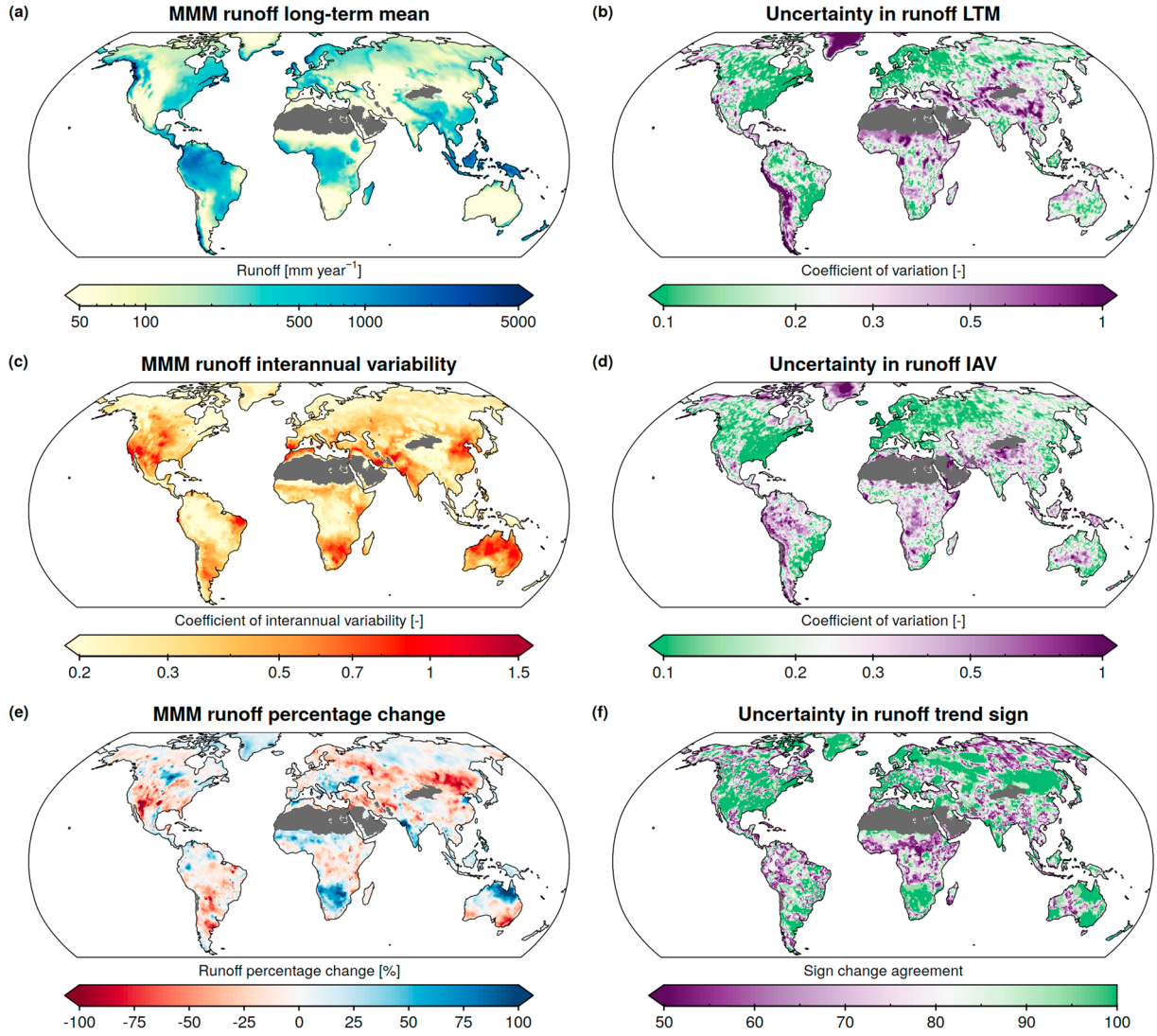

Figure 4. Climatological analysis based on the G-RUN ENSEMBLE for the period 1981–2010. Desert regions with long-term precipitation lower than 100 mm/year are masked in gray. (a) Multi-model median of the runoff long-term mean (LTM) computed for each G-RUN ENSEMBLE member. (b) Coefficient of variation of the ensemble LTM statistics. (c) Multi-model median of the runoff interannual variability (IAV) computed for each G-RUN ENSEMBLE member. (d) Coefficient of variation of the ensemble IAV statistics (e) Multi-model median (MMM) of changes in annual runoff rates, expressed in percentage change over the 30-year period, computed for each G-RUN ENSEMBLE member. (f) Percentage agreement of the runoff trend sign across the 20 G-RUN ENSEMBLE members spanning the period 1981–2010.

spread among the GHMs is much higher than among the G-RUN ENSEMBLE reconstructions (see Figures 4b and S9b for the spatial distribution of uncertainty across the G-RUN ENSEMBLE and the GHMs). The G-RUN ENSEMBLE lies within the GHMs bounds, except in the northern tropics and subtropics. Especially in Southeast Asia, the GHMs simulate higher runoff rates compared to the G-RUN ENSEMBLE (see Figures S1 and S2). Concurrently, in these regions, the G-RUN ENSEMBLE also shows a wider spread in the LTM estimates (see Figure 4b for the LTM spatial uncertainty across the G-RUN ENSEMBLE members), which can partially be related to the uncertainty of precipitation (see Figure S9). The above considerations apply also for the interpretation of the latitudinal profile of the runoff interannual variability (IAV) (Figure 3c). The GHMs tend to have higher annual runoff rates in the northern tropics and subtropics, possibly resulting in higher interannual variability.

Figures S7 and S8 provide a more detailed view on individual G-RUN ENSEMBLE and GHM annual runoff characteristics. As supplementary information, latitudinal profiles, and spatial uncertainty in LTM and IAV of the atmospheric forcing data are reported in Figures S9, S10, S12, and S13.

4.3. Limitations of the G-RUN ENSEMBLE

