Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-386
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 7 August 2018

Hydrology and Earth System Sciences
Open Access
Discussions

Linear Optimal Runoff Aggregate (LORA): A global gridded synthesis runoff product

Sanaa Hobeichi[1,2], Gab Abramowitz[1,3], Jason Evans[1,3] and Hylke E. Beck[4]

[1] Climate Change Research Centre, University of New South Wales, Sydney, NSW 2052, Australia
[2]ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, NSW 2052, Australia
[3]ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW 2052, Australia
[4] Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544, USA

Correspondence to: Sanaa Hobeichi (s.hobeichi@student.unsw.edu.au)

Abstract

No synthesized global gridded runoff product, derived from multiple sources, is available despite such a product being useful to meet the needs of many global water initiatives. We apply an optimal weighting approach to merge runoff estimates from hydrological models constrained with observational streamflow records. The weighting method is based on the ability of the models to match observed streamflow data while accounting for error covariance between the participating products. To address the lack of observed streamflow for many regions, a dissimilarity method was applied to transfer the weights of the participating products to the ungauged basins from the closest gauged basins using dissimilarity between basins in physiographic and climatic characteristics as a proxy for distance. We perform out-of-sample tests to examine the success of the dissimilarity approach and we confirm that the weighted product performs better than its 11 constituents products in a range of metrics. Our resulting synthesized global gridded runoff product is available at monthly time scales, and includes time variant uncertainty, for the period 1980 – 2012 on a 0.5º grid. The synthesized global gridded runoff product broadly agrees with published runoff estimates at many river basins, and represents well the seasonal runoff cycle for most of the globe. The new product, called Linear Optimal Runoff Aggregate (LORA), is a valuable synthesis of existing runoff products and will be freely available for download on geonetwork.nci.org.au.

Introduction

Runoff is the horizontal flow of water on land or through soil before it reaches a stream, river, lake, reservoir or other channels. It has been widely used as a metric for droughts (Shukla and Wood, 2008; van Huijgevoort et al., 2013; Bai et al., 2014; Ling et al., 2016) and to understand the effects of climate change on the hydrological cycle (Ukkola et al., 2016; Zhai and Tao, 2017). Accurate estimates of runoff are critical to inform climate change adaptation strategies, to guide appropriate water management in agriculture (Nyamadzawo et al., 2012), and to enable the assessment of the impact of anthropogenic activities on ecosystems (Vörösmarty et al., 2010), yet direct measurement of runoff at large scales is simply not possible.

While runoff observations do not exist, direct streamflow or river discharge observations - basin integrated runoff - have been archived in many databases. The most comprehensive international streamflow database is the Global Runoff Data Base (GRDB; www.bafg.de), which consists of daily and monthly quality-controlled streamflow records from more than 9500 gauges across the globe. Geospatial Attributes of Gages for Evaluating Streamflow version II (GAGES-II) represents another noteworthy streamflow database, consisting of daily quality-controlled streamflow data from over 9000 US gauges. Hydrological and land surface models are capable of producing gridded runoff estimates for any region across the globe (Sood and Smakhtin, 2015; Liu et al., 2016; Kauffeldt et al., 2016). However, these runoff estimates suffer from uncertainties due to shortcomings in the model structure and parameterization and the meteorological forcing data (Beven, 1989; Beck, 2017a). There are various ways to use streamflow observations for improving the runoff outputs from these models. The conventional approach consists of model parameter calibration using locally observed streamflow data (see review by Pechlivanidis et al., 2011). Another widely used method is through regionalization; that is, the transfer of knowledge (e.g., calibrated parameters) from gauged basins to ungauged basins (see review by Beck et al., 2016). In contrast, several other studies attempted to correct the runoff outputs directly rather than the model parameters, for example by bias-correcting model runoff outputs based on streamflow observations (see review by Ye et al., 2014), or by combining or weighting ensembles of model outputs to obtain improved runoff estimates (e.g., Aires, 2014). There are, however, relatively few continental- and global-scale efforts to improve model estimates using observed streamflow.

A broad array of gridded model-based runoff estimates are freely available, including ECMWF's Interim reanalysis (ERA-Interim ; Dee et al., 2011), NASA's Modern Era Retrospective-analysis for Research and Applications (MERRA) Land (Reichle et al., 2011), the Climate Forecast System Reanalysis (CFSR; Tomy and Sumam, 2016), the second global soil wetness project (GSWP2; Dirmeyer et al., 2006), the Water Model Intercomparison Project (WaterMIP; Haddeland et al., 2011), and the Global Land Data Assimilation System (GLDAS; Rodell et al., 2004). Recently, the eartH2Observe project has made available two ensembles (tier-1 and -2) of state-of-the-art global hydrological and land surface model outputs (http://www.earth2observe.eu/; Beck et al., 2017a; and Schellekens et al., 2017). Although these model simulations represent the only time varying gridded estimates of runoff at the global scale, they are subject to considerable uncertainties, resulting in large differences in runoff simulated by the models. Many studies have therefore evaluated and compared the gridded runoff models (see overview in Table 1 of Beck et al., 2017a).

Despite the demonstrated improved predictive capability of multi-model ensemble approaches (Sahoo et al., 2011; Pan et al., 2012; Bishop and Abramowitz, 2013; Mueller et al., 2013; Munier et al., 2014; Aires, 2014; Rodell et al., 2015; Jiménez et al., 2017; Hobeichi et al., 2018; Zhang et al., 2018), very little has been done to utilise this range of model simulations toward improved runoff estimates. This paper implements the weighting and rescaling method introduced in Bishop and

Abramowitz (2013) and Abramowitz and Bishop (2015) to derive a monthly 0.5° global synthesis runoff product. Briefly summarized, we use a bias correction and weighting approach to merge 11 state-of-the-art gridded runoff products from the eartH2Observe project, constrained by observed streamflow from a variety of sources. This approach also provides us with corresponding uncertainty estimates that are better constrained than the simple range of modelled values. For ungauged regions we employ a dissimilarity method to transfer the product weights to the ungauged basins from the closest basins using dissimilarity between basins as a proxy for distance. Such a synthesis product is in line with the multi-source strategy of Global Energy and Water EXchanges (GEWEX; Morel, 2001) and NASA's Making Earth Science Data Records for Use in Research Environments (MEaSUREs; Earthdata, 2017) initiatives and is particularly useful for studies that aim to close the water budget at the grid scale.

Sections 2.1 describes the observed streamflow data. Section 2.2 presents the participating datasets used to derive the weighted runoff product. Section 2.3 details the weighting method implemented in the gauged basins, while Section 2.4 focuses on the ungauged basins. Section 2.5 examines the approach used to derive the global runoff product. We then present and discuss our results in Section 3 and 4 before concluding.

Data and Methods

2.1 Observed streamflow data

We used observed streamflow from the following four sources: (i) the US Geological Survey (USGS) Geospatial Attributes of Gages for Evaluating Streamflow (GAGES)-II database (Falcone et al., 2010); (ii) the Global Runoff Data Base (GRDB; http://www.bafg.de/GRDC/); (iii) the Australian Peel et al. (2000) database; and (iv) the global Dai (2016) database. We discarded duplicates and from the remaining set of stations discarded those satisfying at least one of the following criteria: (i) basin area <8000 km^2 (fewer than three 0.5° grid cells); (ii) record length <5 y in the period 1980–2012 (not necessarily consecutive); and (iii) low observed streamflow (i.e. around 0) that does not represent the total runoff across the basins due to significant anthropogenic activities. A river basin was identified with significant anthropogenic activities if it has > 20% irrigated area using the Global Map of Irrigation Areas (GMIA-Version 4.0.2; Siebert et al., 2007) or has > 20% classified as "Artificial surfaces and associated areas" according to the Global Land Cover Map (GlobCover-Version 2.3; Bontemps et al., 2011). In total 596 stations (of which 20 are nested in the basins of other stations) were found to be suitable for the analysis (Fig. 1).

2.2 Simulated runoff data

To derive the global monthly 0.5° synthesis runoff product, we used 11 total runoff outputs (from eight different models) and seven streamflow outputs (from six different models) produced as part of tiers 1 and 2 of the eartH2Observe project (available via ftp://wci.earth2observe.eu/). The models and their available variables are presented in Table 1. For tier 1 of eartH2Observe, the models were forced with the WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset (Weedon et al., 2014) corrected using the Climatic Research Unit Timeseries dataset (CRU-TS3.1; Harris et al., 2014). For tier 2, the models were forced using the Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset (Beck et al., 2017b). The runoff and streamflow values are provided in kg $m^{-2}s^{-1}$ and m^3 s^{-1}, respectively. For consistency, the runoff outputs with resolution <0.5° were resampled to 0.5° using bilinear interpolation. In some cases, the river network employed by the model did not correspond with the stream gauge location, in which case we manually selected the grid cell that provided the best match with the observed streamflow.

The runoff outputs were only used if no streamflow output was available and only in basins smaller than 100,000 km^2. To make the runoff data consistent with the streamflow data, we integrated the runoff over the basin areas (termed Ragg, units m^3 s^{-1}). Thus, for basins smaller than 100,000 km^2 the synthesis product was derived from 11 model outputs, whereas for basins larger than 100,000 km^2 the synthesis product was derived from seven outputs.

2.3 Implementing the weighting approach at the gauged basins

At each gauged basin, we built a linear combination μ_{q} of the participating modelled streamflow datasets x (i.e. Ragg in small basins and modelled streamflow, q, in large basins) that minimized the mean square difference with the observed streamflow Q at that basin such that: $\mu_{\mathrm{q}}^{\mathrm{j}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} w_{\mathrm{k}}(x_{\mathrm{k}}^{\mathrm{j}} - b_{\mathrm{k}})$ where $j \in [1, \mathrm{J}]$ are the time steps and $k \in [1, \mathrm{K}]$ represent the participating models, x_k^j (i.e., $\mathrm{Ragg}_{\mathrm{k}}^{\mathrm{j}}$ in small basins and $q_{\mathrm{k}}^{\mathrm{j}}$ in large basins) is the value of the participating dataset in m^3 s^{-1} at the j^{th} time step of the k^{th} participating model, the bias term b_{k} is the mean error of x_{k} in m^3 s^{-1}. The set of weights w_{k} provides an analytical solution to the minimization of $\sum_{\mathrm{j}=1}^{\mathrm{J}}(\mu_{\mathrm{q}}^{\mathrm{j}} - Q^{\mathrm{j}})^2$, where Q^j is the observed streamflow at the j^{th} time step (for derivation see Bishop and Abramowitz (2013)).

