Manuscript hess-2018-386 entitled "Linear Optimal Runoff Aggregate (LORA): A global gridded synthesis runoff product"

We would like to thank the reviewers for their constructive comments on our manuscript. This document outlines our responses to his comments and the improvements made to the manuscript. We have also added the modified manuscript and the new supplement at the end of this document.

Response to Reviewer #1

Overall comments and questions

(1) Please refer early to Fig. 3 in the methods description.
Thanks for this suggestion, we have now referred to the flowchart in Figure 3 before detailing our methods. Its caption has now changed to Figure 2.

(2) Maps of the performance of LORA would be very helpful.

(3) Some example time series, including the time-varying uncertainty would be interesting.

Both (2) and (3) are good ideas. We have now provided a map of the temporal correlation of aggregated LORA runoff and observed streamflow, as well as time series computed over selected basins with different correlation levels, showing LORA uncertainty estimates. We have also added the text:

> *Further, we provide in Fig. S2 the spatial distribution of correlation results from Fig. 6 (c). The basins are colour-coded by their temporal correlation with the observed streamflow and the number of basins in each category is given. Basins in yellow are those where LORA is highly correlated with the observation while dark blue basins are those where LORA exhibits a negative correlation with the observation. It can be noted from Fig. 6(c) that occurrence of negative correlation is extremely unusual which explains why these were considered outliers and were not shown in the box and whisker plot. Likely, low correlation basins are unusual and constitute less than 12% of the number of basins (excluding basins with negative correlation). Also, the median value is above 0.8, which is higher than any constituent estimates. We selected a basin from each correlation range and examined the timeseries of LORA and the observed streamflow more closely (Fig. S3-S7), in particular illustrating the uncertainty estimate of LORA. In Ganges, LORA captures well the observed time-series dynamic with a tendency to over-estimate streamflow peak in August (Fig. S3). Over Madeira basin, LORA is able to represent reasonably well most of the climatic variability found in the observation (Fig. S4). In Congo, the catchment has an irregular time-series dynamic, LORA is in principle able to capture a large part of the climatic variability in the observation (Fig. S5). In*

Lena, the observation shows a peak in June and a second less significant peak in September (Fig. S6). Both peaks are captured by LORA during most of the time series with a tendency to underestimate the late summer peak and overestimate the early summer peak. In the upper Indus, LORA does not capture the magnitudes of observed streamflow and shows a reversed seasonal cycle which explains why it exhibits negative correlation with the observation (Fig.S7). Zhang et al. (2018) found disagreement between simulated runoff from three LSMs and observed streamflow over Indus basin which they expected to be due to errors in the observational data from GRDB dataset.

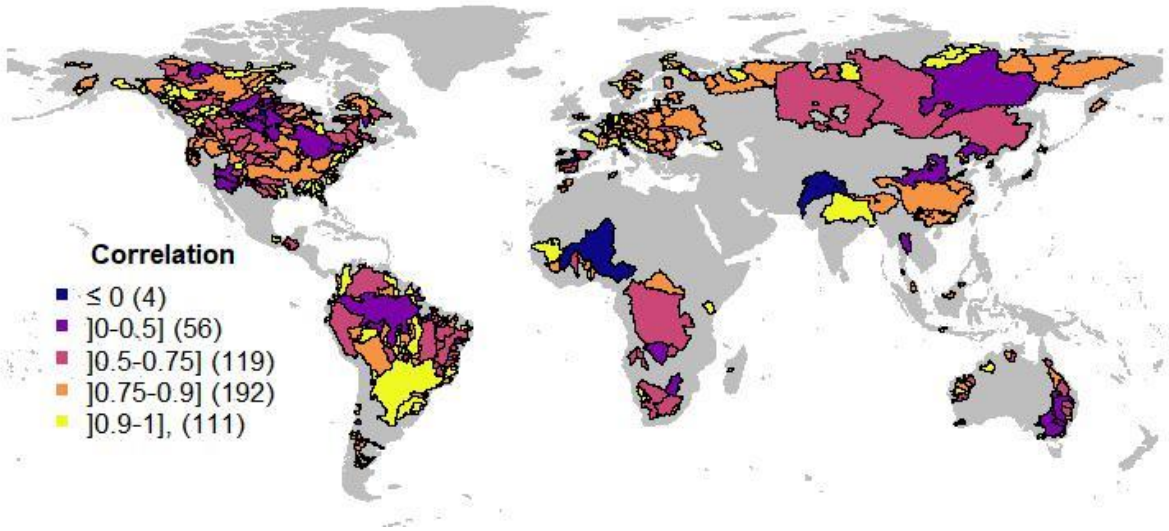


Figure S2: Temporal correlation of LORA with the observed streamflow over the gauged basins. Basins are colour coded by correlation range and their numbers are given in brackets.

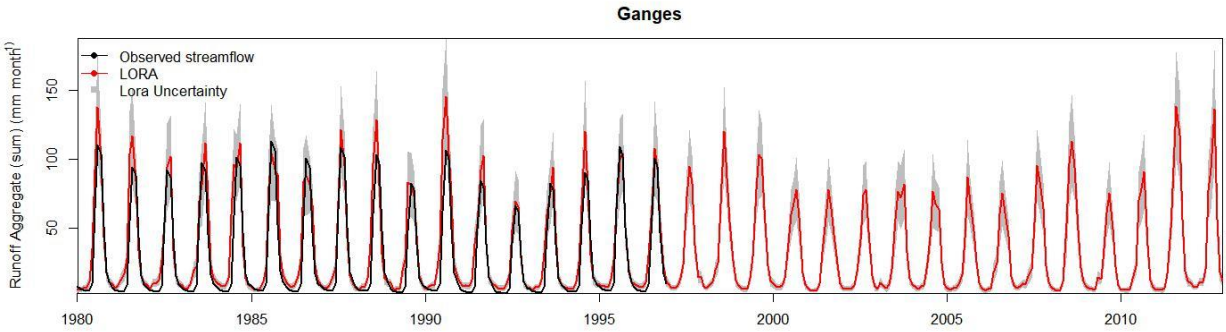


Figure S3: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Ganges basin (in mm month^{-1}). This basin was shown in yellow in Fig. S2, indicating that LORA exhibits a high temporal correlation (≥ 0.9) with the observation.

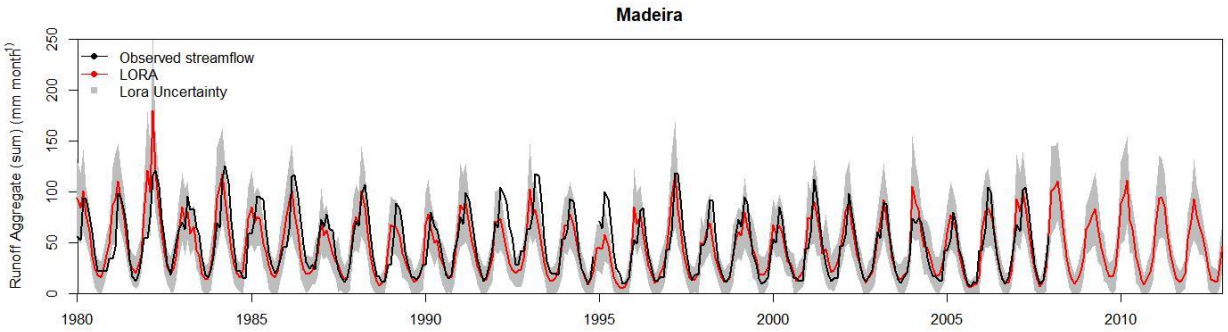


Figure S4: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Madeira basin, i.e. a sub-basin of the Amazonas (in mm month^{-1}). This basin was shown in orange in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.75 – 0.9[with the observation.

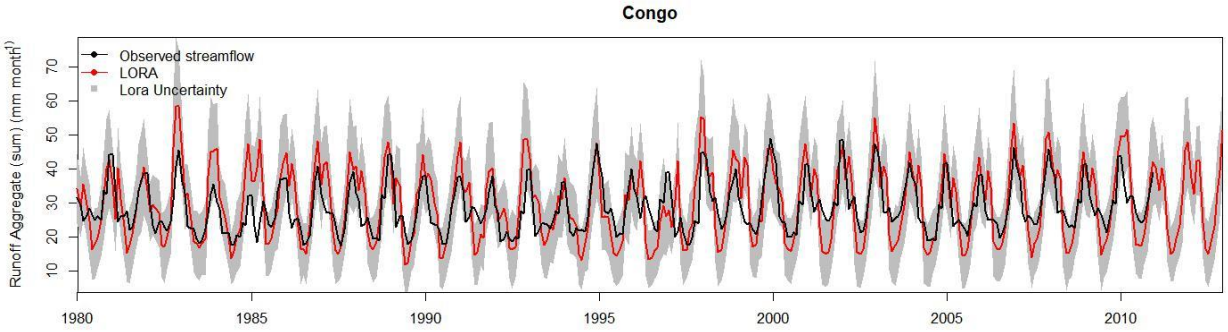


Figure S5: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Congo basin (in mm month^{-1}). This basin was shown in violet in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.5 – 0.75[with the observation.

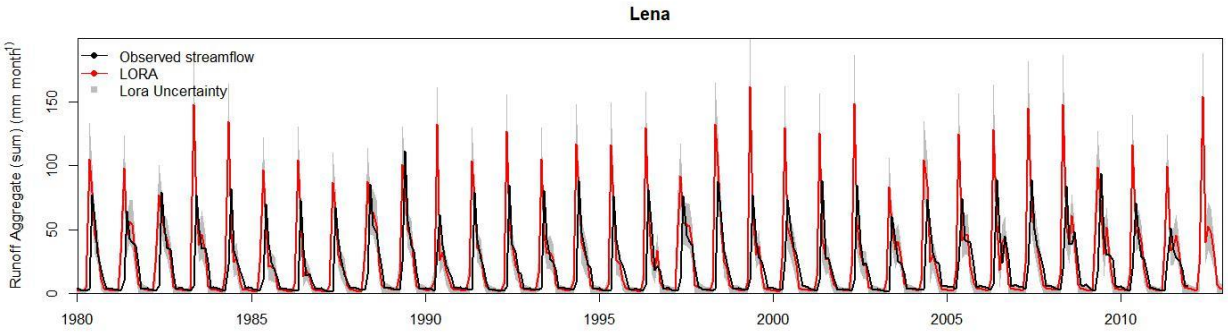


Figure S6: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Lena basin (in mm month^{-1}). This basin was shown in purple in Fig. S2, indicating that LORA exhibits low temporal correlation (<0.5) with the observation.