River discharge observations used for model training have been carefully screened to remove time series presenting unphysical values and possible inhomogeneous behaviors introduced by anthropogenic activities. However, we do not exclude that some river discharge observations that are impacted by human activities might have passed these selection steps, for example, if the magnitude of water abstraction/returns did not alter the monthly hydrograph sufficiently to identify a major changing point or if the time series was not long enough to cover past periods of near-natural streamflow. Because the RFM is solely forced with precipitation and temperature, the G-RUN ENSEMBLE does not explicitly account for effects of human water management such as river flow regulation, water withdrawals, or return flows from groundwater abstraction (Arheimer, Donnelly, & Lindström, 2017; Jaramillo & Destouni, 2015; Nazemi & Wheater, 2015a, 2015b; Veldkamp et al., 2017; Wada et al., 2017; Wada et al., 2010). Therefore, it is our evaluation that the estimates of the G-RUN ENSEMBLE are relatively close to near-natural conditions and would clearly differ with observations in basins heavily impacted by human activities. Additionally, the G-RUN ENSEMBLE is unlikely to provide reliable long-term reconstructions in mountainous areas where an important portion of total monthly runoff comes from glacier melting, since no information on glacier runoff contribution has been fed to the RFM. We also note that the uncertainty of runoff rates in many mountainous regions is likely underestimated due to the large uncertainty in precipitation (see Figure S12) and the fact that the resolution of the meteorological forcings does not capture the sub-grid variability of precipitation and temperature (with consequences for snowmelt volume and timing).

Furthermore, we also note that the design goal of the G-RUN ENSEMBLE differs from the ones of the GHMs. While the G-RUN ENSEMBLE aims at producing the best possible monthly runoff estimates given the available data, the GHMs simultaneously resolve many hydrological fluxes (i.e., evaporation, soil moisture) at a finer hourly/daily temporal resolution, while maintaining balance of water and energy fluxes. Because of the epistemic uncertainty involved in representing all the relevant hydrological processes and the accumulation of input and model errors over time, it appears reasonable that GHM simulations might be characterized by a reduced predictive skill (and a larger ensemble spread) compared to the G-RUN ENSEMBLE.