We then derived the weighted runoff dataset by applying the computed weights on the bias corrected runoff estimates of the participating models. The weighted runoff dataset is expressed as:

$$\mu_{\mathrm{r}}^{\mathrm{j}} = \sum_{\mathrm{k}=1}^{\mathrm{K}} w_{\mathrm{k}}(r_{\mathrm{k}}^{\mathrm{j}} - \mathrm{b'}_{\mathrm{k}})$$

Where $r_{\mathrm{k}}^{\mathrm{j}}$ is the value of runoff estimate in $\mathrm{kg\,m^{-2}s^{-1}}$ of the k^{th} participating model at the j^{th} time step and $\mathrm{b'}_{\mathrm{k}}$ is its runoff bias in $\mathrm{kg\,m^{-2}s^{-1}}$.

To calculate the runoff bias $\mathrm{b'}_{\mathrm{k}}$, we assumed that for each model k and at each time j the bias ratio of a model (defined as the ratio of the model error to the simulated magnitude) is the same for streamflow and runoff estimates Eq. (1). In small basins, the bias ratio of modeled streamflow was calculated by using $\mathrm{Ragg}_{\mathrm{k}}^{\mathrm{j}}$ instead of the modeled streamflow $q_{\mathrm{k}}^{\mathrm{j}}$ Eq. (2).

$$\left[\frac{q_{\mathrm{k}}^{\mathrm{j}} - \mathrm{Q}^{\mathrm{j}}}{q_{\mathrm{k}}^{\mathrm{j}}} = \frac{\mathrm{b'}_{\mathrm{k}}}{r_{\mathrm{k}}^{\mathrm{j}}}\right]_{\mathrm{basin}} \qquad (1)$$

$$[\frac{\mathrm{Ragg}_{\mathrm{k}}^{\mathrm{j}} - Q^{\mathrm{j}}}{\mathrm{Ragg}_{\mathrm{k}}^{\mathrm{j}}} = \frac{\mathrm{b'}_{\mathrm{k}}}{r_{\mathrm{k}}^{\mathrm{j}}}]_{\mathrm{basin}} \qquad (2)$$

To avoid over-fitting when applying the weighting approach, we limited the number of participating models so that the ratio of number of records to number of models does not fall below 10. As a result of this, when required, we discarded the models that had the highest bias (i.e. left terms in Eq (1, 2)) until the threshold was met. The weighting and the bias correction occasionally resulted in negative runoff values, we replaced any negative values with zero.

We implemented the ensemble dependence transformation process detailed in Bishop and Abramowitz (2013) to compute the gridded time-variant uncertainty associated with the derived runoff estimates, following the same approach as in Hobeichi et al. (2018). For any particular gauged basin, we first calculated the spatial aggregate of our weighted runoff estimate, $\mathrm{R}_{\mathrm{agg}_{\mu}}$, then quantified $\mathrm{s_r}^2$, the error variance of $\mathrm{R}_{\mathrm{agg}_{\mu}}$ with respect to the observed streamflow Q over time and space. We then transformed the constituent modelled estimates so that their variance about $\mathrm{R}_{\mathrm{agg}_{\mu}}$ at a given time step $\sigma_{\mathrm{r}}^{2\mathrm{j}}$, averaged over all time steps where we have available streamflow data for the current basin, is equal to $\mathrm{s_r}^2$. This transformed ensemble provides us with uncertainty estimates that (a) are varying in time and space, and (b) accurately reflects our ability to reproduce the observed streamflow. It provides a much more defensible uncertainty estimate than simply calculating the standard deviation of the involved products. We then used $\sqrt{\sigma_{\mathrm{r}}^{2\mathrm{j}}}$ as the spatially and temporally varying estimate of uncertainty standard deviation, which we will refer to below simply as 'uncertainty'. For more details about how this technique was implemented we refer readers to Hobeichi et al. (2018).

2.4 Deriving runoff estimates at the ungauged river basins

Implementing the weighting approach requires observed streamflow to constrain the weighting, which we do not have at ungauged river basins (defined in section 2.1). To address this, we used the modelled and observed streamflow from the three most similar gauged river basins, based on pre-defined physical and climatic characteristics, to derive model weights at each ungauged basin. The selected gauged river basins served as donor basins to the ungauged receptor basins. We then implemented the weighting technique on the ensemble of 11 (in small basins) or eight (in large basins) model outputs by matching Ragg calculated across the selected donor basins with the observed streamflow. Finally, we transferred the weights and bias ratios computed at the donor basins to the receptor basin and subsequently computed the associated uncertainty values.

Most of the gauged river basins were classified as donor basins. Some, however, were excluded from being donors where we found (based on Ragg or modeled streamflow time series and metric values) that none of the models was able to simulate the streamflow dynamics. These basins are mainly located in areas of natural lakes, in mountainous areas covered with snow, or in wet regions with intense rainfall. We therefore (subjectively) decided that those excluded basins should be assigned to a "non–donor and non–receptor" category.

We applied the method presented in Beck et al. (2016) to calculate a similarity index *S* between a donor basin *a* and a receptor basin *b* expressed as:

$$S_{\mathrm{a,b}} = \sum_{\mathrm{p}=1}^{7} \frac{|Z_{\mathrm{p,a}} - Z_{\mathrm{p,b}}|}{\mathrm{IQR_p}} \qquad (3)$$

Where *p* denotes the climatic and physiographic characteristics as in Table 4 of Beck et al. (2016). This includes aridity index, fractions of forest and snow cover, soil clay content, surface slope, and annual averages of precipitation and potential evaporation. $Z_{\mathrm{p,a}}$ and $Z_{\mathrm{p,b}}$ are the values of the characteristic *p* at donor and receptor basins, respectively. $\mathrm{IQR_p}$ is the interquartile range of characteristic *p* calculated over the land surface, excluding deserts (defined by an aridity index > 5, see Table 4 of Beck et al. (2016)) and areas with permanent ice (defined by climate zones Tundra, Subarctic and Ice cap using a simplified climate zones map created by the Esri Education Team for ArcGIS online (World Climate Zones – Simplified; Esri Education Team, 2014)). From Eq. 3 it follows that the most similar donor *a* to a receptor *b* is the one that has the lowest index value with basin *b*. We applied this approach to identify the 3 most similar donors for every receptor basin.

In very large basins, physiographic and climatic heterogeneity can result in misleading basin-mean averages. We therefore excluded highly heterogeneous basins from the list of donors and classified them as 'non-donor and non-receptor' basins, and also broke up large heterogeneous receptor basins by climate groups into smaller basin zones and then treated them as separate basins to effectively receive sets of weights and bias ratios from the donor basins to the separate parts. Here we defined large heterogeneous basins as basins with areas greater than 1,000,000 km^2 and covering climate zones that belong to at least two groups of 1) Tropical Wet, 2) Humid continental, Humid subtropical, Mediterranean and Marine, 3) Tropical Dry, Semi–arid and Arid, 4) Tundra, Subarctic and Ice cap and 5) Highlands. Climate classification is based on the simplified climate zones map (World Climate Zones te zones map; Esri Education Team, 2014) defined above. Figure 2 shows the spatial coverage of the donor basins, receptor basins and non-donor and non-receptor basins, and Fig. 3 summarizes the steps carried out to derive the weighted runoff product for the global land.

2.5 Out-of-sample testing

To test that this approach is producing a runoff estimate at receptor basins (using transferred weights from the most similar gauged basins) that is better than any of the individual models, we performed an out-of-sample test. In this test, we selected a gauged basin and treated it as a receptor basin, constructing model weights by using the three most similar donor basins. We could then compare: (a) observed streamflow; (b) the in-sample weighted product ($\mathrm{WP_{in}}$) derived by using observed streamflow for this basin to weight models; (c) an out-of-sample weighted product ($\mathrm{WP_{out}}$) derived by constructing the weighting at the three most similar basins, and; (d) the individual model estimates at each basin. We calculated four metrics of performance for $\mathrm{WP_{in}}$, $\mathrm{WP_{out}}$ and each of the 11 datasets: Mean Square Error MSE=mean(Ragg – observed streamflow)^2; Mean Bias=mean| Ragg – observed streamflow |; Correlation COR=corr(observed streamflow, Ragg) and Standard Deviation (SD) difference= $\sigma_{\mathrm{Ragg}} - \sigma_{\mathrm{observed\ streamflow}}$. We repeated the out-of-sample test for all the gauged basins (donor basins and non-donor and non-receptor basins).

We displayed the results of the out-sample-test by showing the percentage performance improvement of WP_{out} compared to WP_{in} and each individual model, yielding 12 different values of performance improvement. If the approach is succeeding, we expect that both WP_{out} and WP_{in} will perform better than any of the models used in this study, and also WP_{in} should be in better agreement with the observed streamflow when compared to WP_{out}.

We used box and whisker plots to show the results of performance improvement of WP_{out} calculated relative to WP_{in} and the 11 datasets across all the gauged basins. The lower and upper hinges of a boxplot represent the first (Q_1) and third (Q_3) quartiles respectively of the performance improvement results and the line inside the boxplot shows the median value. The extreme of the lower whisker represents the maximum of 1) min(dataset) and 2) (Q_1 - IQR), while the extreme of the upper whisker is the minimum of 1) max(dataset) and 2) (Q_3 + IQR)), where IQR represents the interquartile range (i.e. Q_3 - Q_1) of the performance improvement results. A median line located above the 0 axis is an indication that the out of sample weighting offers an improvement in more than half of the basins.

3 Results

The results for the out-of-sample test are displayed in the box and whisker plots presented in Fig. 4 (a - d).
The MSE and Mean bias plots in Fig. 4 (a and d) indicate that across almost all the gauged basins WP_{out} performs better than each of the individual models. Similarly, the COR plot in Fig. 3 (c) shows that the out-of-sample weighting has in fact improved the correlation with observational data across almost all the gauged basins. The SD difference plot (Fig. 4 (b)) shows a significant improvement of WP_{out} relative to the models, but the number of basins that benefit from this improvement decreased, perhaps because the variability of the individual members of the weighting ensemble is not necessarily temporally coincident at all the basins, resulting in decreased variability. The negative performance improvement of WP_{out} relative to WP_{in} across all metrics (first boxplot, Fig. 4 (a-d)) indicates that the weighting performs better in-sample than out-of sample, which is to be expected. Critically though, the fact that the weighting delivers improvement over all models when the weights are transferred from similar basins indicate that the dissimilarity technique is succeeding and can be effectively used at the ungauged basins by feeding the weighting with data from the most similar basins with streamflow observations.