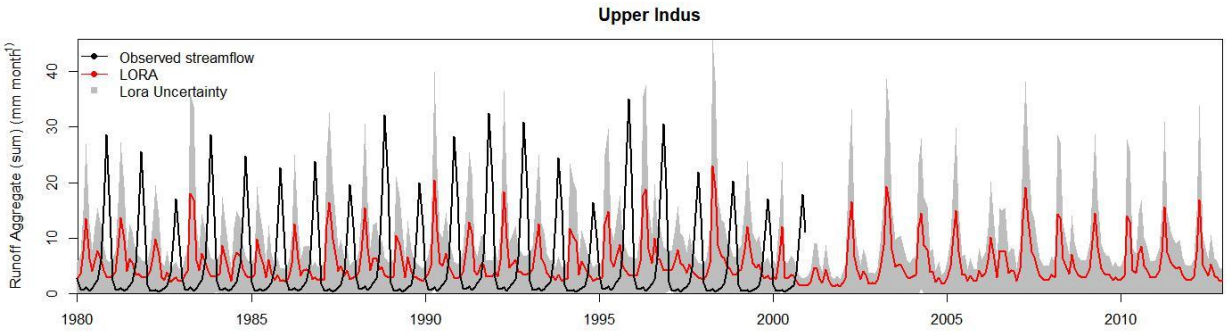


Figure S7: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Indus basin (in mm month^{-1}). This basin was shown in dark blue in Fig. S2, indicating that LORA exhibits a negative temporal correlation with the observation.

(4) Does the fact that the weights are computed based on a discrete set of donor catchments lead to discontinuities in the runoff fields? How would e.g. the field look like for one individual month?

Good point. We obviously don't expect to see discontinuity in the runoff fields within individual gauged basins since the weighting is uniform across the basin. On the other hand, we expect to see discontinuity in runoff fields at ungauged basins, particularly over large heterogeneous ungauged basins, given that different sets of weights were used to derive runoff over different parts of the basins. However, since dividing the heterogeneous basins into different regions was based on climatic differences between these regions, we think that such discontinuity naturally arises from the differences in climatic characteristics anyway, so that a different set of weights for each region is not necessarily problematic. It certainly seems a better use of available information that weighting globally. A visual assessments of LORA monthly maps did not reveal unexpected discontinuities in the runoff fields. We provide in Fig. S8 an example of runoff fields in an individual month (e.g. May, 2003). Since LORA is publicly available it should be easy to verify that this example is not a misrepresentation of the results.

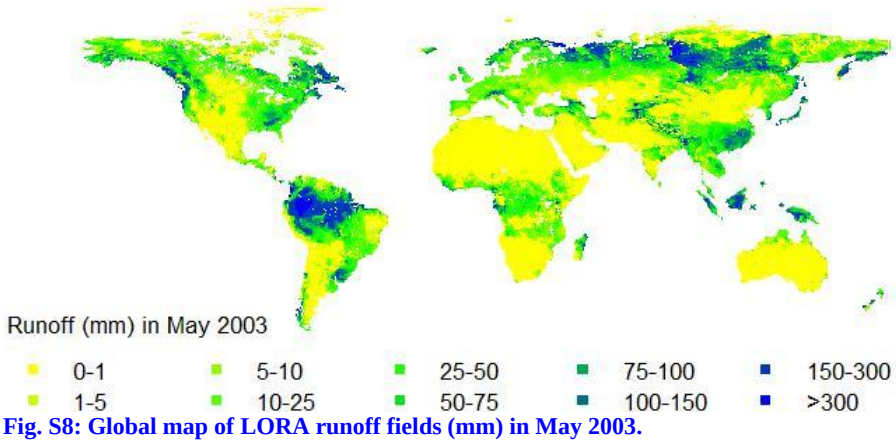


Fig. S8: Global map of LORA runoff fields (mm) in May 2003.

(5) I really appreciate the authors effort to also include uncertainty estimates in their product. Unfortunately, I did not find any validation of this uncertainty estimate or a full interpretation of what it means. I know the uncertainty estimates are introduced elsewhere, but as this is a relatively new approach it requires extra care.

We have now explained in more detail how we calculated the uncertainty estimates:

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. For any given gauged basin, we first calculated the spatial aggregate of our weighted runoff estimate $Ragg_\mu$, then quantified ${s_q}^2$, the error variance of $Ragg_\mu$ with respect to the observed streamflow Q over time as:

$$ {s_q}^2 = \frac{\sum_{j=1}^{J}({Ragg_\mu}^j - Q^j)^2}{J-1} $$

Then, we wished to guarantee that the variance of the constituent modelled estimate ${\sigma_q^2}^j$ about ${Ragg_\mu}^j$ at a given time step, averaged over all time steps where we have available streamflow data, is equal to ${s_q}^2$, such as ${s_q}^2 = \frac{1}{J}\sum_{j=1}^{J}{\sigma_q^2}^j$.

Since the variance of the existing constituent products do not, in general, satisfy this equation. We transformed them so that it does. This involved first modifying the set of weights w to a new set $\widetilde{w}$ such that

$\widetilde{w} = \frac{w^T + (\alpha - 1)\frac{1^T}{K}}{\alpha}$, *where* $\alpha = 1 - K min(w_k)$ *and* $min(w_k)$ *is the smallest negative weight (and* α *is set 1 if all* w_k *are non-negative). This ensures that all the modified weights* $\widetilde{w}_k$ *are positive. We then transform the individual estimates* x_k^j *to* $\tilde{x}_k^j$ *where* $\tilde{x}_k^j = {Ragg_\mu}^j + \beta(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - {Ragg_\mu}^j)$ *and* $\beta = \sqrt{\frac{s_q^2}{\frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\widetilde{w_k}(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - {Ragg_\mu}^j)^2}}$.

The weighted variance estimate of the transformed ensemble can be defined as ${\sigma_q^2}^j = \sum_{k=1}^{K}\widetilde{w}_k(\tilde{x}_k^j - {Ragg_\mu}^j)^2$ *and ensures that the equation* $\frac{1}{J}\sum_{j=1}^{J}{\sigma_q^2}^j = s_q^2$ *holds true. Furthermore,* $\sqrt{{\sigma_q^2}^j}$ *is the temporally varying estimate of uncertainty standard deviation of the transformed ensemble that (a) is varying in time, and (b) accurately reflects our ability to reproduce the observed streamflow.*

We refer the reader to Bishop and Abramowitz (2013) for proofs.

In order to estimate $\sqrt{\sigma_r^{2j}}$, the uncertainty of the runoff attributes μ_r^j at each point in time and space, we first transformed the runoff fields r_k^j to $\tilde{r}_k^j$ by applying the same transformation parameters α and β such that $\tilde{r}_k^j = \mu_r^j + \beta(\bar{r}^j + \alpha(r_k^j - \bar{r}^j) - \mu_r^j)$. We then calculated the error variance $\sigma_r^{2j} = \sum_{k=1}^{K} \widetilde{w}_k(\tilde{r}_k^j - \mu_r^j)^2$.

Finally, we used $\sqrt{\sigma_r^{2j}}$ as the spatially and temporally varying estimate of runoff uncertainty standard deviation, which we will refer to below simply as 'uncertainty'. It provides a much more defensible uncertainty estimate than simply calculating the standard deviation of the involved products.

We note that for a given basin, $\sqrt{\sigma_q^{2j}}$ represents the uncertainty of the modelled streamflow i.e. $Ragg_\mu{}^j$, while $\sqrt{\sigma_r^{2j}}$ represents the uncertainty of modelled runoff at each grid cell across the basin. This means that at every time step, there is one value for $\sqrt{\sigma_q^{2j}}$ per basin, and one value for $\sqrt{\sigma_r^2}$ per grid across the basin.

(5.1) For example: It is not clear to me whether this uncertainty estimate is related to the "confidence interval" (i.e. an estimate for the range of the "true" ensemble mean") or whether it is related to the "prediction interval" (i.e. an estimate of the range in which new observations would fall).

(5.1) It follows from the above that $\sqrt{\sigma_q^{2j}}$ can be considered a 'prediction interval' if it was calculated over an ungauged basin. So $\sqrt{\sigma_r^{2j}}$, the uncertainty estimate associated to the runoff fields, is also considered a prediction interval. The applied transformation can be viewed as a tool to extrapolate the uncertainty estimates in two directions: 1) basin-to-basin: from gauged to ungauged basins, and 2) basin-to-grid: from runoff aggregate in a basin to grid cells across the same basin.

(5.2) if the uncertainty is an "prediction interval", an evaluation of the widths of the "uncertainty bounds" with respect to the distribution of the residuals would be helpful. Especially, compared to the spread of the input ensemble. For this, metrics from ensemble forecasting (e.g. the continuous ranked probability score (CRPS) or "reliability plots/histograms" might be helpful).