5. G-RUN ENSEMBLE Applications

Figure 4 provides example applications of the G-RUN ENSEMBLE, offering a global view of freshwater resources. Each analysis reports the multi-model median and is complemented with an uncertainty quantification metric made possible through the multi-forcing nature of the G-RUN ENSEMBLE. A similar analysis for the available GHM simulations is provided in Figure S11. Figure 4a displays the multi-model median of the long-term mean annual runoff estimates over the period 1981–2010. Runoff rates vary by 3 orders of magnitude across the Earth, with the highest rates in the tropics and large mountain ranges and lowest rates in the subtropics and major world deserts such as the Sahara. The uncertainty in long-term runoff rates (Figure 4b) is estimated by computing the robust coefficient of variation (defined as the interquartile range divided by the median) across the G-RUN ENSEMBLE members statistics. The G-RUN ENSEMBLE long-term runoff estimates agree well over North America, Europe, Russia, Australia, Southern Africa, and Eastern South America. Disagreement between the ensemble members occurs along the Andes Mountain Ranges, the mountains of Western United States, Central and East Asia, and in correspondence with uncertainty in the precipitation and temperature forcing data (see Figure S12 and S13). The same analysis performed on the GHMs (Figure S11b) reveals the much higher uncertainty of the long-term mean of the GHMs compared to the G-RUN ENSEMBLE. Figure 4c highlights locations with large fluctuations in freshwater availability across the period 1981–2010, as indicated by values higher than one of the coefficient of interannual variability (defined as the standard deviation of yearly runoff, divided by the long-term mean). Interannual variability is particularly high in regions that have experienced long-lasting droughts in the last decades: for example, the Fertile Crescent (Trigo, Gouveia, & Barriopedro, 2010), Australia (van

Dijk et al., 2013), California (Seager et al., 2015; Swain et al., 2014) and South Africa (Blamey et al., 2018). Figure 4e reveals long-term trends in annual freshwater availability for the period 1981–2010. Following Stahl et al. (2012) trends are computed using the Sen's slope (Sen, 1968) and expressed in percentage change over the 29-year period. Figure 4d shows the percentage agreement on the sign of trends among the G-RUN ENSEMBLE reconstructions. The pattern of change obtained from the GHMs (Figure S11e) is very similar. Note that since the proper choice of trend tests for runoff time series remains the subject of a scientific debate (Chen & Grasby, 2009; Cohn & Lins, 2005; Radziejewski & Kundzewicz, 2004) and because a pixel-wise application of statistical tests in large data sets can yield spurious results (Wilks, 2016) we follow here the best practice for grid cell scale runoff estimation (Marx et al., 2017; Stahl et al., 2010; Stahl et al., 2012; Thober et al., 2018) and refrain from reporting *p*-values. For a comprehensive assessment of runoff trends the reader is referred to Gudmundsson et al. (2019) and Gudmundsson et al. (2021). Note also that observed river flow trends can be subject to significant decadal variability and that magnitude and sign of trends can depend on the considered period (Gudmundsson et al., 2019).