Based on the improvement that the weighting approach implemented in both gauged and ungauged basins offers over Ragg estimates computed for 11 individual model runoff estimates, in terms of MSE, SD difference, COR and Mean Bias against observed streamflow data, we now present details of the mosaic of the individual weighted runoff estimates derived across all the basins that we name LORA. At the gauged basins, the weighting was trained with the Ragg of the modelled runoff at the individual basins and constrained with the observed streamflow. At ungauged basins, the dissimilarity approach was first implemented to find the three most similar basins, then the weighting was trained on the combined datasets from these three basins. Subsequently, weights were transferred to the ungauged basins and applied to combine the runoff estimates at the individual basins.

The eight modelled runoff datasets listed in Table 1 as part of the tier1 ensemble were recently included in a global evaluation by Beck et al. (2017a). In their analysis, they computed a summary performance statistic that they termed OS by incorporating several long-term runoff behavioural signatures defined in Table 3 of Beck et al. (2017a) and found that the mean of runoff estimates from four models only (LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) performed the best in terms of $\overline{OS}$ (i.e. mean of OS over all the basins included in their study) relative to each individual modelled runoff estimates and the mean of all the modelled runoff estimates. In this study, we calculated the mean runoff from the four best products found by Beck et al. (2017a), that is (LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG. Hereafter, we refer this as "Best4", and we calculated four statistics (RMSE, SD difference, COR and Mean bias defined here as mean(dataset-obs)) for Ragg computed from LORA, Best4 and each of the 11 runoff datasets across all the gauged basins. The boxplots in Fig. 5 (a-d) display the results.

The RMSE plot in Fig. 5(a) shows that LORA has the lowest RMSE values with the observed streamflow. All of the component models exhibit a similar performance in RMSE. Similarly, LORA has overall the least SD difference with observations (Fig. 5 b) across more than half of the basins. The Mean bias plot in Fig. 5(d) shows a non-significant positive bias in LORA relative to the observation at the majority of the basins. Best4, HBV-SIMREG, PCR-GLOBWB and particularly LISFLOOD exhibit a positive mean bias across most of the basins but with much higher bias magnitude compared to that of LORA. HTESSEL and SURFEX estimates from both tiers (i.e tier1 and tier2) together with JULES (tier2) and WGAP3 show negative and positive bias distributed evenly across the basins. LORA shows the highest temporal correlation with the observed streamflow at more than half of gauged basins (Fig. 5 (c)). The low RMSE and Mean bias values relative to the other estimates is partly due to the bias correction applied before the weighting. While all the performance metrics calculated here show that LORA outperforms Best4, these metrics do not allow us to assess how well LORA performs in terms of bias in the runoff timing, replicating the peaks or representing quick runoff, with the exception of the correlation metric. These aspects were studied in more detail in Beck et al. (2017a) and showed that Best4 performs well in these performance metrics.

All the models involved in deriving LORA with the exception of HBV-SIMREG were found in the study of (Beck et al., 2017a) to show early spring snowmelt peak and an overall significant underestimation of runoff in the snow-dominated basins. To see how well LORA performs at high latitudes, we examined the gauged basins located at higher latitudes (>60°) and we calculated two statistics – COR and mean bias – as in Fig. 5 (c-d) but this time for the snow-dominated basins only. We display the results in Fig. 6.

The temporal correlation plot in Fig. 6 (a) shows that LORA is in better agreement with observed streamflow at snow-dominated basins compared to the ensemble of all the gauged basins on the globe (Fig. 5 (c)) with an overall average improvement of 7%. Similarly, HBV-SIMREG shows an improved correlation with the observed streamflow at snow-dominated basins with an average improvement of 14%, this agrees with the results reported by Beck et al. (2017a) who attributed the improved performance of HBV-SIMREG in snow-dominated regions to a snowfall gauge undercatch correction. The overall performance of Best4 and LISFLOOD do not change in terms of spatial correlation; on the contrary, all the remaining products show a degraded performance. Figure 6 (b) shows that LORA exhibits small biases across snow-dominated basins relative to participating models. Conversely, with the exception of LISFLOOD, all the tier1 products including Best4 show a negative mean bias across more than half of the snow-dominated basin, in particular HTESSEL, JULES, SURFEX and W3RA show a large negative bias at most of these basins. This agrees with the negative bias found in the study of Beck et al. (2017a) in all tier1 products except LISFLOOD. These results indicate that LORA is likely to slightly overestimate runoff in high latitudes whereas all tier1 products with the exception of LISFLOOD tend to underestimate runoff in these regions, and that this underestimation is larger for HTESSEL, JULES, SURFEX and W3RA. Tier2 products show both positive and negative bias across the basins. Their bias is of a lower magnitude than that found in tier1 products. That is probably because the forcing precipitation used to derive tier 2 outputs (i.e. MSWEP) has less biases than that used to derive tier1 estimates (i.e. WFDEI corrected using CRU-TS3.1). We also calculated the two metrics, SD difference and mean bias as in Fig. 5 (a and b), but we found no noticeable differences in the performance of any of the products relative to that found globally in Fig. 5 (a and b). The results displayed in Fig. 5 and Fig. 6 are discussed further below.

We calculated the seasonal relative uncertainty expressed as the ratio of average uncertainty to mean runoff (i.e. $\frac{\text{mean runoff uncertainty}}{\text{mean runoff}}$) for the period 1980 – 2012. This metric is intended to show some indication of the reliability of the derived runoff, with results displayed in Fig. 7. Regions in red show grid cells that satisfy $\frac{\text{mean runoff uncertainty}}{\text{mean runoff}} < 1$, while those shown in yellow are regions where the value of mean runoff uncertainty are larger than the value of the associated mean runoff itself. Regions in blue are grid cells that have a zero mean runoff and hence an undetermined relative uncertainty. The global maps in Fig. 7 show a consistent low reliability in Sahel, Indus basin, Parana, the semi-arid regions of Eastern Argentina, Doring basin in South Africa, red river sub-basin of the Mississippi, Burdekin and Fitzroy basins in North-East Australia and many regions of the Arab Peninsula. The areas at the higher latitudes in Asia and North America show high reliability during Jun-Jul-Aug and low reliability during the rest of the year. Parts of Madeiry sub-basin – a major sub-basin of the Amazon – show low reliability during June-Nov. The basins in Central America show high reliability in all seasons except in Mar-May while River basins in Somalia show low reliability during the austral summer and winter. River basins in the far east show low reliability in spring and autumn and a higher reliability in winter and summer.

Figure 8 displays the seasonal cycles of Ragg for LORA and Best4 and the observed streamflow over 11 major river basins. To generate this plot, we calculated the average Ragg for each month over the period of availability of observed streamflow. The shaded regions represent the range of uncertainty aggregates associated with the derived runoff. In the Amazon basin, LORA overestimates runoff in the wet season and underestimates it in the dry season, but the observed streamflow during the dry season still lies within the error bounds of LORA. LORA shows good agreement with the observed cycle in the Mississippi. In the Niger and Murray-Darling basins, while LORA overestimates the observed streamflow, it shows a much better agreement compared to Best4 which strongly overestimates runoff. In the Parana basin, LORA underestimates the observed streamflow in all seasons except summer. In the subarctic basins, LORA shows different behavior within the individual basins. In Pechora and Olenek, LORA represents well the seasonal cycle and the magnitude of runoff, whereas in the Amur, Lena and Yenisei, LORA shows an early shift of the runoff peak and an overall overestimation of runoff. In the Indigirka, LORA overestimates the spring peak, but the observed seasonal cycle lies within the error bounds.

Finally, we compared our mean annual runoff (mm/year) with those estimated by a well-known land surface hydrological model the Variable Infiltration Capacity (VIC; Liang et al., 1994) model in the study of Zhang et al. (2018) over comparable temporal and spatial scale for 16 large basins chosen from different climate zones on the globe. The mean annual runoff was computed over the period 1984 - 2010 instead of 1980 – 2012 to maximize the temporal agreement with the study of Zhang et al. (2018).

Table 2 shows that for some basins VIC and LORA agree well in estimating mean annual runoff (i.e. difference between LORA and at least one of VIC and VIC adjusted for budget closure <10%). This threshold is met in the Amazon, Columbia, Congo, Danube, Mackenzie and Mississippi. The basins that show a larger difference between VIC and LORA but show that VIC estimates lie within the uncertainty bounds of LORA (i.e. between LORA-uncertainty and LORA+uncertainty) include the Indigirka, Olenek, Parana, Pechora, Yenisei and Yukon. Large discrepancies between VIC and LORA are found in Lena and the Murray-Darling.

4 Discussion

The results of the out-of-sample test suggest that deriving runoff estimates in an ungauged basin by training the weighting with streamflow data from similar basins - in terms of climatic and physiographic characteristics - is successful. While the runoff product derived by using weights from external basins outperforms the runoff estimates from the individual models, the weighted runoff derived in-sample offers overall even more capable runoff estimates.

It follows from Fig. 2 and Fig. 7 that the runoff values computed over dry climates tend to be less reliable than those in other regimes. This is perhaps due to the biases in the WFDEI precipitation forcing that intensify in the arid and semi-arid regions and propagate in the simulated runoff (Beck et al., 2017a). Also, due the lower density of gauged basins in the arid and semi-arid climates compared to other regimes, receptor basins are dominant over dry climates, which reduces the skill of the weighting to produce good runoff estimates. This is also in line with our conclusions from Fig. 3 that the weighting provides more reliable results in the gauged basins.

All the tier1 model outputs involved in this study with the exception of HBV-SIMREG were found by Beck et al. (2017a) to show early spring snowmelt in the snow-dominated basins. Both the Yenisei and the Lena are large basins (2.6 and 2.4 million km^2, respectively), and hence – as noted in Sect. 2.2 – only models that had estimates of both streamflow and runoff were used to derive LORA at these basins, and therefore HBV-SIMREG – whose inclusion would have improved the weighting - was excluded. Beck et al. (2017a) also found that LISFLOOD has the best square root-transformed mean annual runoff among the tier1 datasets and perfoms well in terms of temporal correlation in all climates, this agrees with the high temporal correlation of LISFLOOD seen in Fig. 5 (c) and Fig. 6 (a), and also explains the highest weights attributed to LISFLOOD in the majority of snow-dominated basins. Because of this, and because LISFLOOD tends to overestimate runoff across half of the snow-dominated basins (as shown in Fig.6 (b)) LORA exhibits a positive bias across half of the snow-dominated basins (Fig. 6 (b)) and particularly in Lena, Amur and Yenisei basins (Fig. 8) .