We don't really have observations for runoff, so we can't test our method for deriving uncertainty on runoff, however we can test it on streamflow (i.e. runoff aggregates). At the gauged basins, our method for deriving the uncertainty estimates guarantees -by design- that the

uncertainty estimates are equal to the RMSE of runoff aggregates against observed streamflow. Therefore, we see that the good performance of uncertainty estimates at the gauged basins is obvious. However, to test that our approach is also succeeding over the ungauged basins, we have now performed out-of-sample tests to show that the distribution of the errors over the gauged basin is similar to the distribution of their errors when they are considered ungauged. We have now explained how we have performed this test and showed the results in the manuscript:

> *The uncertainty estimates computed at the gauged basins represent the deviation of (the spatial aggregate of) our weighted product ($Ragg_{\mu}$) from the observed streamflow, since the in-sample uncertainty estimates are calculated from the variance of the transformed ensemble, which by design equals MSE of $Ragg_{\mu}$ against observations (i.e. error variance of $Ragg_{\mu}$). To test if the uncertainty estimates perform well out-of-sample (i.e. at the ungauged basins), we took a gauged basin, but instead of constraining the weighting using observed streamflow from this basin, we constructed model weights by using the three most similar donor basins. We could then calculate MSE of $Ragg_{\mu}$ against observations from the three donor basins, denoted by MSE_{in}, which provides us with the uncertainty estimates calculated in-sample ($\sqrt{MSE_{in}}$), since the observational data used in this case is the same dataset that was used to train the weighting. We also calculated the MSE of the aggregated weighted product against the actual observation of the gauged basin and denoted this MSE_{out}. $\sqrt{MSE_{out}}$ represents the uncertainty estimates computed out-of-sample, since the comparison was performed against observational data that has not been used to train the weighting. We repeated the out-of-sample test for all the gauged basins.*
>
> *We displayed the results of the out-sample-test by showing the ratio $\sqrt{MSE_{out}}$ /$\sqrt{MSE_{in}}$. If the approach is succeeding, we expect that this ratio is around one, indicating that the values of MSE_{in} and MSE_{out} are close to each other. We used a box and whisker plot, where each sample is a different basin, to show the results.*

We have also commented on the results:

> 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. *Furthermore, the boxplot in Fig 5 shows that, overall, when the uncertainty estimates are computed out-of-sample they are very similar to what they would have been if they were computed in-sample. This demonstrates that the dissimilarity technique can be effectively used to derive not only the weighting product but also its associated uncertainties at the ungauged basin.*

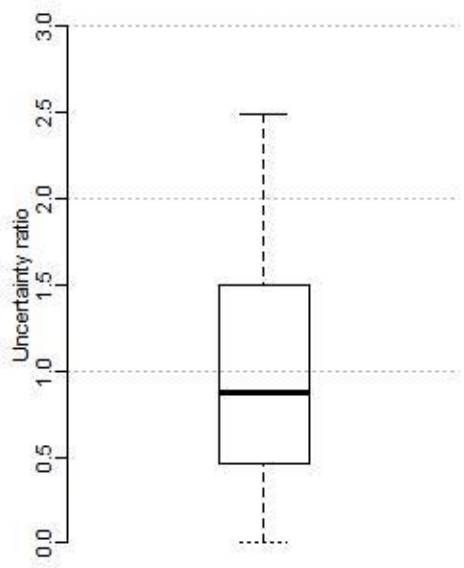


Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample to (2) the uncertainties of the spatial aggregate of the weighted computed in-sample.

(5.3) the uncertainty bounds can produce negative runoff. Should this be the case or is this an artefact?

Good point. When runoff is smaller than its associated uncertainty, the uncertainty bounds will certainly produce negative runoff. This is an artefact that arises from using a generic process to determine an uncertainty range. A negative runoff is obviously not physical and requires some interpretation from the side of the user, to make sure that there are hard boundaries.

(6) I would appreciate some more information (figures, tables) on the actual "weights" and "biases"

We thank the reviewer for his suggestion. We have now provided in table S1 the weights and bias ratios calculated for the participating products over a range of river basins:

> *Table S1 shows examples of weights and bias ratios calculated for the participating models over a range of river basins. It shows that HBVS, JULES1, JULES2 and SURF2 didn't participate in the weighting over the large basins (i.e. Amur, Indigirka, Mississippi, Murray-Darling, Olenek, Parana, Pechora and Yenisei) since these models don't have estimates for streamflow which are needed to construct the weights over large basins. For the smaller Copper River basin, however, runoff estimates from all models participated in deriving weighted runoff estimates. Table S1 also shows that in many cases, models were assigned negative weights. While this might not be expected in typical performance-based weighting, it is possible when weighting is based on error covariance as well as their performance differences in this formulation. We show below how the weights can be modified to non-negative weights.*

Table S1: Example of weights (w) and bias ratios (r) computed for the participating products over a range of river basins.

	HBVS		HTESS1		HTESS2		JULES1		JULES2		LISF		PCRG		SURF2		SURF1		WGAP3		W3RA	
	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r
Amur			-1.22	-0.78	0.14	-0.18					0.46	0.11	1.75	0.09			0.51	-1.35	-0.71	-0.10	0.08	-0.46
Copper	0.33	-0.32	-0.35	-0.42	0.59	0.02	-0.47	-0.35	-0.33	0.20	0.14	-0.18	0.89	-0.19	0.32	0.42	0.84	-0.01	0.07	0.03	-1.02	-0.20
Indigirka			-0.35	-0.80	0.03	0.02					1.23	-0.27	0.42	-0.10			0.79	-0.95	-0.02	-0.04	-1.10	-1.15
Mississippi			0.33	-0.14	0.02	-0.45					-0.09	0.39	0.31	0.28			-0.13	-0.06	0.25	0.02	0.30	0.24
Murray-Darling			1.01	0.75	0.08	0.74					-0.12	0.91	0.01	0.93			0.08	0.81	0.34	0.40	-0.41	0.84
Olenek			-0.35	-0.76	0.08	-0.12					2.07	-0.16	-0.95	-0.18			0.28	-0.70	0.10	-0.03	-0.22	-0.83
Parana			0.26	0.01	0.16	-0.19					-0.88	0.34	0.18	0.48			0.20	-0.12	1.29	0.06	-0.21	0.36
Pechora			-0.38	-0.34	0.33	-0.15					0.84	-0.13	0.82	-0.21			-0.01	-0.42	-0.47	-0.01	-0.14	-0.41
Yenisei			-0.71	-0.71	-0.10	-0.21					2.20	-0.09	-0.96	-0.09			0.52	-0.65	0.39	-0.01	-0.33	-0.72

(7) some of the references have artefacts, e.g. "nan" values instead of page numbers…

Thanks for spotting this out. We have now fixed this in the manuscript

Response to specific comments and questions

Section 2.1: The authors might also be interested in the following global-scale data source: Do et al (2018, doi:10.5194/essd-10-765-2018) & Gudmundsson et al (2018, doi: 10.5194/essd-10-787-2018). Sorry for this self-citation.

Thanks for pointing us to this very impressive dataset. It is not immediately clear how we can use the GSIM in the current analysis since we need continuous discharge time series. However, we will work towards using the dataset in the future to improve our dissimilarity index.

Section 2.2: What about other comprehensive model ensembles, such as ISIMIP2a (http://dx.doi.org/10.5880/PIK.2017.010)

We agree with the reviewer that ISIMIP2a model outputs provide a suite of valuable datasets for runoff and discharge. However, we haven't included any of those in our analysis because at the beginning of our project, these datasets spanned up to 2005 only while our employed datasets spanned up to 2012.

We also note that three out of the eight models that we employed in this study are members of the ISIMIP2a models. In the future, we aim to include additional datasets, and we might consider including datasets from ISIMIP2a if we sort out the difference in the temporal coverage period (i.e. up to 2010 now for ISIMIP2a ensemble, while up to 2012 for LORA)

page 4, line 25: Is there some empirical evidence supporting the assumption of constant bias ratio?

Good point. The literature doesn't provide any empirical evidence that supports or contradicts this assumption. However, this assumption was a part of our whole approach that we have tested in section (2.5). The results from Figures 4 and 5 indicate that our overall approach was succeeding.

We have now clarified this point in section 2.3:

> *We note that there is no empirical evidence in the literature that the assumptions presented in Eq 1 and Eq 2 are valid or invalid. However, they are a core part of our overall approach which we tested and demonstrated to be successful later in this paper.*

Generall for methods: I found the mixed use of Ragg_kˆj and q_kˆj confusing. Would it be possible to clarify the difference/commonality once and then only use one of them?

Over a given basin and for a participating model *k* $Ragg_k$ refers to the spatial aggregate of simulated runoff across a basin, while q_k refers to modelled streamflow at a grid cell underlying a stream gauge.
We have now clarified these terms in the text.

> *where* $j \in [1, J]$ *are the time steps and* $k \in [1, K]$ *represent the participating models,* x_k^j *(i.e., integrated runoff* $Ragg_k^j$ *over the basin areas in small basins and modelled streamflow at a gauge location* q_k^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.*

page 5, line 2: not really clear what you mean with this sentence. what is the number of records?

The number of records refers to the total number of available monthly observations available for a basin. We have now clarified this in the text

> *To avoid over-fitting when applying the weighting approach, we limited the number of participating models so that the ratio of number of records (i.e. total number of available monthly observations within the period of study) to number of models does not fall below ten.*

page 5, line 4: An alternative approach for dealing with negative values might be to log transform runoff/stream flow before doing the computations and back transform it in the end. See e.g. Gudmundsson & Seneviratne (2015, doi: 10.5194/hess-19-2859-2015; 2016, doi:10.5194/essd-8-279-2016). Again, my apologies for self-citation.

Thanks for sharing this study. This is something to look at in the future versions of LORA and requires testing its applicability with our methods for deriving the uncertainty estimates.

page 5, line 10: Some more details on the transformation process would be appreciated.

I know it is published in Hobeichi et al. (2018) but it would make the paper easier to understand if it would be outlined in more detail here.

We thank the reviewer for his comments, we have addressed this earlier in this document. Please see our response to (5).

page 5, line 20: transfer of "weights" from the 3 most similar basins; why 3 and not a larger/smaller number? Any empirical motivation for this? Would optimizing this hyper-parameter help to get even better results?

We agree this is a subjective choice. The dissimilarity technique has been previously applied to find 10 donors for 1 receptor. Given that all the selected donors must have very close similarity indices, we found by trial and error that increasing the number of donor basins might introduce donor basins that have a significantly different similarity index, and that setting the number of donor basins to three seemed most appropriate. Informal optimisation. We have now clarified this in the text:

> *The dissimilarity technique has been previously applied to find ten donors for one receptor. Given that all the selected donors must have very close similarity indices, we found by trial and error that increasing the number of donor basins might introduce donor basins that have a significantly different similarity index, and that setting the number of donor basins to three seemed most appropriate. Informal optimisation.*

page 6. line 8: Note that "Tundra" and "Subarctic" will not always have permanent ice!
Thanks for spotting this, we have now changed the text to read:

> *areas covered with ice during most of the year (defined by climate zones Tundra, Subarctic and Ice cap)*

Also: The chosen climate zone map is very uncommon, and most readers will not be familiar with it. Therefore, it needs to be presented in a figure. Alternative: why not use a common climate-zone definition (e.g. Köppen-Geiger?)
We used this particular climate map because it comprises only 12 broad climate classes (compared to more than 30 in other climate maps e.g. Köppen–Geiger). This reduced the divisions made to large heterogenous basins, while ensuring that the resultant basin zones within individual basins have very distinct climate characteristics.