6. Conclusions

This study builds on the established methodology presented by Gudmundsson and Seneviratne (2015) and Ghiggi et al. (2019) and derives monthly runoff estimates from an ensemble of atmospheric forcing data. To this end, a machine learning algorithm is trained with runoff observations from a global collection of in situ streamflow observations separately for each atmospheric forcing data set. The resulting multi-forcing ensemble of runoff reconstructions, termed G-RUN ENSEMBLE, allows us to quantify the uncertainty associated to model input data. This publicly available data set is provided on a 0.5° × 0.5° World Geodetic System 1984 (WGS84) grid, and the reconstructions span a period from 1902 to 2019. The G-RUN ENSEMBLE reconstructions were benchmarked against a comprehensive set of global-scale hydrological model (GHM) simulations, using a large database of river discharge observations as a reference, which can serve as basis also for future global hydrological model intercomparison studies. Overall, the G-RUN ENSEMBLE shows higher accuracy than most GHMs evaluated in this study, especially with respect to the reproduction of the dynamics and seasonality of monthly runoff rates. The analysis also revealed that the accuracy of the reconstruction is dependent on the quality of the forcing data. However, we found that the spread imposed by the atmospheric forcing in the G-RUN ENSEMBLE is small compared to the spread found within a comprehensive ensemble of GHM simulations driven with a smaller subset of possible forcing data. Possible explanations for this behavior include a higher sensitivity of GHMs to biases in the meteorological forcing and a possible accumulation of small errors throughout integration of the governing equations. In summary, the multi-forcing nature of the G-RUN ENSEMBLE allows to quantify the uncertainty associated with the currently available atmospheric forcing data, thereby paving the way for more robust and reliable water resources assessments, climate change attribution studies, hydro-climatic process investigations as well as evaluation, calibration and refinement of the GHMs. We conclude by highlighting that the production of the G-RUN ENSEMBLE would not have been possible without the mobilization of national and international hydrological archives. We call for a continuation of the international efforts to reduce political and technical barriers for the exchange of hydrometeorological data across the scientific community.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The G-RUN ENSEMBLE reconstructions and their associated realizations are publicly available at https://doi.org/10.6084/m9.figshare.12794075 under Creative Commons Attribution 4.0 International License (CC_BY_4.0). Preliminary data repository available at: https://figshare.com/s/ad6d5cdfbba945d93ad2, DOI has been reserved but will be activated once the manuscript will be published. Full data (>400 GB) will be uploaded before final publication when it will be possible to add the full manuscript reference and DOI to the netCDF-4 attributes.

Acknowledgments
L. Gudmundsson and S. I. Seneviratne acknowledge partial support from the European Union's Horizon 2020 Research and Innovation Program (grant agreement 821003 (4C)). V. Humphrey was supported by a Postdoc Mobility fellowship of the Swiss National Science Foundation (P400P2_180784). The authors thank Martin Hirschi and Richard Wartenburger for the help in downloading and preprocessing some of the data sets used in this study. The authors also acknowledge Hylke Beck for kindly providing the MSWEPv2.2 precipitation data set.

References

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., & Siebert, S. (2003). Development and testing of the WaterGAP 2 global model of water use and availability. *Hydrological Sciences Journal*, *48*(3), 317–337. https://doi.org/10.1623/hysj.48.3.317.45290

Alfieri, L., Lorini, V., Hirpa, F. A., Harrigan, S., Zsoter, E., Prudhomme, C., & Salamon, P. (2020). A global streamflow reanalysis for 1980–2018. *Journal of Hydrology X*, *6*, 100049. https://doi.org/10.1016/j.hydroa.2019.100049

Allen, G. H., David, C. H., Andreadis, K. M., Hossain, F., & Famiglietti, J. S. (2018). Global estimates of river flow wave travel times and implications for low-latency satellite data. *Geophysical Research Letters*, *45*(15), 7551–7560. https://doi.org/10.1029/2018GL077914

Arheimer, B., Donnelly, C., & Lindström, G. (2017). Regulation of snow-fed rivers affects flow regimes more than climate change. *Nature Communications*, *8*(1), 62. https://doi.org/10.1038/s41467-017-00092-8

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., et al. (2015). ERA-Interim/Land: A global land surface reanalysis data set. *Hydrology and Earth System Sciences*, *19*(1), 389–407. https://doi.org/10.5194/hess-19-389-2015

Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., & Ziese, M. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth System Science Data*, *5*(1), 71–99. https://doi.org/10.5194/essd-5-71-2013

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., & Schellekens, J. (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrology and Earth System Sciences*, *21*(6), 2881–2903. https://doi.org/10.5194/hess-21-2881-2017

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., et al. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/BAMS-D-17-0138.1