Pan et al. (2012) and Sheffield et al. (2009) assumed that the errors in the measured streamflow are inversely proportional to the area of the basins and ranges between 5% and 10%. Whereas Di Baldassarre and Montanari (2009) analyzed the overall error affecting streamflow observations and found that these errors range between 6% and 42%. In earlier studies, the errors in streamflow measurement were estimated to range from 10% to 20% (Rantz, 1982; Dingman, 1994). In the study of Zhang et al. (2018), the error ratios of VIC were set to be 5%. In this study, we used the weighting approach to compute gridded uncertainty values based on the discrepancy between the Ragg of the derived runoff and the associated observational dataset in each gauged basin or alternatively, based on the discrepancy between Ragg of the derived runoff and the associated observational dataset from three similar basins in the case of ungauged basins. The derived gridded uncertainty changes in time and space. Our uncertainty estimates show higher values than those set for VIC, and additionally the estimated values and their reliability change with climate and season (Fig. 7). It follows from Table 2 that in most of the basins the mean annual runoff uncertainty exceeds 30% of the values of the associated runoff itself. In fact, when the values of runoff approach zero (i.e. in arid and semi-arid regions during the hot climate or in the snow dominated basins during winter) it is expected that the uncertainty values become very close to the associated runoff estimates and eventually the error ratio becomes high. It is not surprising that the estimated relative uncertainties exceed the error ratios of the observations. Also the change of the uncertainty values with time and space is consistent with the fact that the individual datasets that were used to derive LORA exhibit performance differences in different climates and terrains (Beck et al., 2017a).

Figure 9 shows the Mean seasonal runoff (mm/year) calculated for the period 1980 – 2012. There is consistently low runoff in arid regions and high runoff in wet regions across all the seasons. High latitudes in America and Asia exhibit no runoff during the snow season and high runoff during Mar-Aug when snow melts. Overall, there is a clear agreement between the spatial distribution of runoff and the different climate regimes. This is particularly reflected in Madagascar where the differences in runoff pattern match the different climate regimes across the island. LORA captures the high wetness in the monsoonal seasons and exhibits a shift in magnitude during the wet monsoon in the lower Amazon during Oct-May, the upper Amazon during Jun-Aug, South Asia during Jun-Nov, Central Sahel in August and Guinea Coasts in June, July, September and October.

As discussed in Hobeichi et al. (2018), the weighting approach has its own advantages and drawbacks. One limitation is that a common imperfection in all the individual products is likely to propagate into the derived product. The early spring runoff peak found in both LORA and the datasets that were used to derive it is an example of this limitation. On the other hand, the seasonal runoff cycle of LORA in both Pechora and Olenek (i.e. two snow-dominated basins) indicate that LORA was able to capture the seasonal signal and the timing of the runoff peak very well as opposed to the constituent products and Best4, which also suggests that the weighting has the ability to overcome the weaknesses of the individual products. Additionally, it was shown in Beck et al. (2017a) that tier1 products consistently overestimate runoff in arid and semi-arid regions due to a bias in the WFDEI precipitation forcing, this appears in the massive overestimation exhibited by Best4 in Niger and Murray-Darling (Fig. 8), however the weighting was able to eliminate a large amount of this overestimation, which also emphasizes the ability of the weighting approach to mitigate limitations in individual models. Another limitation arises from the scarcity of observed streamflow particularly in the arid regions and from the quality of the observational data itself. As noted earlier, the errors in GRDB dataset were reported to range between 10% and 20% and were found by Di Baldassarre and Montanari (2009) to have an average value that exceed 25% across all the studied river basins.

The weighting technique allows the addition of new runoff estimates when they become available. This will be particularly beneficial if the future estimates represent reasonably the runoff peak in the snow-dominated regions.

5 Conclusion

In this study, we presented LORA, a new global monthly runoff product with associated uncertainty. LORA was derived for 1980–2012 with monthly temporal resolution at 0.5° spatial resolution by applying a weighting approach that accounts for both performance differences and error covariance between the constituent products.

To ensure full global coverage, we used a similarity index to transfer weights and bias ratios constructed from gauged basins with similar climatic and physiographic characteristics to ungauged basins. This allows the derivation of runoff in areas where we do not have observed streamflow.

We showed that this approach is succeeding, that LORA performs better than any of its constituent modelled products in a range of metrics, across basins globally and especially in the higher latitudes. However, LORA tends to overestimate runoff and shows an early snow-melt peak in some snow-dominated basins. LORA was not found to significantly overestimate runoff in arid and semi-arid regions as opposed to the constituent products.

The approach and product detailed here offers the opportunity for improvement as new streamflow and modelled runoff datasets become available. It presents a new, relatively independent estimate of a key component of the terrestrial water budget, with a justifiable and well constrained uncertainty estimate.

6 Competing interests

The authors declare that they have no conflict of interest.

7 Acknowledgment

Sanaa Hobeichi acknowledges the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028). Gab Abramowitz and Jason Evans acknowledge the support of the Australian Research Council Centre of Excellence for Climate Extremes (CE170100023). Hylke Beck was supported by the U.S. Army Corps of Engineers' International Center for Integrated Water Resources Management (ICIWaRM), under the auspices of UNESCO. This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We are grateful to the Global Runoff Data Centre (GRDC) for providing observed streamflow data. We thank the participants of the eartH2Observe project for producing and making available the model simulations. We also acknowledge that the HydroBASINS product has been developed on behalf of World Wildlife Fund US (WWF), with support from, and in collaboration with: the EU BioFresh project, Berlin, Germany; the International Union for Conservation of Nature (IUCN), Cambridge, UK; and McGill University, Montreal, Canada. Major funding for this project was provided to WWF by Sealed Air Corporation; additional funding was provided by BioFresh and McGill University.

8 References

Abramowitz, G. and Bishop, C. H.: Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections, J. Clim., 28(6), 2332–2348, doi:10.1175/JCLI-D-14-00364.1, 2015.

Aires, F.: Combining Datasets of Satellite-Retrieved Products. Part I: Methodology and Water Budget Closure, J. Hydrometeorol., 15(4), 1677–1691, doi:10.1175/JHM-D-13-0148.1, 2014.

Bai, Y., Xu, H. and Ling, H.: Drought-flood variation and its correlation with runoff in three headstreams of Tarim River, Xinjiang, China, Environ. Earth Sci., 71(3), 1297–1309, doi:10.1007/s12665-013-2534-5, 2014.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: A quantitative analysis, Hydrol. Earth Syst. Sci., 13(6), 913–921, doi:10.5194/hess-13-913-2009, 2009.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M. and Betts, A. K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, J. Hydrometeorol., 10(3), 623–643, doi:10.1175/2008JHM1068.1, 2009.

Balsamo, G., Pappenberger, F., Dutra, E., Viterbo, P. and van den Hurk, B.: A revised land hydrology in the ECMWF model: A step towards daily water flux prediction in a fully-closed water cycle, Hydrol. Process., 25(7), 1046–1054, doi:10.1002/hyp.7808, 2011.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., Mcvicar, T. R., Schellekens, J. and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resour. Res., 52, 3599–3622, doi:10.1002/2015WR018247, 2016.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R. and Schellekens, J.: Global evaluation of runoff from ten state-of-the-art hydrological models, Hydrol. Earth Syst. Sci. Discuss., 21, 2881–2903, doi:10.5194/hess-2016-124, 2017a.

Beck, H. E., Van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B. and De Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data, Hydrol. Earth Syst. Sci., 21(1), 589–615, doi:10.5194/hess-21-589-2017, 2017b.

Van Beek, L. P. H. and Bierkens, M. F. P.: The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification. [online] Available from: http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf, 2009.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Menard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B. and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), Model description – Part 2: Carbon fluxes and vegetation, Geosci. Model Dev. Discuss., 4(1), 641–688, doi:10.5194/gmdd-4-641-2011, 2011.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41(3–4), 885–900, doi:10.1007/s00382-012-1610-y, 2013.

Burek, P., van der Knijff, J., de Roo, A.: LISFLOOD DistributedWater Balance and Flood Simulation Model e Revised User Manual 2013,. JRC Technical Reports. Joint Research Centre of the European Commission, Publ. Off. Eur. Union, Luxemb., doi:10.2788/24719, 2013.

Dai, A.: Historical and Future Changes in Streamflow and Continental Runoff: A Review, in Terrestrial Water Cycle and Climate Change: Natural and Human-Induced Impacts, edited by Q. T. and T. Oki, pp. 17–37, John Wiley & Sons, Inc., Hoboken, NJ, USA., 2016.

Decharme, B., Boone, A., Delire, C. and Noilhan, J.: Local evaluation of the Interaction between Soil Biosphere Atmosphere soil multilayer diffusion scheme using four pedotransfer functions, J. Geophys. Res. Atmos., 116(20), 1–29, doi:10.1029/2011JD016002, 2011.

Decharme, B., Martin, E. and Faroux, S.: Reconciling soil thermal and hydrological lower boundary conditions in land surface models, J. Geophys. Res. Atmos., 118(14), 7819–7834, doi:10.1002/jgrd.50631, 2013.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

Van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y. and Tregoning, P.: A global water cycle reanalysis (2003-2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, Hydrol. Earth Syst. Sci., 18(8), 2955–2973, doi:10.5194/hess-18-2955-2014, 2014.

Van Dijk, A. and Warren, G.: The Australian Water Resources Assessment System. Technical Report 4. landscape Model (version 0.5) evaluation Against Observations. [online] Available from: http://www.clw.csiro.au/publications/waterforahealthycountry/2010/wfhc-awras-evaluation-against-observations.pdf, 2010.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T. and Hanasaki, N.: GSWP-2: Multimodel analysis and implications for our perception of the land surface, Bull. Am. Meteorol. Soc., 87(10), 1381–1397, doi:10.1175/BAMS-87-10-1381, 2006.

Falcone, J. A., Carlisle, D. M., Wolock, D. M. and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Ecology, 91(2), 621, doi:10.1890/09-0889.1, 2010.

Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F. and Alcamo, J.: Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study, Glob. Environ. Chang., 23(1), 144–156, doi:10.1016/j.gloenvcha.2012.10.018, 2013.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P. and Yeh, P.: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results, J. Hydrometeorol., 12(5), 869–884, doi:10.1175/2011JHM1324.1, 2011.

Harris, I., Jones, P. D., Osborn, T. J. and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, Int. J. Climatol., 34(3), 623–642, doi:10.1002/joc.3711, 2014.