We have now clarified this in the text and provided a figure of this climate map in the supplemental material

> *We used this particular climate map because it comprises only 12 broad climate groups (compared to more than 30 in other climate maps e.g. Köppen–Geiger). This reduced the divisions made to large heterogenous basins, while ensuring that the resultant basin zones of individual basins have very distinct climate characteristics.*

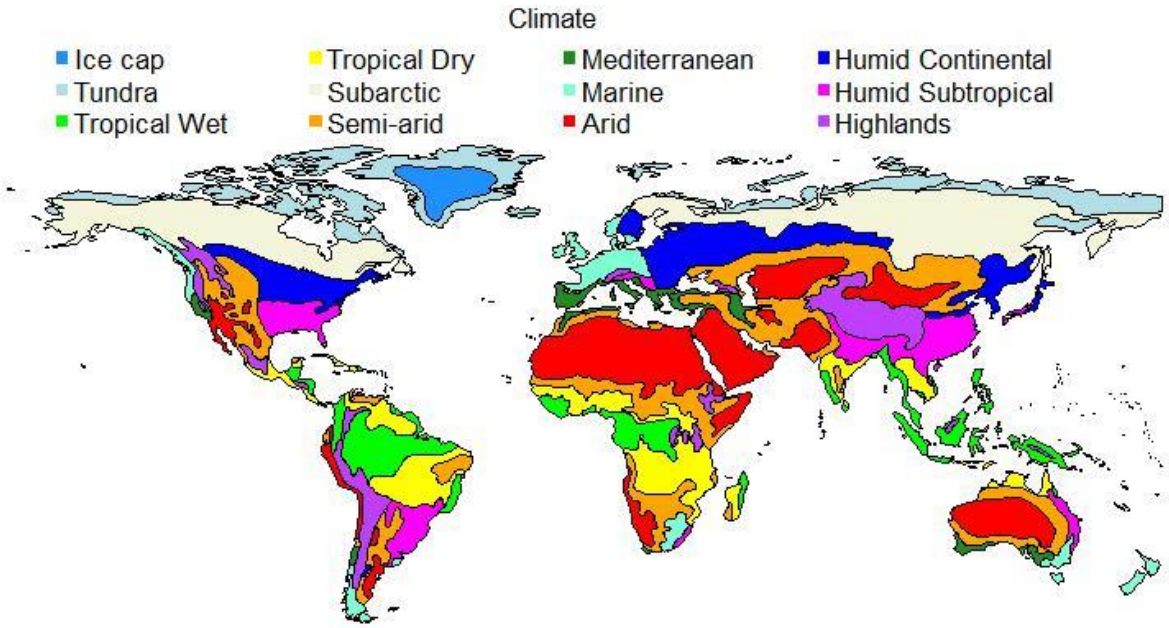


Figure S1: Climate map used in this study (available from ArcGIS online). It is a simplified climate zones map consisting of 12 broad climate classes

page 7. l 15: should be fig 4 not fig 3
Thanks for spotting this. We made the changes in the text

Fig 4: I found the "relative" improvement difficult to grasp upon first reading. Suggestion: just show the performance for WPin WPout (and omit individual models, as they are shown later)
We agree with the reviewer that the plot in Figure 4 needs careful examination, and that one might think that Figure 4 and Figure 5 (now Figure 6) provide redundant information. However, while both Figures look similar, each achieves a different purpose. For instance, the plot in Figure 4 provides evidence that our approach (i.e. using transferred weights from the most similar gauged basins to derive runoff estimates at the receptor basins) is succeeding by the fact that WP_{out} offers improvement over the individual products. On the other hand, Figure 5 (now Figure 6) compares the performance of the individual product with that of WP_{in} (i.e. partially LORA) over the gauged basins. We therefore believe that it is worth providing both plots to the reader.

page 9, line 18:
(1) "mean runoff uncertainty". What was exactly calculated? Note that simply averaging sqrt(sigma) will yield wrong results. Instead the rules for propagation of uncertainty should be used (https://en.wikipedia.org/wiki/Propagation_of_uncertainty). If the authors did already consider this, I apologize for this comment.
(2) Note that there are strong perceptions on what "reliability" means in ensemble forecasting.
See e.g. "reliability diagrams" and "reliability histograms"
page 9, line 31 & Fig 8: how did you compute the mean uncertainty? (refer to my

comment above)

Great point. We understand that averaging sqrt(sigma) cannot be used to derive the uncertainty of runoff fields from the uncertainty of streamflow or vice versa, but rather there are various rules for propagation of uncertainty that can be applied. In our response to (5) We have now explained our method for deriving uncertainty estimates for both runoff fields and their spatial aggregate (streamflow) which had not been clearly detailed in the manuscript. Here "mean runoff uncertainty" refers to the seasonal mean runoff uncertainty (climatology), we agree with the reviewer that this should be specified to avoid confusion with the spatial mean of uncertainty which does not really make sense. We have now clarified this in the text:

> *We calculated the seasonal relative uncertainty expressed as the ratio of the seasonal average uncertainty to seasonal mean runoff (i.e.* $\frac{mean\ runoff\ uncertainty}{mean\ runoff}$*) over the period 1980 – 2012.*

page 10, line 31: not Fig 3...

Thanks for spotting this out, we have now fixed this in the text.

Response to Reviewer #2

(1) I miss a few explicit examples explaining why runoff is at all useful, especially at 0.5 degrees. The argument is that accurate estimates of runoff are critical to inform climate change adaptation strategies, to guide appropriate water management in agriculture and to enable the assessment of the impact of anthropogenic activities on ecosystems. However, what does runoff at 50x50 km resolution even mean? It is in terms of scale too far off from being operationally relevant. Thus, a stronger justification using examples is called for.

We thank the reviewer for their suggestion: We have replaced the text with:

> *Characterizing its dynamics and magnitudes is a major research aim of hydrology and hydrometeorology and a critical importance to improve our understanding of the current conditions of the large-scale water cycle and predict its future states. More accurate estimates also provide additional constraint for climate model evaluation.*

(2) I also miss some references to the earliest work on runoff fields, e.g. Fekete et al. 2002: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/1999GB001254

We agree with the reviewer, the study of Fekete et al. (2002) is an important example of how streamflow observation and model outputs can be combined to generate runoff fields. We have now referred to their study in the text

> *…, 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 (Fekete et al., 2002; Ye et al., 2014),*

(3) Regarding to previous work and scales, I would like to call the authors' attention to a recent publication by Barbarossa et al.(https://www.nature.com/articles/sdata201852). They provide discharge estimates at 1 km resolution. I move that these are probably more representative for local runoff than those obtained from GHMs at half degree resolution?

We thank the reviewer for pointing out this study. We of course agree that FLO1k better represents small streams due to its higher spatial resolution. However, FLO1k does not necessarily provide more accurate estimates for large rivers. Additionally, FLO1k only provides information about the mean, minimum, and maximum annual flow, which limits its usefulness. In contrast, LORA provides valuable information about flow timing and the seasonal runoff distribution. The two datasets are thus quite different and in some ways complementary. In the revised paper we now cite Barbarossa et al. (2018).

(4) Section 2.3: I find that there is too little info on the method used to compute the weights. I don't think that it should be necessary read another paper to comprehend the essentials

of the methods used. So, I would want some more explanation on how the weights are calculated.

As noted in our response to Reviewer 1 and below, we have included significantly more detail about this in the revised manuscript.

For instance:
- How are correlations between models accounted for?
An error covariance matrix is calculated for the participating models. The weights are functions of this error covariance matrix. We provide more details below
- Are the weights allowed to be negative?
Yes, weights can be negative

- Is the sum of the weights adding up to one?
Yes, the weights add up to one.
If this is the case, one has to add another equation and transform a constrained optimization to an unconstrained one using Lagrange multipliers.
This is correct, we constrained the weights to sum up to one, and we transformed the problem of minimizing $\sum_{j=1}^{J}(\mu_q^j - Q^j)^2$ to a problem of minimizing a function that involves a Lagrange multiplier.

We have now detailed the weighting method in the text:

> *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: $\mu_q^j = \sum_{k=1}^{K} w_k(x_k^j - b_k)$ where $j \in [1, J]$ are the time steps and $k \in [1, K]$ represent the participating models, x_k^j (i.e., integrated runoff $Ragg_k^j$ over the basin areas in small basins and modelled streamflow at a gauge location q_k^j in large basins) is the value of the participating dataset in m3 s-1 at the j^{th} time step of the kth 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 $\sum_{j=1}^{J}(\mu_q^j - Q^j)^2$ subject to the constraint that $\sum_{k=1}^{K} w_k = 1$, where Q^j is the observed streamflow at the j^{th} time step. This minimization problem can be solved using the method of Lagrange multipliers by finding a minima for*
>
> $$F(w, \lambda) = \frac{1}{2}\left[\frac{1}{(J-1)}\sum_{j=1}^{J}(\mu_q^j - Q^j)^2\right] - \lambda\left(\left(\sum_{k=1}^{K} w_k\right) - 1\right).$$
>
> *The solution to the minimization of $F(w, \lambda)$ can be expressed as $= \frac{A^{-1}1}{1^T A^{-1}1}$, where $1^T = \overbrace{[1,1,\ldots,1]}^{k\ elements}$ and A is the $k \times k$ error covariance matrix of the participating datasets*

(after bias correction), i.e. A=$\begin{pmatrix} c_{1,1} & \cdots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \dots & c_{k,k} \end{pmatrix}$*. A is symmetric and the term* $c_{a,b}$ *is the covariance of the* a^{th} *and* b^{th} *bias corrected dataset after subtracting the observed dataset, while each diagonal term* $c_{k,k}$ *is the error variance of dataset k. We note here that the solution presented here is based on the performance of the participating products (diagonal terms of A) and the dependence of their errors (accounted for by the non-diagonal terms of A). 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:

$$\mu_r^j = \sum_{k=1}^{K} w_k (r_k^j - b'_k)$$

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

Line 8-10 page 5: This seems to assume that travel times are less than a month because it neglects routing?

Yes, this is worth mentioning. It is a limitation that has possibly led to an overestimation in the computed uncertainties over large basins. We have now added this to the text.

Given that there are no direct observations for runoff, uncertainties were computed from the discrepancy between the modelled runoff aggregates and observed streamflow. This ignored the lag time between LORA integrated runoff and observed streamflow at the mouth of the river and induced biases that possibly led to overestimated uncertainty over large gauged basins.