Blamey, R. C., Kolusu, S. R., Mahlalela, P., Todd, M. C., & Reason, C. J. C. (2018). The role of regional circulation features in regulating El Niño climate impacts over southern Africa: A comparison of the 2015/2016 drought with previous events. *International Journal of Climatology*, *38*(11), 4276–4295. https://doi.org/10.1002/joc.5668

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, *95*(9), 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1

Boogaard, H., de Wit, A., Lazebnik, J., Schubert, J., & van der Grijn, G. (2020). *AgERA5: Agrometeorological indicators from 1979 to 2018 derived from reanalysis*. Copernicus Climate Change Service (C3S). https://doi.org/10.24381/cds.6c68c9bb

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, Z., & Grasby, S. E. (2009). Impact of decadal and century-scale oscillations on hydroclimate trend analyses. *Journal of Hydrology*, *365*(1–2), 122–133. https://doi.org/10.1016/j.jhydrol.2008.11.031

Cohn, T. A., & Lins, H. F. (2005). Nature's style: Naturally trendy. *Geophysical Research Letters*, *32*(23), L23402. https://doi.org/10.1029/2005GL024476

Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A., & Arheimer, B. (2020). Lessons learnt from checking the quality of openly accessible river flow data worldwide. *Hydrological Sciences Journal*, *65*(5), 699–711. https://doi.org/10.1080/02626667.2019.1659509

Cucchi, M., Weedon, G., Amici, A., Bellouin, N., Lange, S., Schmied, H. M., et al. (2020). WFDE5: Bias adjusted ERA5 reanalysis data for impact studies. *Earth System Science Data Discussions*, *12*(3), 2097–2120. https://doi.org/10.5194/essd-2020-28

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata. *Earth System Science Data*, *10*(2), 765–785. https://doi.org/10.5194/essd-10-765-2018

Döll, P., Kaspar, F., & Lehner, B. (2003). A global hydrological model for deriving water availability indicators: Model tuning and validation. *Journal of Hydrology*, *270*(1–2), 105–134. https://doi.org/10.1016/S0022-1694(02)00283-4

Dutra, E., Gianpaolo, B., Jean-Christophe, C., Munier, S., Burke, S., Fink, G., & Polcher, J. (2017). *Report on the improved water resources reanalysis Deliverable (D5.2)*. Retrieved from https://www.mendeley.com/catalogue/c454cca9-1593-34c5-a376-a5aa472e21ad/

Fekete, B. M., Looser, U., Pietroniro, A., & Robarts, R. D. (2012). Rationale for monitoring discharge on the ground. *Journal of Hydrometeorology*, *13*(6), 1977–1986. https://doi.org/10.1175/JHM-D-11-0126.1

Fekete, B. M., Robarts, R. D., Kumagai, M., Nachtnebel, H.-P., Odada, E., & Zhulidov, A. V. (2015). Time for in situ renaissance. *Science*, *349*(6249), 685–686. https://doi.org/10.1126/science.aac7358

Fekete, B. M., & Vörösmarty, C. J. (2007). The current status of global river discharge monitoring and potential new technologies complementing traditional discharge measurements. *IAHS Publication*, *309*, 129–136.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., & Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Ghiggi, G., Humphrey, V., Seneviratne, S. I., & Gudmundsson, L. (2019). GRUN: An observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, *11*(4), 1655–1674. https://doi.org/10.5194/essd-11-1655-2019

Gudmundsson, L., Boulange, J., Do, H. X., Gosling, S. N., Grillakis, M. G., Koutroulis, A. G., & Zhao, F. (2021). Globally observed trends in mean and extreme river flow attributed to climate change. *Science*, *371*(6534), 1159–1162. https://doi.org/10.1126/science.aba3996

Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment. *Earth System Science Data*, *10*(2), 787–804. https://doi.org/10.5194/essd-10-787-2018