Hobeichi, S., Abramowitz, G., Evans, J. and Ukkola, A.: Derived Optimal Linear Combination Evapotranspiration (DOLCE): a global gridded synthesis ET estimate, Earth Syst. Sci, 225194, 1317–1336, doi:10.5194/hess-22-1317-2018, 2018.

van Huijgevoort, M. H. J., Hazenberg, P., van Lanen, H. A. J., Teuling, A. J., Clark, D. B., Folwell, S., Gosling, S. N., Hanasaki, N., Heinke, J., Koirala, S., Stacke, T., Voss, F., Sheffield, J. and Uijlenhoet, R.: Global Multimodel Analysis of Drought in Runoff for the Second Half of the Twentieth Century, J. Hydrometeorol., 14(5), 1535–1552, doi:10.1175/JHM-D-12-0186.1, 2013.

Jiménez, C., Martens, B., Miralles, D. M., Fisher, J. B., Beck, H. E. and Fernández-Prieto, D.: Local tower-based merging of two land evaporation products, Earth Syst. Sci. Discuss, 1–41, doi:10.5194/hess-2017-573, 2017.

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P. and Thielen, J.: Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level, Environ. Model. Softw., 75, 68–76, doi:10.1016/j.envsoft.2015.09.009, 2016.

Van Der Knijff, J. M., Younis, J. and De Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, Int. J. Geogr. Inf. Sci., 24(2), 189–212, doi:10.1080/13658810802549154, 2010.

Liang, X., Lettenmaier, D. P., Wood, E. F. and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99(D7), 14415, doi:10.1029/94JD00483, 1994.

Ling, H., Deng, X., Long, A. and Gao, H.: The multi-time-scale correlations for drought–flood index to runoff and North Atlantic Oscillation in the headstreams of Tarim River, Xinjiang, China, Hydrol. Res., 47(6), 1–12, doi:10.2166/nh.2016.166, 2016.

Liu, C., Nepf, H., Gray, W., Yates, D., Velde, Y. Van Der, Bertuzzo, E., Botter, G., Turner, S. W. D., Marlow, D., Ekstr, M., Jeffrey, P. J., Valley, Y., Porporato, A., Calabrese, S., Turner, S. W. D., Marlow, D., Ekstr, M., Jeffrey, P. J., Velde, Y. Van Der, Bertuzzo, E., Botter, G., Index, M., Valley, Y., Velde, Y. Van Der, Bertuzzo, E., Botter, G., Turner, S. W. D., Marlow, D., Ekstr, M., Jeffrey, P. J., Valley, Y., Index, M., Porporato, A., Calabrese, S., Velde, Y. Van Der, Bertuzzo, E., Botter, G., Valley, Y., Porporato, A., Calabrese, S. and Index, M.: Global hydrology 2015: State, trends, and directions, Water Resour. Res., (February), 600–612, doi:10.1002/2015WR017173.Received, 2016.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y. and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., doi:10.5194/hess-17-3707-2013, 2013.

Nyamadzawo, G., Nyamugafata, P., Wuta, M., Nyamangara, J. and Chikowo, R.: Infiltration and runoff losses under fallowing and conservation agriculture practices on contrasting soils, Zimbabwe, Water SA, 38(2), 233–240, doi:10.4314/wsa.v38i2.8, 2012.

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, A. E. F.: Multisource estimation of long-term terrestrial water budget for major global river basins, J. Clim., 25(9), 3191–3206, doi:10.1175/JCLI-D-11-00300.1, 2012.

Pechlivanidis, I. G., Jackson, B. M., Mcintyre, N. R. and Wheater, H. S.: Catchment Scale Hydrological Modelling: A Review Of Model Types, Calibration Approaches And Uncertainty Analysis Methods In The Context Of Recent Developments In Technology And Applications, Glob. NEST J., 13(3), 193–214, doi:10.1002/hyp, 2011.

Peel, M. C., Chiew, F. H. S., Western, A. W. and McMahon, T. A.: Extension of Unimpaired Monthly Streamflow Data and Regionalisation of Parameter Values to Estimate Streamflow in Ungauged Catchments., 2000.

Rantz, S. E.: Measurement and computation of stream flow. Volume 2: Computation of discharge, US Geol. Surv. water-supply Pap., 2(2175), 631, doi:10.1029/WR017i001p00131, 1982.

Reichle, R. H., Koster, R. D., De Lannoy, G. J. M., Forman, B. A., Liu, Q., Mahanama, S. P. P. and Touré, A.: Assessment and Enhancement of MERRA Land Surface Hydrology Estimates, J. Clim., 24(24), 6322–6338, doi:10.1175/JCLI-D-10-05033.1, 2011.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin*, J. K., Walker, J. P., Lohmann, D. and Toll, D.: The Global Land Data Assimilation System, Bull. Am. Meteorol. Soc., 85(3), 381–394, doi:10.1175/BAMS-85-3-381, 2004.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, E. F.: Reconciling the global terrestrial water budget using satellite remote sensing, Remote Sens. Environ., 115(8), 1850–1865, doi:10.1016/j.rse.2011.03.009, 2011.

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Weiland, F. S., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., Van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W. and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe

Tier-1 dataset, Earth Syst. Sci. Data, 95194, 389–413 [online] Available from: https://www.earth-syst-sci-data.net/9/389/2017/essd-9-389-2017.pdf (Accessed 5 September 2017), 2017.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F. and McCabe, M. F.: Closing the terrestrial water budget from satellite remote sensing, Geophys. Res. Lett., 36(7), n/a-n/a, doi:10.1029/2009GL037338, 2009.

Shukla, S. and Wood, A. W.: Use of a standardized runoff index for characterizing hydrologic drought, Geophys. Res. Lett., 35(2), 1–7, doi:10.1029/2007GL032487, 2008.

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

Tomy, T. and Sumam, K. S.: Determining the Adequacy of CFSR Data for Rainfall-Runoff Modeling Using SWAT, Procedia Technol., 24, 309–316, doi:10.1016/j.protcy.2016.05.041, 2016.

Ukkola, A. M., Prentice, I. C., Keenan, T. F., van Dijk, A. I. J. M., Viney, N. R., Myneni, R. B. and Bi, J.: Reduced streamflow in water-stressed climates consistent with CO2 effects on vegetation, Nat. Clim. Chang., 6(1), 75–78, doi:10.1038/nclimate2831, 2016.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J. and Viterbo, P.: Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505–7514, doi:10.1002/2014WR015638.Received, 2014.

Ye, A., Duan, Q., Yuan, X., Wood, E. F. and Schaake, J.: Hydrologic post-processing of MOPEX streamflow simulations, J. Hydrol., 508, 147–156, doi:10.1016/j.jhydrol.2013.10.055, 2014.

Zhai, R. and Tao, F.: Contributions of climate change and human activities to runoff change in seven typical catchments across China, Sci. Total Environ., 605–606, 219–229, doi:10.1016/j.scitotenv.2017.06.210, 2017.

Zhang, Y., Pan, M., Sheffield, J., Siemann, A. L., Fisher, C. K., Liang, M., Beck, H. E., Wanders, N., Maccracken, R. F., Houser, P. R., Zhou, T., Lettenmaier, D. P., Pinker, R. T., Bytheway, J., Kummerow, C. D. and Wood, E. F.: A Climate Data Record (CDR) for the global terrestrial water, Earth Syst. Sci, 225194, 241–263, doi:10.5194/hess-22-241-2018, 2018.

Beven, K. J., Changing ideas in hydrology: The case of physically-based models, J. Hydrol., 105(1–2), 157–172, 1989.

Dingman, S. L., Physical Hydrology, 575 pp., Prentice-Hall, Old Tappan, N. J., 1994.

Morel, P., “Why GEWEX? The agenda for a global energy and water cycle research program”. GEWEX News 11(1), 7–11, 2001.

Earthdata: MEaSUREs project, [online] Available from: https://earthdata.nasa.gov/community/community-data-system-programs/measures-projects (Accessed 31 May 2018), 2017.

Baumgardner, A. and E. Reichel: The World Water Balance. Transl. R. Lee. 179 pp., 31 plates. Elsevier Scientific, Amsterdam, 1975.

Esri Education Team. World Climate Zones – Simplified [Esri shapefile]. Scale Not Given. “MappingOurWorld” Februray 2014. http://services.arcgis.com/BG6nSlhZSAWtExvp/arcgis/rest/services/WorldClimateZonesSimp/FeatureServer. (February 14, 2016). Using: *ArcGIS* [GIS software]. National Geographic.

Siebert S., Döll P., Feick S., Hoogeveen J. and Frenken K.: Global map of irrigation areas version 4.0. 1. Johann Wolfgang Goethe University, Frankfurt am Main, Germany/Food and Agriculture Organization of the United Nations, Rome, Italy. 2007.

Bontemps, S., Defourny, P., Bogaert, E.V., Arino, O., Kalogirou, V. and Perez, J.R.: GLOBCOVER 2009-Products description and validation report, 2011.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R. and Davies, P.M., Global threats to human water security and river biodiversity. Nature, 467(7315), 555-561, 2010.

Tables

Table 1: Model outputs from Tiers 1 and 2 of eartH2Observe project used to derive the synthesis runoff product.

Model	Tier	Our abbreviation	Variables	Spatial Resolution	Reference

HTESSEL	1	HTESS1	Streamflow & Total runoff	0.5°	(Balsamo et al., 2009, 2011)
	2	HTESS2	streamflow & Total runoff	0.25°	(Balsamo et al., 2009, 2011)
JULES	1	JULES1	Total runoff	0.5°	(Best et al., 2011)
	2	JULES2	Total runoff	0.25°	(Best et al., 2011)
LISFLOOD	1	LISF	Streamflow & Total runoff	0.5°	(Burek, P., van der Knijff, J., de Roo, 2013; Van Der Knijff et al., 2010)
PCR-GLOBWB	1	PCRG	Streamflow & Total runoff	0.5°	(Van Beek and Bierkens, 2009)
SURFEX	1	SURF1	Streamflow & Total runoff	0.5°	(Decharme et al., 2011, 2013)
	2	SURF2	Total runoff	0.25°	(Decharme et al., 2011, 2013)
W3RA	1	W3RA	Streamflow & Total runoff	0.5°	(Van Dijk et al., 2014; Van Dijk and Warren, 2010)
WaterGAP3	1	WGAP3	Streamflow & Total runoff	0.5°	(Flörke et al., 2013)
HBV-SIMREG	1	HBVS	Total runoff	0.5°	(Beck et al., 2016)

Table 2: A comparison of mean annual runoff (mm/year) of 16 major basins covering different climate zones around the world for LORA and VIC (Zhang et al., 2018), the mean annual uncertainty values associated with LORA runoff are shown and the adjusted VIC annual runoff values within 5% error bounds for water budget closure are displayed.