(6) Lines 11-13 page 5: "It provides better estimates than simply calculating the standard deviation of the involved products". Is that really the case? If your weighting method is optimal and you have bias-corrected correctly would the following estimator for each pixel not be unbiased (i and j are different products bias corrected):

$$\hat{\sigma}_r^2 = \frac{1}{N \times 11 \times 11} \sum_{i=1}^{11} \sum_{j=1}^{11} \sum_{t=1}^{T} w_i w_j (r_{i,t} - \mu_r)(r_{j,t} - \mu_r)$$

By moving a window T over time you get your time varying variance

We thank the reviewer for his suggestion. We think that while the suggested formula provides time varying uncertainty estimates associated to the weighted runoff, it does not account for the dependence between $r_{i,t}$ and $r_{j,t}$ which is likely to lead to an overestimation of uncertainty. Meanwhile, the ensemble dependence transformation process that we applied in this paper to the participating products transforms the dependent estimates to statistically independent estimates.

We don't really have observations for runoff, so we can't test our method for deriving uncertainty on runoff, however we can test it on streamflow (i.e. runoff aggregates). We have performed out-of-sample tests to show that the distribution of the errors over the gauged basin is similar to the distribution of their errors when they are considered ungauged. We have now explained how we have performed this test and showed the results in the manuscript:

> *The uncertainty estimates computed at the gauged basins represent the deviation of (the spatial aggregate of) our weighted product ($Ragg_{\mu}$) from the observed streamflow, since the in-sample uncertainty estimates are calculated from the variance of the transformed ensemble, which by design equals MSE of $Ragg_{\mu}$ against observations (i.e. error variance of $Ragg_{\mu}$). To test if the uncertainty estimates perform well out-of-sample (i.e. at the ungauged basins), we took a gauged basin, but instead of constraining the weighting using observed streamflow from this basin, we constructed model weights by using the three most similar donor basins. We could then calculate MSE of $Ragg_{\mu}$ against observations from the three donor basins, denoted by MSE_{in}, which provides us with the uncertainty estimates calculated in-sample ($\sqrt{MSE_{in}}$), since the observational data used in this case is the same dataset that was used to train the weighting. We also calculated the MSE of the aggregated weighted product against the actual observation of the gauged basin and denoted this MSE_{out}. $\sqrt{MSE_{out}}$ represents the uncertainty estimates computed out-of-sample, since the comparison was performed against observational data that has not been used to train the weighting. We repeated the out-of-sample test for all the gauged basins.*
>
> *We displayed the results of the out-sample-test by showing the ratio $\sqrt{MSE_{out}}$ /$\sqrt{MSE_{in}}$. If the approach is succeeding, we expect that this ratio is around one, indicating that the values of MSE_{in} and MSE_{out} are close to each other. We used a box and whisker plot, where each sample is a different basin, to show the results.*

We have also commented on the results:

> 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. *Furthermore, the boxplot in Fig 5 shows that, overall, when the uncertainty estimates are computed out-of-sample they are very similar to what they would have been if they were computed in-sample. This demonstrates that the dissimilarity technique can be effectively used to derive not only the weighting product but also its associated uncertainties at the ungauged basin.*

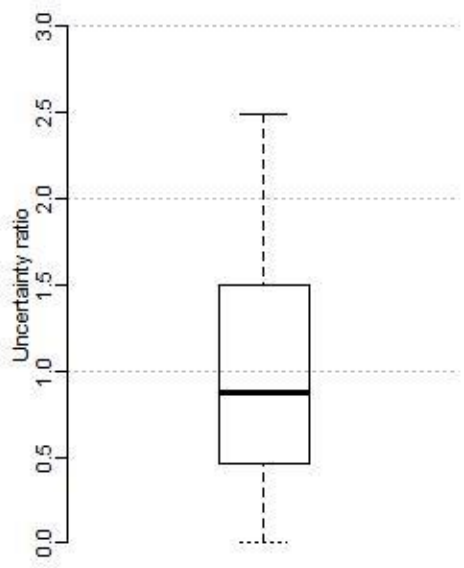


Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample to (2) the uncertainties of the spatial aggregate of the weighted computed in-sample.

(7) Line 20-25 page 5: transferring the weights from donors to receptors. Are one set of weights obtained jointly for the three donor catchments or are three sets of weights averaged and transferred?

Yes, this was not clear in the text. One set of weights is obtained jointly from the three donor catchments. We clarified this in the text.

> *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. This resulted in one set of weights and bias ratios obtained jointly from the three donor basins.*

(8) Line 11 on page 10. Why compare only with VIC? Why not GLDAS (4 models).
Good question. At the time of analysis, GLDAS version1 model outputs had either a very short common period with LORA or a coarse resolution (1°) and showed a significant disagreement with observation when we interpolated them to a 0.5° grid. We clarified this in the text.

> *Other global estimates of total runoff are also available such as GLDAS and Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger et al., 2016), however we haven't compared LORA with these datasets because they either have a short common period with LORA, or a coarser resolution (i.e. 1o) and showed a significant disagreement with observation when interpolated to a 0.5° grid.*

(9). Table 2: why not add the estimated total runoff volumes from GRDC (also globally in km^3). It would be good to see what the global runoff volume is in this product compared to other estimates.

Good idea. We have now added the average total yearly volume of discharged water from LORA and observation

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 yearly volume of LORA runoff aggregates (i.e. flow in Km^3) and observed annual flow (Km^3) over the basins and 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. Observed annual flow is given only if data from all contributing stations is available over a whole year over for at least 17 years out of 33 years covered in this study.

Basin	VIC mm/year	VIC adjusted for water budget closure mm/year	LORA (Runoff) mm/year	LORA (uncertainty) mm/year	LORA yearly flow ± uncertainty Km^3	Observed yearly flow Km^3	Dominant climate
Amazon	1048	1029	1151	360	6763 ± 2115	-	Tropical wet
Amur	135	129	219	115	428 ± 225	325	Humid continental and semi-arid
Columbia	318	293	333	101	218 ± 66	209	Semi-arid and highlands
Congo	407	404	358	147	1292 ± 532	1240	Tropical wet and tropical dry
Danube	272	265	260	125	199 ± 95	205	Marine Humid, continental and humid subtropical
Indigirka	132	120	228	171	78 ± 59	53	Subarctic
Lena	142	134	301	137	731 ± 332	557	Subarctic
Mackenzie	189	173	191	110	323 ± 186	294	Subarctic
Mississippi	220	215	212	123	616 ± 359	581	Humid continental and humid subtropical
Murray-Darling	42	41	15	6	12 ± 5	-	Arid and semi-arid
Niger	198	194	106	41	239 ± 87	170	Arid, semi-arid and tropical dry
Olenek	114	106	230	208	48 ± 43	40	Subarctic
Parana	278	279	189	97	471 ± 247	600	Marine and humid subtropical

Pechora	*342*	*308*	*420*	*420*	*131 ± 131*	*153*	*Tundra and subarctic*
Yenisei	*217*	*195*	*324*	*203*	*828 ± 520*	*612*	*Subarctic*
Yukon	*149*	*139*	*229*	*102*	*188 ± 83*	*214*	*Subarctic*

(10) Line 27 on page 10: reduced performance in dry climates. Apart from the reasons mentioned, another possible cause could be the fact that GHMs are probably less proficient in representing runoff processes in arid basins where often runoff is local and will not always be turned into streamflow

Great point. We have now added this in the text to read:

> *It follows from Fig. 8 that the runoff values computed over dry climates tend to be less reliable than those in other regimes. This is perhaps due to biases in the WFDEI precipitation forcing that are propagated and intensified in the simulated runoff (Beck et al., 2017a). Another possible reason is the reduced proficiency of models in representing runoff dynamics in arid climates where runoff tends to be highly non-linearly related to rainfall and often evaporates locally without reaching a river system (Ye et al., 1997).*

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 0.5° 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.

1 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). Characterizing its dynamics and magnitudes is a major research aim of hydrology and hydrometeorology and a critical importance to improve our understanding of the current conditions of the large-scale water cycle and predict its future states. More accurate estimates also provide additional constraint for climate model evaluation., 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; Falcone et al., 2010) 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;Bierkens, 2015; 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 (Fekete et al., 2002; 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 but not limited to 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 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.

2 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.

We detail in sections 2.3 and 2.4 our methods to derive the weighted runoff product for the global land. A flowchart summarizing the process is provided in Fig. 2.

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 $j \in [1, J]$ are the time steps and $k \in [1, K]$ represent the participating models, x_k^j (i.e., integrated runoff Ragg_k^j over the basin areas in small basins and modelled streamflow at a gauge location q_k^j in large basins) is the value of the participating dataset in m3 s-1 at the j^{th} time step of the kth participating model, the bias term b_{k} is the mean error of x_{k} in $\mathrm{m^3\ s^{-1}}$. The set of weights w_{k} provides an analytical solution to the minimization of $\sum_{\mathrm{j}=1}^{\mathrm{J}}(\mu_{\mathrm{q}}^{\mathrm{j}} - Q^{\mathrm{j}})^2$ subject to the constraint that $\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{w_k} = 1$, where Q^j is the observed streamflow at the j^{th} time step. This minimization problem can be solved using the method of Lagrange multipliers by finding a minima for

$$F(w, \lambda) = \frac{1}{2}\left[\frac{1}{(J-1)}\sum_{j=1}^{J}(\mu_q^j - Q^j)^2\right] - \lambda\left(\left(\sum_{k=1}^{K} w_k\right) - 1\right).$$

The solution to the minimization of $F(w, \lambda)$ can be expressed as $\mathrm{w} = \frac{\mathrm{A^{-1}1}}{\mathrm{1^T A^{-1} 1}}$, where $1^{\mathrm{T}} = \overbrace{[1,1,\ldots,1]}^{\text{k elements}}$ and A is the k × k error covariance matrix of the participating datasets (after bias correction), i.e. $\mathrm{A}=\begin{pmatrix} c_{1,1} & \cdots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \ldots & c_{k,k} \end{pmatrix}$. A is symmetric and the term $c_{a,b}$ is the covariance of the $\mathrm{a^{th}}$ and $\mathrm{b^{th}}$ bias corrected dataset after subtracting the observed dataset, while each diagonal term $c_{k,k}$is the error variance of dataset k. We note here that the solution presented here is based on the performance of the participating products (diagonal terms of A) and the dependence of their errors (accounted for by the non-diagonal terms of A). 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:

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

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 $\mathrm{b'_k}$ is its runoff bias in $\mathrm{kg\ m^{-2} s^{-1}}$.