Gudmundsson, L., Leonard, M., Do, H. X., Westra, S., & Seneviratne, S. I. (2019). Observed trends in global indicators of mean and extreme streamflow. *Geophysical Research Letters*, *46*(2), 756–766. https://doi.org/10.1029/2018GL079725

Gudmundsson, L., & Seneviratne, S. I. (2015). Towards observation-based gridded runoff estimates for Europe. *Hydrology and Earth System Sciences*, *19*(6), 2859–2879. https://doi.org/10.5194/hess-19-2859-2015

Gudmundsson, L., & Seneviratne, S. I. (2016). Observation-based gridded runoff estimates for Europe (E-RUN version 1.1). *Earth System Science Data*, *8*(2), 279–295. https://doi.org/10.5194/essd-8-279-2016

Gudmundsson, L., Seneviratne, S. I., & Zhang, X. (2017). Anthropogenic climate change detected in European renewable freshwater resources. *Nature Climate Change*, *7*(11), 813–816. https://doi.org/10.1038/nclimate3416

Gudmundsson, L., Wagener, T., Tallaksen, L. M., & Engeland, K. (2012). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resources Research*, *48*(11), 1–20. https://doi.org/10.1029/2011WR010911

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., et al. (2011). Large-scale river flow archives: Importance, current status and future needs. *Hydrological Processes*, *25*(7), 1191–1200. https://doi.org/10.1002/hyp.7794

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., & Pappenberger, F. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present. *Earth System Science Data*, *12*(3), 2043–2060. https://doi.org/10.5194/essd-12-2043-2020

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 109. https://doi.org/10.1038/s41597-020-0453-3

He, X., Pan, M., Wei, Z., Wood, E. F., & Sheffield, J. (2020). A global drought and flood catalogue from 1950 to 2016. *Bulletin of the American Meteorological Society*, *101*(5), E508–E535. https://doi.org/10.1175/BAMS-D-18-0269.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 1–51. https://doi.org/10.1002/qj.3803

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., & Dadson, S. J. (2018). Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data. *Journal of Hydrology*, *566*(July), 595–606. https://doi.org/10.1016/j.jhydrol.2018.09.052

Hobeichi, S., Abramowitz, G., Evans, J., & Beck, H. E. (2019). Linear Optimal Runoff Aggregate (LORA): A global gridded synthesis runoff product. *Hydrology and Earth System Sciences*, *23*(2), 851–870. https://doi.org/10.5194/hess-23-851-2019

Humphrey, V., & Gudmundsson, L. (2019). GRACE-REC: A reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data*, *11*(3), 1153–1170. https://doi.org/10.5194/essd-11-1153-2019

Humphrey, V., Gudmundsson, L., & Seneviratne, S. I. (2017). A global reconstruction of climate-driven subdecadal water storage variability. *Geophysical Research Letters*, *44*(5), 2300–2309. https://doi.org/10.1002/2017GL072564

Hunger, M., & Döll, P. (2008). Value of river discharge data for global-scale hydrological modeling. *Hydrology and Earth System Sciences*, *12*(3), 841–861. https://doi.org/10.5194/hess-12-841-2008

Jaramillo, F., & Destouni, G. (2015). Local flow regulation and irrigation raise global human water consumption and footprint. *Science*, *350*(6265), 1248–1251. https://doi.org/10.1126/science.aad1010

Jones, P. W. (1999). First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, *127*(9), 2204–2210. https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2

Kim, H., Watanabe, S., Chang, E. C., Yoshimura, K., Hirabayashi, J., Famiglietti, J., & Oki, T. (2017). *Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1)*. Data Integration and Analysis System (DIAS). https://doi.org/10.20783/DIAS.501

Lange, S. (2019). *EartH2Observe, WFDEI and ERA-Interim data Merged and Bias-corrected for ISIMIP (EWEMBI). V. 1.1*. GFZ Data Services. https://doi.org/10.5880/pik.2019.004

Laudon, H., Spence, C., Buttle, J., Carey, S. K., McDonnell, J. J., McNamara, J. P., et al. (2017). Save northern high-latitude catchments. *Nature Geoscience*, *10*(5), 324–325. https://doi.org/10.1038/ngeo2947