Basin	VIC mm/year	VIC adjusted for water budget closure mm/year	LORA (Runoff) mm/year	LORA (uncertainty) mm/year	Dominant climate
Amazon	1048	1029	1151	357	Tropical wet
Amur	135	129	219	115	Humid continental and semi arid
Columbia	318	293	333	101	Semi-arid and highlands
Congo	407	404	358	147	Tropical wet and tropical dry
Danube	272	265	260	125	Marine Humid, continental and humid subtropical

Indigirka	132	120	228	171	Subarctic
Lena	142	134	301	137	Subarctic
Mackenzie	189	173	191	110	Subarctic
Mississippi	220	215	212	123	Humid continental and humid subtropical
Murray-Darling	42	41	15	6	Arid and semi-arid
Niger	198	194	106	41	Arid, semi-arid and tropical dry
Olenek	114	106	230	208	Subarctic
Parana	278	279	189	97	Marine and humid subtropical
Pechora	342	308	420	420	Tundra and subarctic
Yenisei	217	195	324	203	Subarctic
Yukon	149	139	229	102	Subarctic

Figures

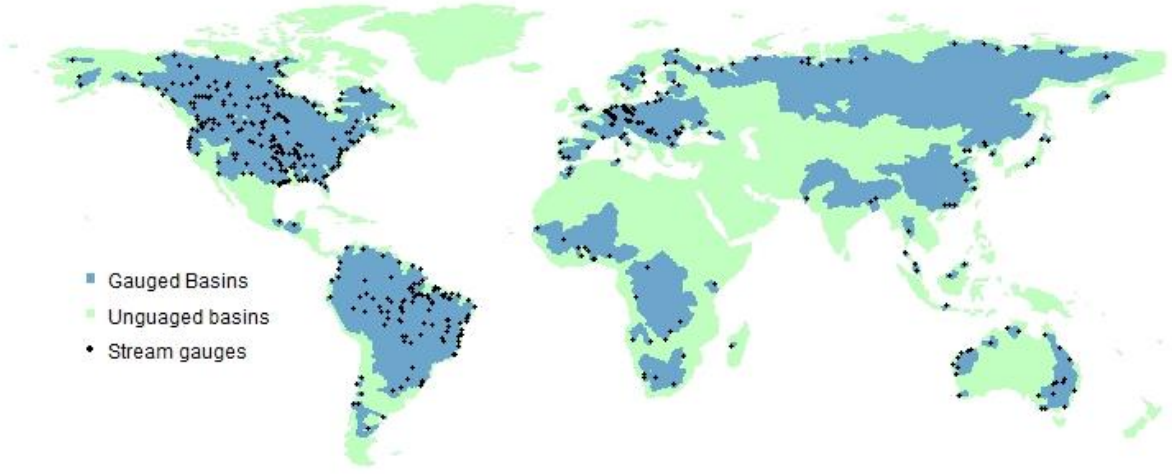

Figure 1: Spatial coverage of gauged and ungauged river basins and location of stream gauges.

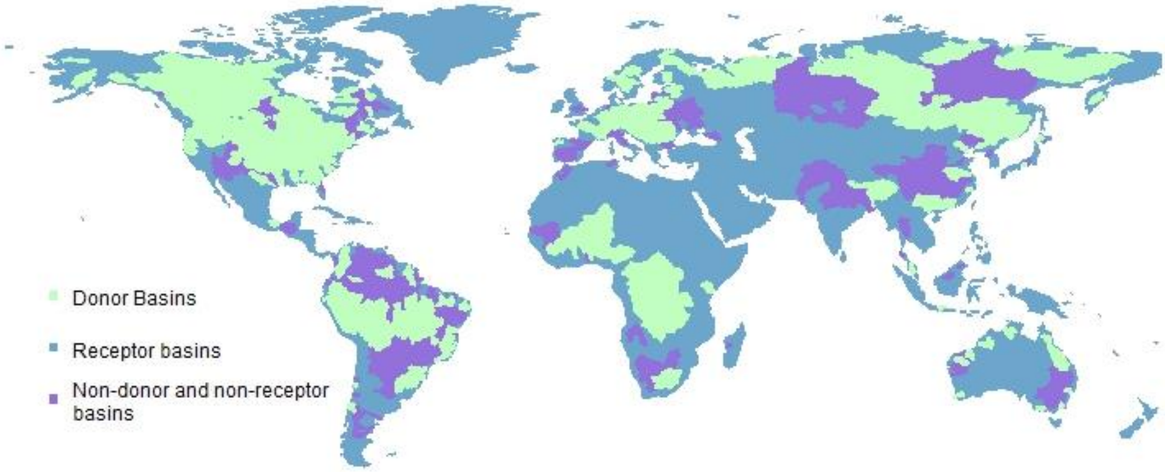

Figure 2: Spatial coverage of donor basins, receptor basins and non-donor and non-receptor basins.

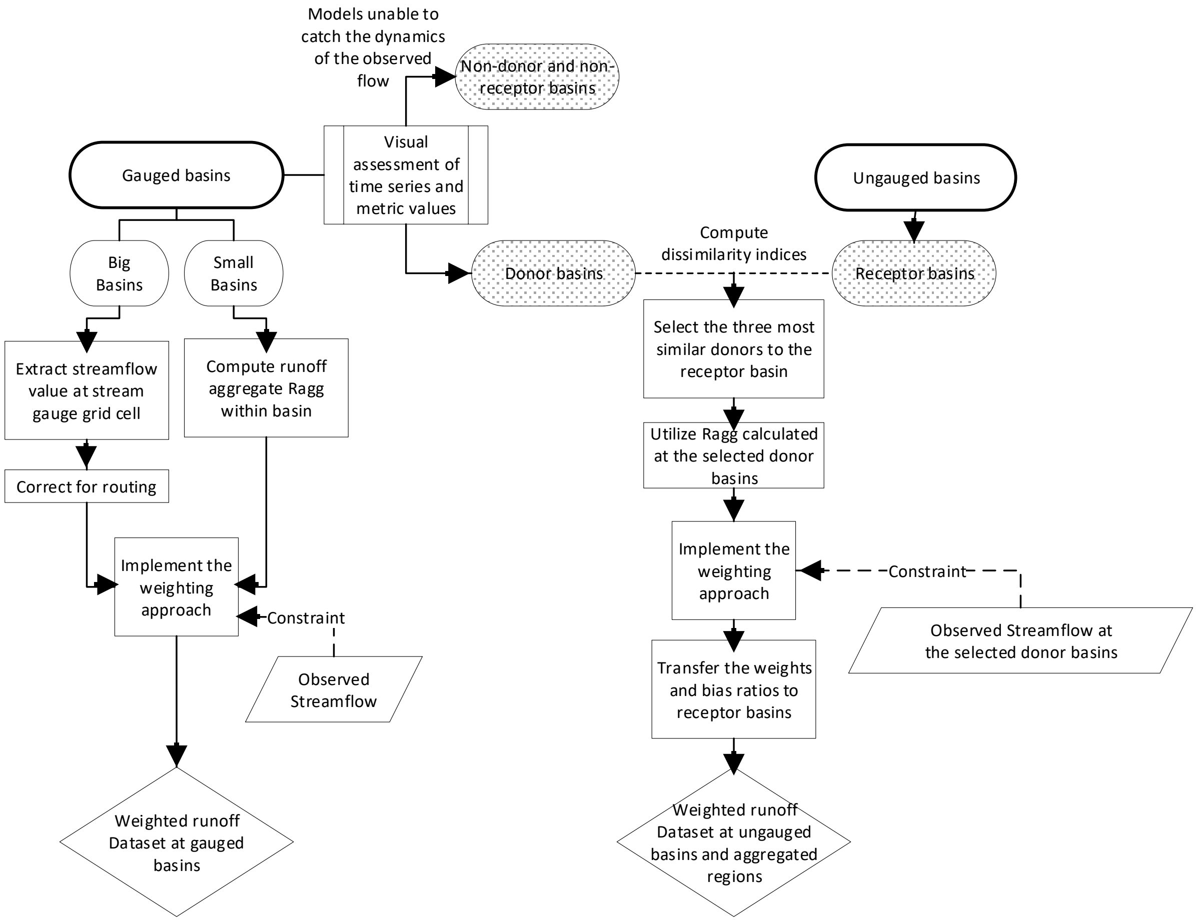
Models unable to catch the dynamics of the observed flow
Non-donor and non-receptor basins
Gauged basins
Visual assessment of time series and metric values
Ungauged basins
Big Basins
Small Basins
Donor basins
Compute dissimilarity indices
Receptor basins
Extract streamflow value at stream gauge grid cell
Compute runoff aggregate Ragg within basin
Select the three most similar donors to the receptor basin
Utilize Ragg calculated at the selected donor basins
Correct for routing
Implement the weighting approach
Constraint
Implement the weighting approach
Constraint
Observed Streamflow at the selected donor basins
Observed Streamflow
Transfer the weights and bias ratios to receptor basins
Weighted runoff Dataset at gauged basins
Weighted runoff Dataset at ungauged basins and aggregated regions

Figure 3. Flow chart summarizing the steps carried out to derive the weighted runoff product for the global land surface.