To calculate the runoff bias $\mathrm{b'_k}$, 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 $\mathrm{Ragg_k^j}$ instead of the modeled streamflow $q_{\mathrm{k}}^{\mathrm{j}}$ Eq. (2).

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

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

We note that there no empirical evidence in the literature that the assumptions presented in Eq. 1 and Eq. 2 are valid. However, given that these assumptions constitute a part of our overall approach that we tested and proved its success later in this paper, the validity of these assumptions is very likely to hold true.

To avoid over-fitting when applying the weighting approach, we limited the number of participating models so that the ratio of number of records (i.e. total number of available monthly observations within the period of study) to number of models does not fall below ten. 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. Table S1 shows examples of weights and bias ratios calculated for the participating models over a range of river basins. It shows that HBVS, JULES1, JULES2 and SURF2 didn't participate in the weighting over the large basins (i.e. Amur, Indigirka, Mississippi, Murray-Darling, Olenek, Parana, Pechora and Yenisei) since these models don't have estimates for streamflow which are needed to construct the weights over large basins. For the smaller Copper River basin, however, runoff estimates from all models participated in deriving weighted runoff estimates. Table S1 also shows that in many cases, models were assigned negative weights. While this might not be expected in typical performance-based weighting, it is possible when weighting is based on error covariance as well as their performance differences in this formulation. We show below how the weights can be modified to non-negative weights.

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. For any given gauged basin, we first calculated the spatial aggregate of our weighted runoff estimate Ragg_{μ}, then quantified ${s_q}^2$, the error variance of Ragg_{μ} with respect to the observed streamflow Q over time as:

$${s_q}^2 = \frac{\sum_{j=1}^{J}(\mathrm{Ragg_{\mu}}^j - Q^j)^2}{J-1}$$

Then, we wished to guarantee that the variance of the constituent modelled estimate ${\sigma_q^2}^j$ about ${\mathrm{Ragg}_{\mu}}^j$ at a given time step, averaged over all time steps where we have available streamflow data, is equal to ${s_q}^2$, such as ${s_q}^2 = \frac{1}{J}\sum_{j=1}^{J}{\sigma_q^2}^j$.

Since the variance of the existing constituent products do not, in general, satisfy this equation. We transformed them so that it does. This involved first modifying the set of weights w to a new set $\widetilde{w}$ such that

$\widetilde{w} = \frac{w^T + (\alpha-1)\frac{1^T}{K}}{\alpha}$, where $\alpha = 1 - Kmin(w_k)$ and $min(w_k)$ is the smallest negative weight (and α is set 1 if all w_k are non-negative). This ensures that all the modified weights $\widetilde{w}_k$ are positive. We then transform the individual estimates x_k^j to $\tilde{x}_k^j$ where $\tilde{x}_k^j = {\mathrm{Ragg}_{\mu}}^j + \beta(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - {\mathrm{Ragg}_{\mu}}^j)$ and $\beta = \sqrt{\frac{s_q^2}{\frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\widetilde{w_k}(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - {\mathrm{Ragg}_{\mu}}^j)^2}}$.

The weighted variance estimate of the transformed ensemble can be defined as $\sigma_q^{2j} = \sum_{k=1}^{K} \widetilde{w}_k(\tilde{x}_k^j - \text{Ragg}_\mu{}^j)^2$ and ensures that the equation $\frac{1}{J}\sum_{j=1}^{J} \sigma_q^{2j} = s_q^2$ holds true. Furthermore, $\sqrt{\sigma_q^{2j}}$ is the temporally varying estimate of uncertainty standard deviation of the transformed ensemble that (a) is varying in time, and (b) accurately reflects our ability to reproduce the observed streamflow. We refer the reader to Bishop and Abramowitz (2013) for proofs.

In order to estimate $\sqrt{\sigma_r^{2j}}$, the uncertainty of the runoff attributes μ_r^j at each point in time and space, we first transformed the runoff fields r_k^j to $\tilde{r}_k^j$ by applying the same transformation parameters α and β such that $\tilde{r}_k^j = \mu_r^j + \beta(\bar{r}^j + \alpha(r_k^j - \bar{r}^j) - \mu_r^j)$. We then calculated the error variance $\sigma_r^{2j} = \sum_{k=1}^{K} \widetilde{w}_k(\tilde{r}_k^j - \mu_r^j)^2$.

Finally, we used $\sqrt{\sigma_r^{2j}}$ as the spatially and temporally varying estimate of runoff uncertainty standard deviation, which we will refer to below simply as 'uncertainty'. It provides a much more defensible uncertainty estimate than simply calculating the standard deviation of the involved products.

We note that for a given basin, $\sqrt{\sigma_q^{2j}}$ represents the uncertainty of the modelled streamflow i.e. $\text{Ragg}_\mu{}^j$, while $\sqrt{\sigma_r^{2j}}$ represents the uncertainty of modelled runoff at each grid cell across the basin. This means that at every time step, there is one value for $\sqrt{\sigma_q^{2j}}$ per basin, and one value for $\sqrt{\sigma_r^2}$ per grid across the basin.

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. This resulted in one set of weights and bias ratios obtained jointly from the three donor basins. 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:

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

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. $Z_{p,a}$ and $Z_{p,b}$ are the values of the characteristic p at donor and receptor basins, respectively. 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 covered with ice during most of the year (defined by climate zones Tundra, Subarctic and Ice cap using a simplified climate zones map (Fig. S1) 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 three most similar donors for every receptor basin. The dissimilarity technique has been previously applied to find ten donors for one receptor. Given that all the selected donors must have very close similarity indices, we found by trial and error that increasing the number of donor basins might introduce donor basins that have a significantly different similarity index, and that setting the number of donor basins to three seemed most appropriate.

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 climate zones map; Esri Education Team, 2014) defined above. We used this particular climate map because it comprises only 12 broad climate groups (compared to more than 30 in other climate maps e.g. Köppen–Geiger). This reduced the divisions made to large heterogenous basins, while ensuring that the resultant basin zones of individual basins have very distinct climate characteristics. Figure 3 shows the spatial coverage of the donor basins, receptor basins and non-donor and non-receptor basins.

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 (WP_{in}) derived by using observed streamflow for this basin to weight models; (c) an out-of-sample weighted product (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 WP_{in}, 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= $\sigma_{\text{Ragg}} - \sigma_{\text{observed streamflow}}$. 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.

The uncertainty estimates computed at the gauged basins represent well the deviation of the spatial aggregate of our weighted product ($Ragg_\mu$) from the observed streamflow, since the in-sample uncertainty estimates are calculated from the variance of the transformed ensemble, which by design equals MSE of $Ragg_\mu$ against the observation (i.e. error variance of $Ragg_\mu$). To test if the uncertainty estimates perform well out-of-sample (i.e. at the ungauged basins), we performed another out of sample test. In this test, we took a gauged basin, but instead of constraining the weighting using observed streamflow from this basin, we constructed model weights by using the three most similar donor basins. We could then calculate MSE of $Ragg_\mu$ against observation from the three donor basins, and we denoted this by MSE_{in}, which represents the uncertainty estimates calculated in-sample, since the observational data used in this case is the same dataset that was used to train the weighting. We also calculated the MSE of the aggregated weighted product against the actual observation of the gauged basin and we denoted this by MSE_{out}. MSE_{out} represents the uncertainty estimates computed out-of-sample, since the comparison was performed against observational data that has not been used to train the weighting. We repeated the out-of-sample test for all the gauged basins.

We displayed the results of the out-sample-test by showing the ratios of MSE_{in} to MSE_{out}. If the approach is succeeding, we expect that this ratio is around one, indicating that the values of MSE_{in} and MSE_{out} are close to each other. We used a box and whisker plots to show the results.

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. 4 (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. Furthermore, the boxplot in Fig. 5 shows that, overall, when the uncertainty estimates are computed out-of-sample they are very similar to what they would have been if they were computed in-sample. This demonstrates that the dissimilarity technique can be effectively used to derive not only the weighting product but also its associated uncertainties at the ungauged basin.

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. 6 (a-d) display the results.

The RMSE plot in Fig. 6 (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. 6 (b)) across more than half of the basins. The Mean bias plot in Fig. 6 (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. 6 (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. 6 (c-d) but this time for the snow-dominated basins only. We display the results in Fig. 7.

The temporal correlation plot in Fig. 7 (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. 6 (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 7 (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. 6 (a and b), but we found no noticeable differences in the performance of any of the products relative to that found globally in Fig. 6 (a and b). The results displayed in Fig. 6 and Fig. 7 are discussed further below.

We calculated the seasonal relative uncertainty expressed as the ratio of the seasonal average uncertainty to seasonal mean runoff (i.e. $\frac{\text{mean runoff uncertainty}}{\text{mean runoff}}$) over the period 1980 – 2012. This metric is intended to show some indication of the reliability of the derived runoff, with results displayed in Fig. 8. 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. 8 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 9 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 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.

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 as well as adjusted VIC estimates after enforcing the physical constraints of the water budget 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). We also showed the average annual volume of water that discharges from these basins computed from LORA and the observational data.

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 Indigirka, Olenek, Parana, Pechora, Yenisei and Yukon. Large discrepancies between VIC and LORA are found in Lena and the Murray-Darling. Other global estimates of total runoff are also available such as GLDAS and Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger et al., 2016), however we haven't compared LORA with these datasets because they either have a short common period with LORA, or a coarser resolution (i.e. 1°) and showed a significant disagreement with observation when interpolated to a 0.5° grid.

Finally, we provide in Fig. S8 an example of runoff fields in an individual month (e.g. May, 2003).