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., et al. (2019). Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resources Research*, *55*(8), 6499–6516. https://doi.org/10.1029/2019WR025287

Marx, A., Kumar, R., Thober, S., Zink, M., Wanders, N., Wood, E. F., & Samaniego, L. (2017). Climate change alters low flows in Europe under a 1.5, 2, and 3 degree global warming. *Hydrology and Earth System Sciences*, *22*(2), 1017–1032. https://doi.org/10.5194/hess-22-1017-2018

Mekonnen, M., & Hoekstra, Y. A. (2016). Four billion people experience water scarcity. *Science Advances*, *2*(2), e1500323. https://doi.org/10.1126/sciadv.1500323

Mengel, M., Treu, S., Lange, S., & Frieler, K. (2020). ATTRICI 1.0 – counterfactual climate for impact attribution. *Geoscientific Model Development Discussions*, *2020*(June), 1–26. https://doi.org/10.5194/gmd-2020-145

Miralles, D. G., Jiménez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., & Fernández-Prieto, D. (2016). The WACMOS-ET project – Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences*, *20*(2), 823–842. https://doi.org/10.5194/hess-20-823-2016

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., et al. (2013). Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrology and Earth System Sciences*, *17*(10), 3707–3720. https://doi.org/10.5194/hess-17-3707-2013

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data Discussions*. https://doi.org/10.5194/essd-2021-82

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

Nazemi, A., & Wheater, H. S. (2015a). On inclusion of water resource management in Earth system models – Part 1: Problem definition and representation of water demand. *Hydrology and Earth System Sciences*, *19*(1), 33–61. https://doi.org/10.5194/hess-19-33-2015

Nazemi, A., & Wheater, H. S. (2015b). On inclusion of water resource management in Earth system models – Part 2: Representation of water supply and allocation and opportunities for improved modeling. *Hydrology and Earth System Sciences*, *19*(1), 63–90. https://doi.org/10.5194/hess-19-63-2015

Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F. (2001). Predicting the Discharge of Global Rivers. *Journal of Climate*, *14*(15), 3307–3323. https://doi.org/10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2

Padrón, R. S., Gudmundsson, L., Decharme, B., Ducharne, A., Lawrence, D. M., Mao, J., et al. (2020). Observed changes in dry-season water availability attributed to human-induced climate change. *Nature Geoscience*, *13*(7), 477–481. https://doi.org/10.1038/s41561-020-0594-1

Radziejewski, M., & Kundzewicz, Z. W. (2004). Detectability of changes in hydrological records/Possibilité de détecter les changements dans les chroniques hydrologiques. *Hydrological Sciences Journal*, *49*(1), 39–51. https://doi.org/10.1623/hysj.49.1.39.54002

Reichle, R. H., Liu, Q., Koster, R. D., Draper, C. S., Mahanama, S. P. P., & Partyka, G. S. (2017). Land surface precipitation in MERRA-2. *Journal of Climate*, *30*(5), 1643–1664. https://doi.org/10.1175/JCLI-D-16-0570.1

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., Van Beek, L. P. H., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(6), E1080–E1089. https://doi.org/10.1073/pnas.1704665115

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., et al. (2017). A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset. *Earth System Science Data*, *9*(2), 389–413. https://doi.org/10.5194/essd-9-389-2017

Seager, R., Hoerling, M., Schubert, S., Wang, H., Lyon, B., Kumar, A., et al. (2015). Causes of the 2011–14 California drought. *Journal of Climate*, *28*(18), 6997–7024. https://doi.org/10.1175/JCLI-D-14-00860.1

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389.