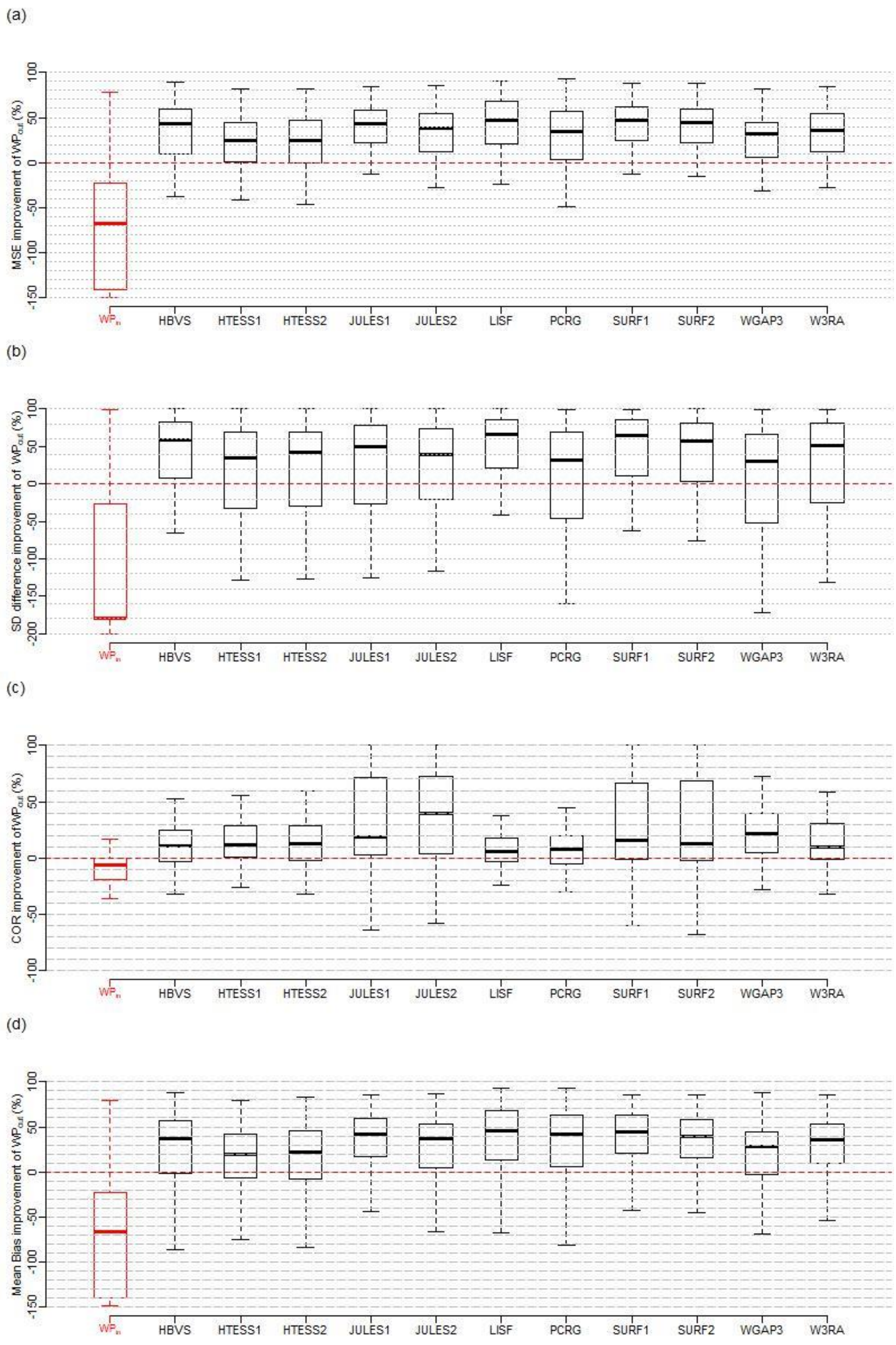
(a)
MSE improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
(b)
SD difference improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
-200
(c)
COR improvement of WP$_{out}$ (%)
100
50
0
-50
-100
(d)
Mean Bias improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
WP$_{in}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

Figure 4: Box and whisker plots displaying the percentage improvement that the weighted product ($\mathrm{WP_{out}}$) offers when tested out-of-sample, using four metrics: MSE (a), SD difference (b), COR (c) and Mean bias (d), when compared to the weighted product derived from in-sample data ($\mathrm{WP_{in}}$), and each runoff product involved in this study. Box and whisker plots represent values calculated at 482 gauged basins. See Table 1 for dataset abbreviations. The lower and upper hinges of a boxplot represent the first (Q_1) and third (Q_3) quartiles respectively of the performance improvement results and the line inside the boxplot shows the median value. The extreme of the lower whisker represents the maximum of 1) min(dataset) and 2) (Q_1 - IQR), while the extreme of the upper whisker is the minimum of 1) max(dataset) and 2) (Q_3 + IQR)), where IQR represents the interquartile range (i.e. Q_3 - Q_1) of the performance improvement results. A median line located above the 0 axis is an indication that the out of sample weighting offers an improvement in more than half of the basins.

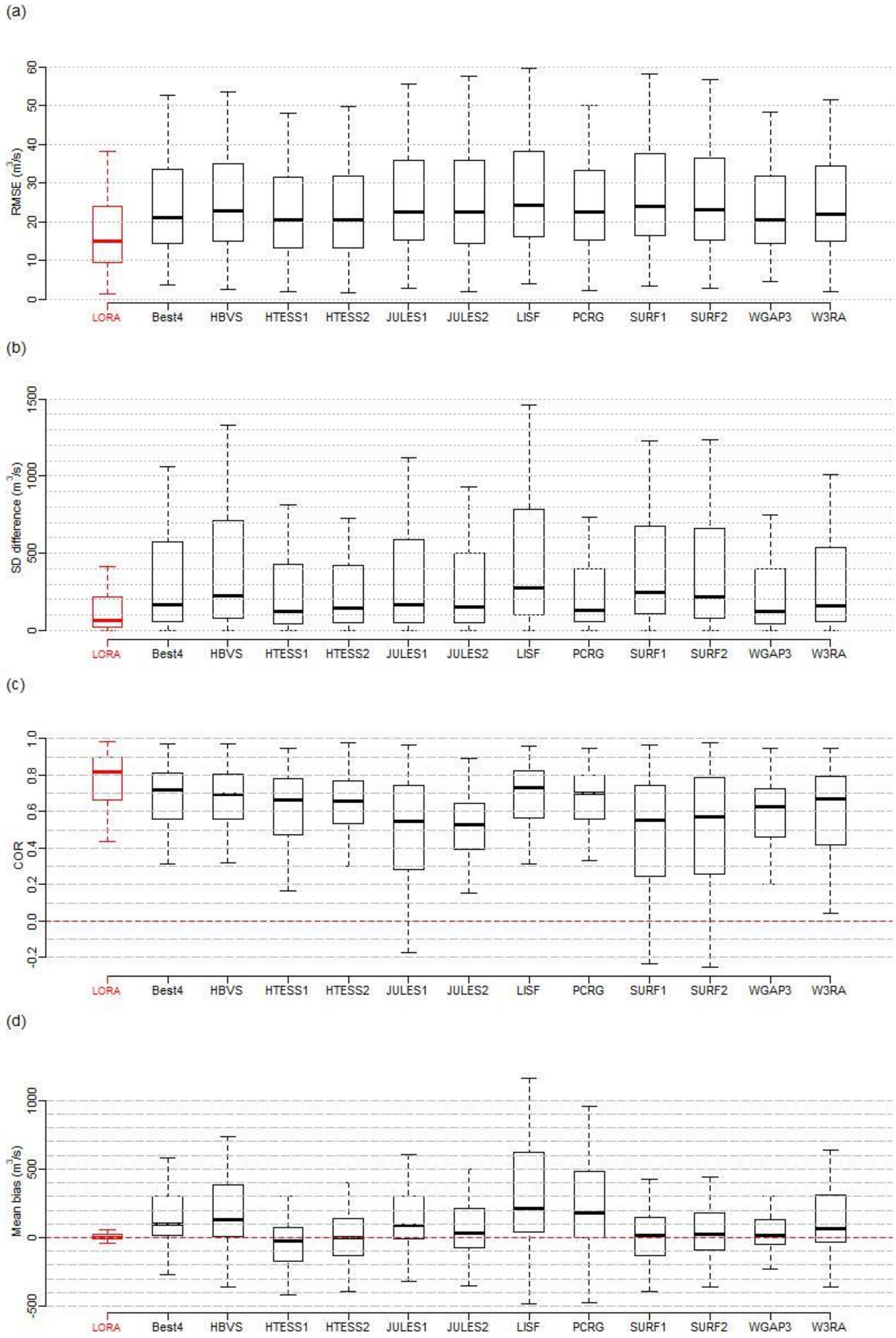
(a)
RMSE (m³/s)
0
10
20
30
40
50
60
LORA
Best4
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA
(b)
SD difference (m³/s)
0
500
1000
1500
LORA
Best4
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA
(c)
COR
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
LORA
Best4
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA
(d)
Mean bias (m³/s)
-500
0
500
1000
LORA
Best4
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

Figure 5: Four statistics, (a) RMSE, (b) SD difference, (c) COR and (d) Mean bias, calculated for LORA, Best4 (i.e. the simple average of runoff estimates from LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) and each runoff product involved in this study at the gauged basins. See Table 1 for dataset abbreviations.