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. 8 that the runoff values computed over dry climates tend to be less reliable than those in other regimes. This is perhaps due to biases in the WFDEI precipitation forcing that are propagated and intensified in the simulated runoff (Beck et al., 2017a). Another possible reason is the reduced proficiency

of models in representing runoff dynamics in arid climates where runoff tends to be highly non-linearly related to rainfall and often evaporates locally without reaching a river system (Ye et al., 1997). 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. 4 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. 6 (c) and Fig. 7 (a), and also explains the highest weights attributed to LISFLOOD in the majority of snow-dominated basins (Table S1). Because of this, and because LISFLOOD tends to overestimate runoff across half of the snow-dominated basins (as shown in Fig.7 (b)) LORA exhibits a positive bias across half of the snow-dominated basins (Fig. 7 (b)) and particularly in Lena, Amur and Yenisei basins (Fig. 9).

Further, we provide in Fig. S2 the spatial distribution of correlation results from Fig. 6 (c). The basins are colour-coded by their temporal correlation with the observed streamflow and the number of basins in each category is given. Basins in yellow are those where LORA is highly correlated with the observation while dark blue basins are those where LORA exhibits a negative correlation with the observation. It can be noted from Fig. 6(c) that occurrence of negative correlation is extremely unusual which explains why these were considered outliers and were not shown in the box and whisker plot. Likely, low correlation basins are unusual and constitute less than 12% of the number of basins (excluding basins with negative correlation). Also, the median value is above 0.8, which is higher than any constituent estimates. We selected a basin from each correlation range and examined the timeseries of LORA and the observed streamflow more closely (Fig. S3-S7), in particular illustrating the uncertainty estimate of LORA. In Ganges, LORA captures well the observed time-series dynamic with a tendency to over-estimate streamflow peak in August (Fig. S3). Over Madeira basin, LORA is able to represent reasonably well most of the climatic variability found in the observation (Fig. S4). In Congo, the catchment has an irregular time-series dynamic, LORA is in principle able to capture a large part of the climatic variability in the observation (Fig. S5). In Lena, the observation shows a peak in June and a second less significant peak in September (Fig. S6). Both peaks are captured by LORA during most of the time series with a tendency to underestimate the late summer peak and overestimate the early summer peak. In the upper Indus, LORA does not capture the magnitudes of observed streamflow and shows a reversed seasonal cycle which explains why it exhibits negative correlation with the observation (Fig.S7). Zhang et al. (2018) found disagreement between simulated runoff from three LSMs and observed streamflow over Indus basin which they expected to be due to errors in the observational data from GRDB dataset.

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. 8). 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 10 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. 9), 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. Also, given that there are no direct observations for runoff, uncertainties were computed from the discrepancy between the modelled runoff aggregates and observed streamflow. This ignored the lag time between LORA integrated runoff and observed streamflow at the mouth of the river and induced biases that possibly led to overestimated uncertainty over large gauged 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 Data availability

LORA v1.0 can be downloaded from geonetwork.nci.org.au and its DOI is http://dx.doi.org/10.25914/5b612e993d8ea

8 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.

9 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, Department of Physical Geography, Utrecht University, Utrecht, The Netherlands. [online] Available from: http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf, 2008.

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.

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resour. Res., 51, 4923–4947, doi:10.1002/2015WR017173, 2015.

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

Burek, P., van der Knijff, J., de Roo, A.: LISFLOOD, distributed water balance and flood simulation model revised user manual, Joint Research Centre of the European Commission., 2013.

Dai, A.: Historical and Future Changes in Streamflow and Continental Runoff: A Review, Terr. Water Cycle Clim. Chang. Nat. Human-Induced Impacts, Geophys. Monogr., 221, 17–37, doi:10.1002/9781118971772.ch2, 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, r a Heal. Ctry. Natl. Res. Flagsh., 3(5), 100 pp, 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.

Fekete, B. M., Vörösmarty, C. J. and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cycles, 16(3), 15-1-15–10, doi:10.1029/1999GB001254, 2002.

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, Hydrol. Earth Syst. Sci, 22(2), 1317–1336, doi:https://doi.org/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, Hydrol. 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. Atmos., 99(D7), 14415--14428, 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., 48(1), 1–12, doi:10.2166/nh.2016.166, 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., 17, 3707–3720, doi:10.5194/hess-17-3707-2013, 2013.

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, Report to the National Land and Water Resources Audi., 2000.

Rantz, S. E.: Measurement and computation of stream flow. Volume 2: Computation of discharge, US Geol. Surv. water-supply Pap., 2(2175), 631 pp, 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, 9(2), 389–413, doi:https://doi.org/10.5194/essd-9-389-2017, 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), 1–5, 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.

Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M. and Jakeman, A. J.: Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resour. Res., 33(1), 153–166, 1997.

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.

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.

Huntzinger, D.N., C.R. Schwalm, Y. Wei, R.B. Cook, A.M. Michalak, K. Schaefer, A.R. Jacobson, M.A. Arain, P. Ciais, J.B. Fisher, D.J. Hayes, M. Huang, S. Huang, A. Ito, A.K. Jain, H. Lei, C. Lu, F. Maignan, J. Mao, N. Parazoo, C. Peng, S. Peng, B. Poulter, D.M. Ricciuto, H. Tian, Xiaoying Shi, W. Wang, N. Zeng, F. Zhao, Q. Zhu, J. Yang, and B. Tao. 2016. NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (version 1) in Standard Format. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1225.

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, 2008)
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 yearly volume of LORA runoff aggregates (i.e. flow in Km^3) and observed annual flow (Km^3) over the basins and 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. Observed annual flow is given only if data from all contributing stations is available over a whole year over for at least 17 years out of 33 years covered in this study.

Basin	VIC mm/year	VIC adjusted for water budget closure mm/year	LORA (Runoff) mm/year	LORA (uncertainty) mm/year	LORA yearly flow ± uncertainty Km^3	Observed yearly flow Km^3	Dominant climate
Amazon	1048	1029	1151	360	6763 ± 2115	-	Tropical wet
Amur	135	129	219	115	428 ± 225	325	Humid continental and semi-arid
Columbia	318	293	333	101	218 ± 66	209	Semi-arid and highlands
Congo	407	404	358	147	1292 ± 532	1240	Tropical wet and tropical dry
Danube	272	265	260	125	199 ± 95	205	Marine Humid, continental and humid

							subtropical
Indigirka	132	120	228	171	78 ± 59	53	Subarctic
Lena	142	134	301	137	731 ± 332	557	Subarctic
Mackenzie	189	173	191	110	323 ± 186	294	Subarctic
Mississippi	220	215	212	123	616 ± 359	581	Humid continental and humid subtropical
Murray-Darling	42	41	15	6	12 ± 5	-	Arid and semi-arid
Niger	198	194	106	41	239 ± 87	170	Arid, semi-arid and tropical dry
Olenek	114	106	230	208	48 ± 43	40	Subarctic
Parana	278	279	189	97	471 ± 247	600	Marine and humid subtropical
Pechora	342	308	420	420	131 ± 131	153	Tundra and subarctic
Yenisei	217	195	324	203	828 ± 520	612	Subarctic
Yukon	149	139	229	102	188 ± 83	214	Subarctic

Figures

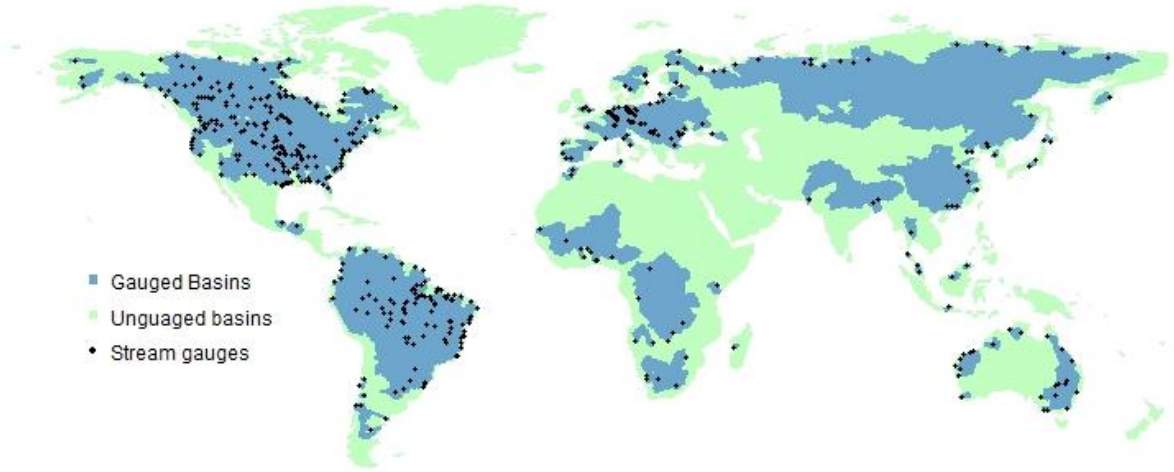


Figure 1: Spatial coverage of gauged and ungauged river basins and location of stream gauges.