Sheffield, J., Goteti, G., & Wood, E. F. (2006). Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *Journal of Climate*, *19*(13), 3088–3111. https://doi.org/10.1175/JCLI3790.1

Shiklomanov, A. I., Lammers, R. B., & Vörösmarty, C. J. (2002). Widespread decline in hydrological monitoring threatens pan-Arctic research. *Eos, Transactions American Geophysical Union*, *83*(2), 13. https://doi.org/10.1029/2002EO000007

Stahl, K., Hisdal, H., Hannaford, J., Tallaksen, L. M., Van Lanen, H. A. J., Sauquet, E., et al. (2010). Streamflow trends in Europe: Evidence from a dataset of near-natural catchments. *Hydrology and Earth System Sciences*, *14*(12), 2367–2382. https://doi.org/10.5194/hess-14-2367-2010

Stahl, K., Tallaksen, L. M., Hannaford, J., & Van Lanen, H. A. J. (2012). Filling the white space on maps of European runoff trends: Estimates from a multi-model ensemble. *Hydrology and Earth System Sciences*, *16*(7), 2035–2047. https://doi.org/10.5194/hess-16-2035-2012

Swain, D. L., Tsiang, M., Haugen, M., Singh, D., Charland, A., Rajaratnam, B., & Diffenbaugh, N. S. (2014). The extraordinary California drought of 2013-2014: Character, context, and the role of climate change. *Bulletin of the American Meteorological Society*, *95*(9), S3–S7.

Thober, S., Kumar, R., Wanders, N., Marx, A., Pan, M., Rakovec, O., et al. (2018). Multi-model ensemble projections of European river floods and high flows at 1.5, 2, and 3 degrees global warming. *Environmental Research Letters*, *13*(1), 014003. https://doi.org/10.1088/1748-9326/aa9e35

Trigo, R. M., Gouveia, C. M., & Barriopedro, D. (2010). The intense 2007–2009 drought in the Fertile Crescent: Impacts and associated atmospheric circulation. *Agricultural and Forest Meteorology*, *150*(9), 1245–1257. https://doi.org/10.1016/j.agrformet.2010.05.006

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., et al. (2013). The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, *49*(2), 1040–1057. https://doi.org/10.1002/wrcr.20123

Veldkamp, T. I. E., Wada, Y., Aerts, J. C. J. H., Döll, P., Gosling, S. N., Liu, J., et al. (2017). Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. *Nature Communications*, *8*(1), 15697. https://doi.org/10.1038/ncomms15697

Viglione, A., Borga, M., Balabanis, P., & Blöschl, G. (2010). Barriers to the exchange of hydrometeorological data in Europe: Results from a survey and implications for data policy. *Journal of Hydrology*, *394*(1–2), 63–77. https://doi.org/10.1016/j.jhydrol.2010.03.023

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., et al. (2010). Global threats to human water security and river biodiversity. *Nature*, *467*(7315), 555–561. https://doi.org/10.1038/nature09440

Wada, Y., Bierkens, M. F. P., de Roo, A., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., et al. (2017). Human–water interface in hydrological modelling: Current status and future directions. *Hydrology and Earth System Sciences*, *21*(8), 4169–4193. https://doi.org/10.5194/hess-21-4169-2017

Wada, Y., van Beek, L. P. H., van Kempen, C. M., Reckman, J. W. T. M., Vasak, S., & Bierkens, M. F. P. (2010). Global depletion of groundwater resources. *Geophysical Research Letters*, *37*(20). https://doi.org/10.1029/2010GL044571

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences*, *111*(9), 3228–3232. https://doi.org/10.1073/pnas.1312330110

Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., et al. (2018). Evapotranspiration simulations in ISIMIP2a—Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. *Environmental Research Letters*, *13*(7), 075001. https://doi.org/10.1088/1748-9326/aac4bb

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*(9), 7505–7514. https://doi.org/10.1002/2014WR015638

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., et al. (2011). Creation of the WATCH forcing data and its use to assess global and regional reference crop evaporation over land during the twentieth century. *Journal of Hydrometeorology*, *12*(5), 823–848. https://doi.org/10.1175/2011JHM1369.1

Wilks, D. S. (2016). The stippling shows statistically significant grid points. *Bulletin of the American Meteorological Society*, *97*(12), 2263–2274.