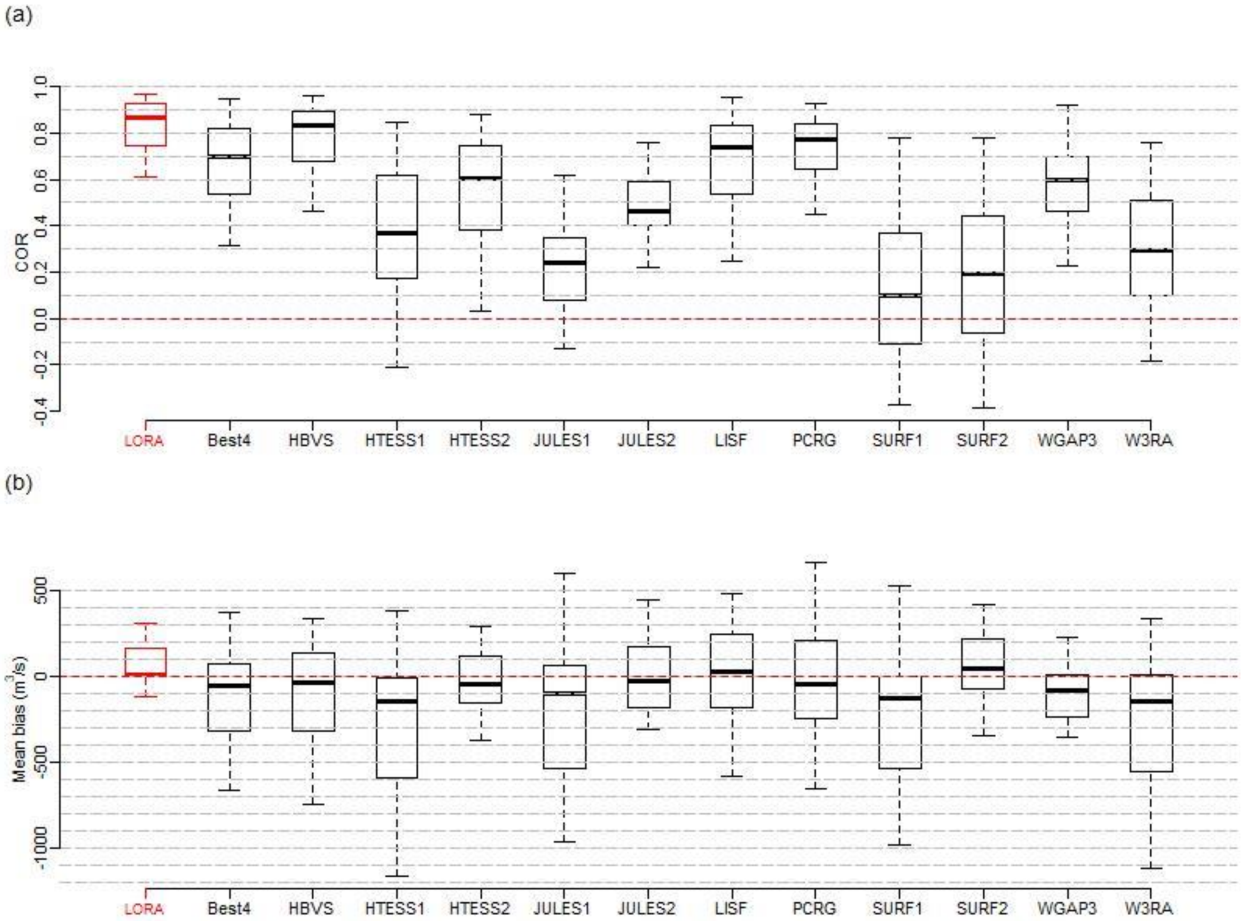

Figure 6: Two statistics, (a) COR and (b) Mean bias, calculated for LORA, Best4 (i.e. the simple average of runoff estimates from LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) and each runoff product involved in this study at the gauged basins located at the high latitudes (>60°). See Table 1 for dataset abbreviations.

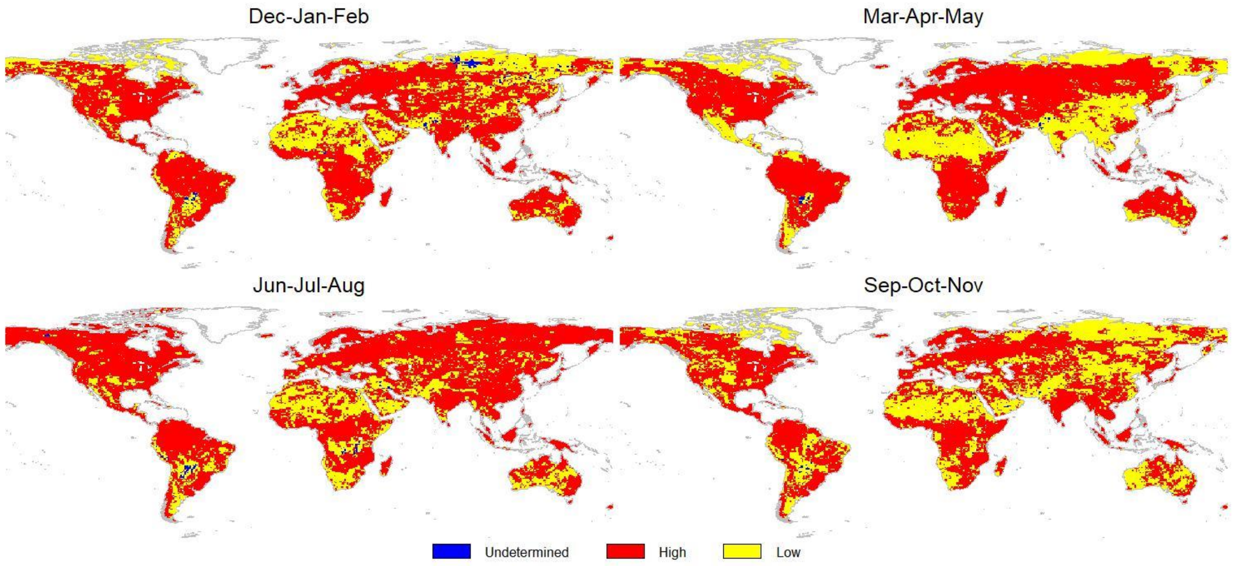

Figure 7: Seasonal reliability, defined as high ($\frac{men\ runoff\ uncertainty}{mean\ runoff} < 1$, in red), low ($\frac{men\ runoff\ uncertainty}{mean\ runoff} \geq 1$, in yellow) and undetermined (mean runoff = 0, in blue).

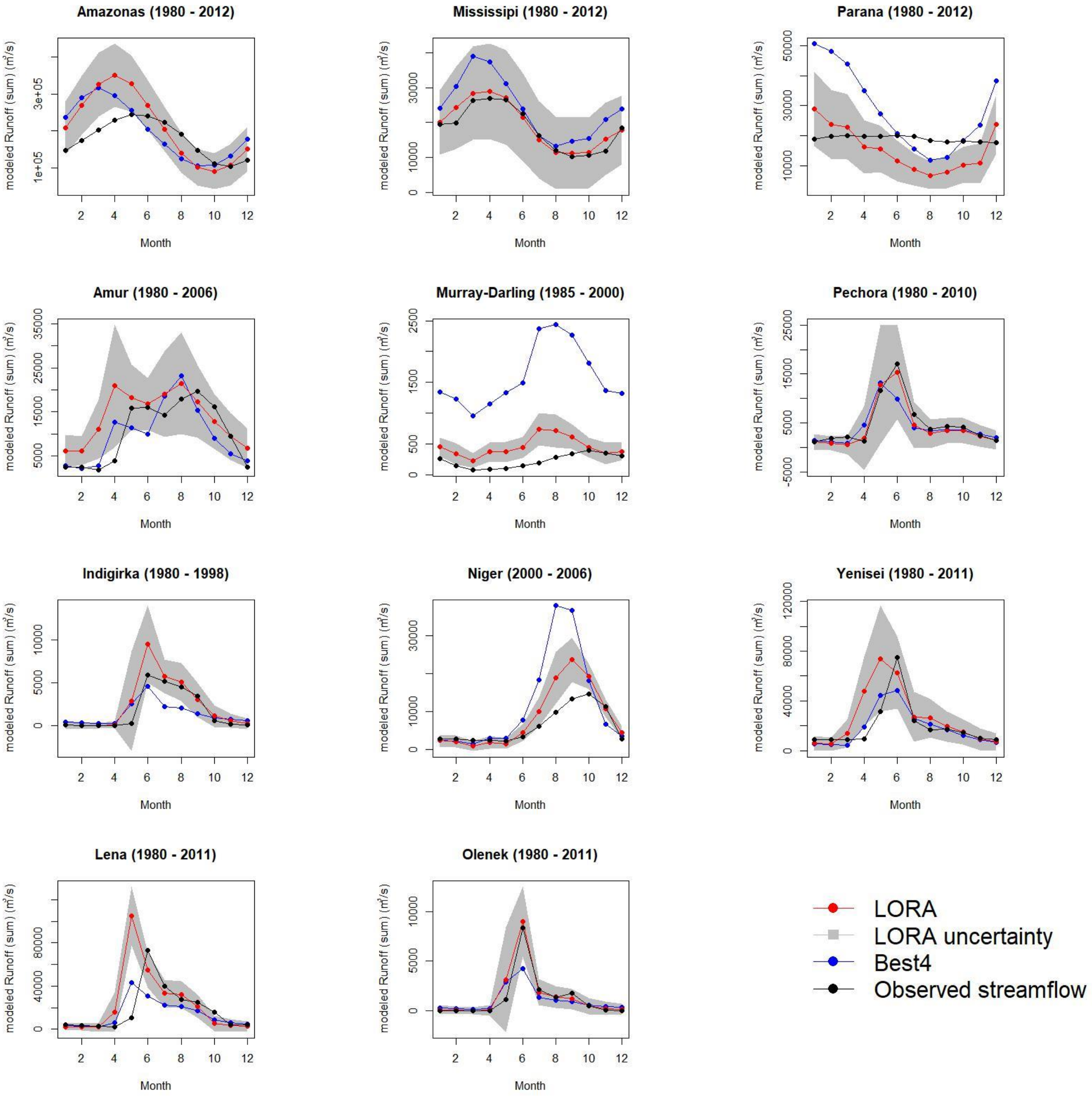

Figure 8: Seasonal cycle of Runoff aggregates from LORA and Best4 compared with the observed streamflow over 11 major basins. Runoff aggregates and the observed streamflow were averaged for each month across the period of availability of observation. The shaded regions shows the aggregated uncertainty derived for LORA.

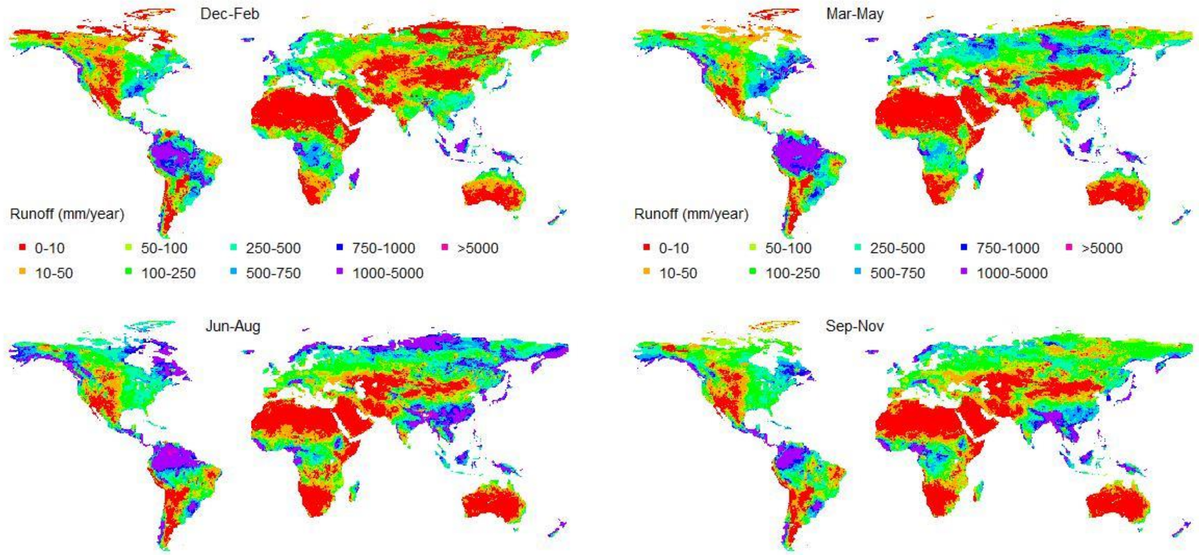

Figure 9: Mean seasonal runoff calculated for the period 1980 – 2012