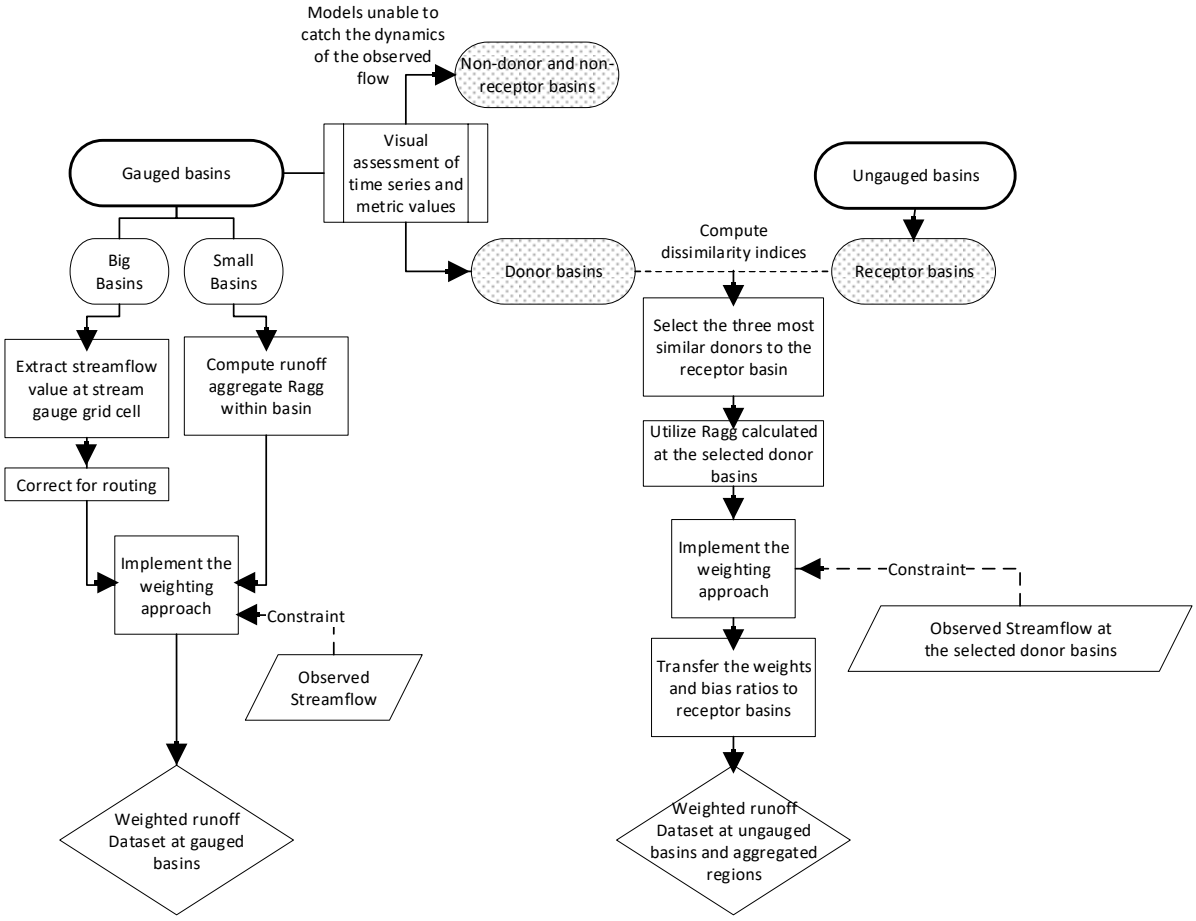


Figure 2. Flowchart summarizing the steps carried out to derive the weighted runoff product for the global land surface.

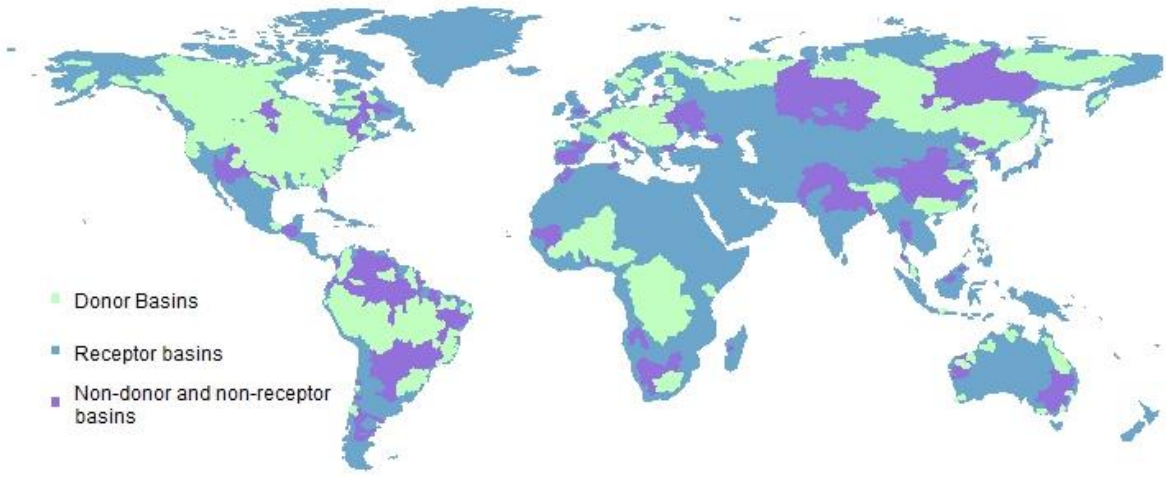
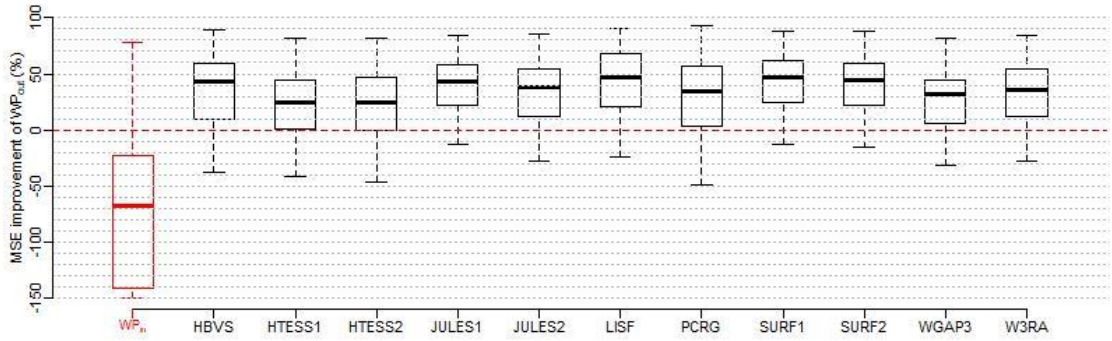
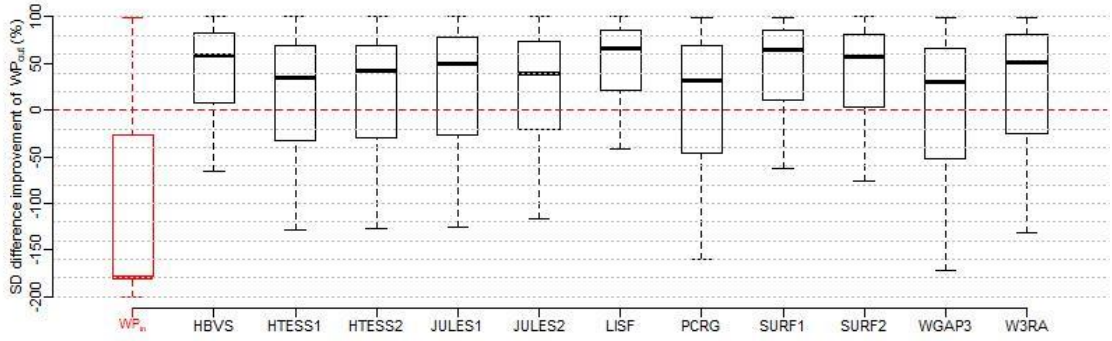


Figure 3: Spatial coverage of donor basins, receptor basins and non-donor and non-receptor basins.

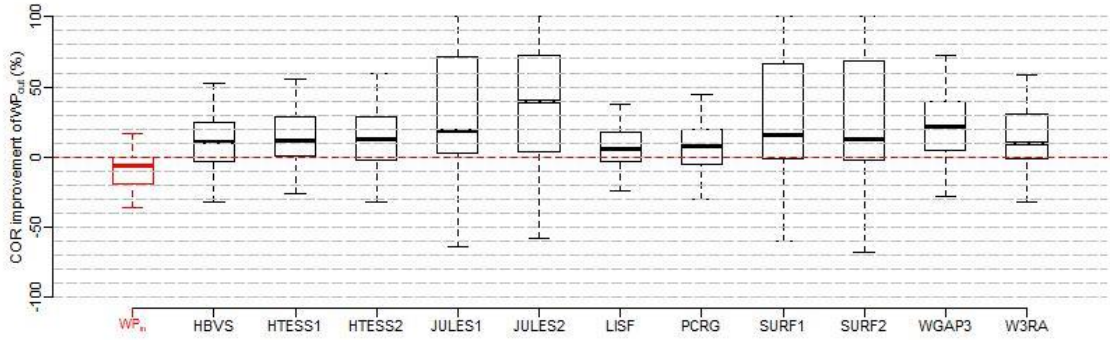
(a)


MSE improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
WP$_{in}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(b)


SD difference improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
-200
WP$_{in}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(c)


COR improvement ofWP$_{out}$ (%)
100
50
0
-50
-100
WP$_{in}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(d)

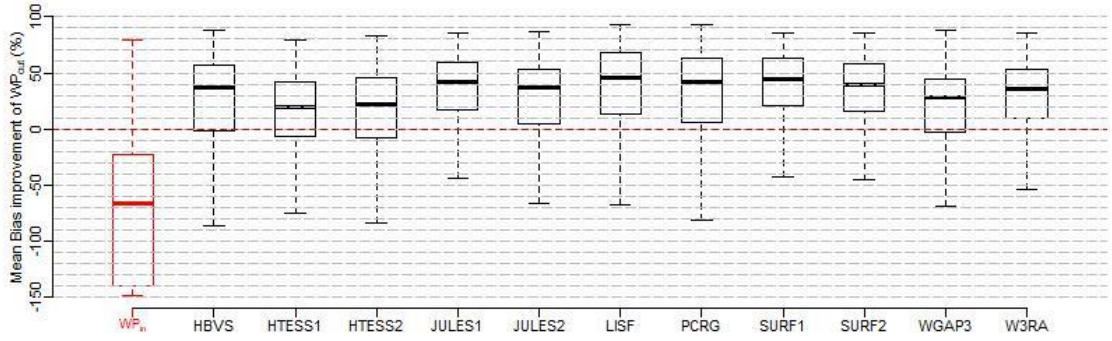

Mean Bias improvement of WP$_{out}$ (%)
100
50
0
-50
-100
-150
WP$_{in}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

Figure 4: Box and whisker plots displaying the percentage improvement that the weighted product (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 (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.

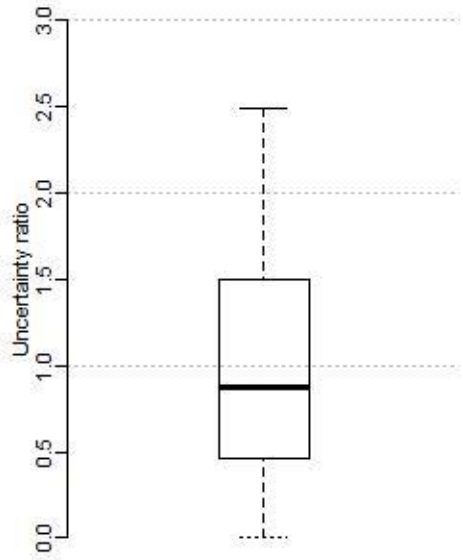


Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted computed in-sample to (2) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample

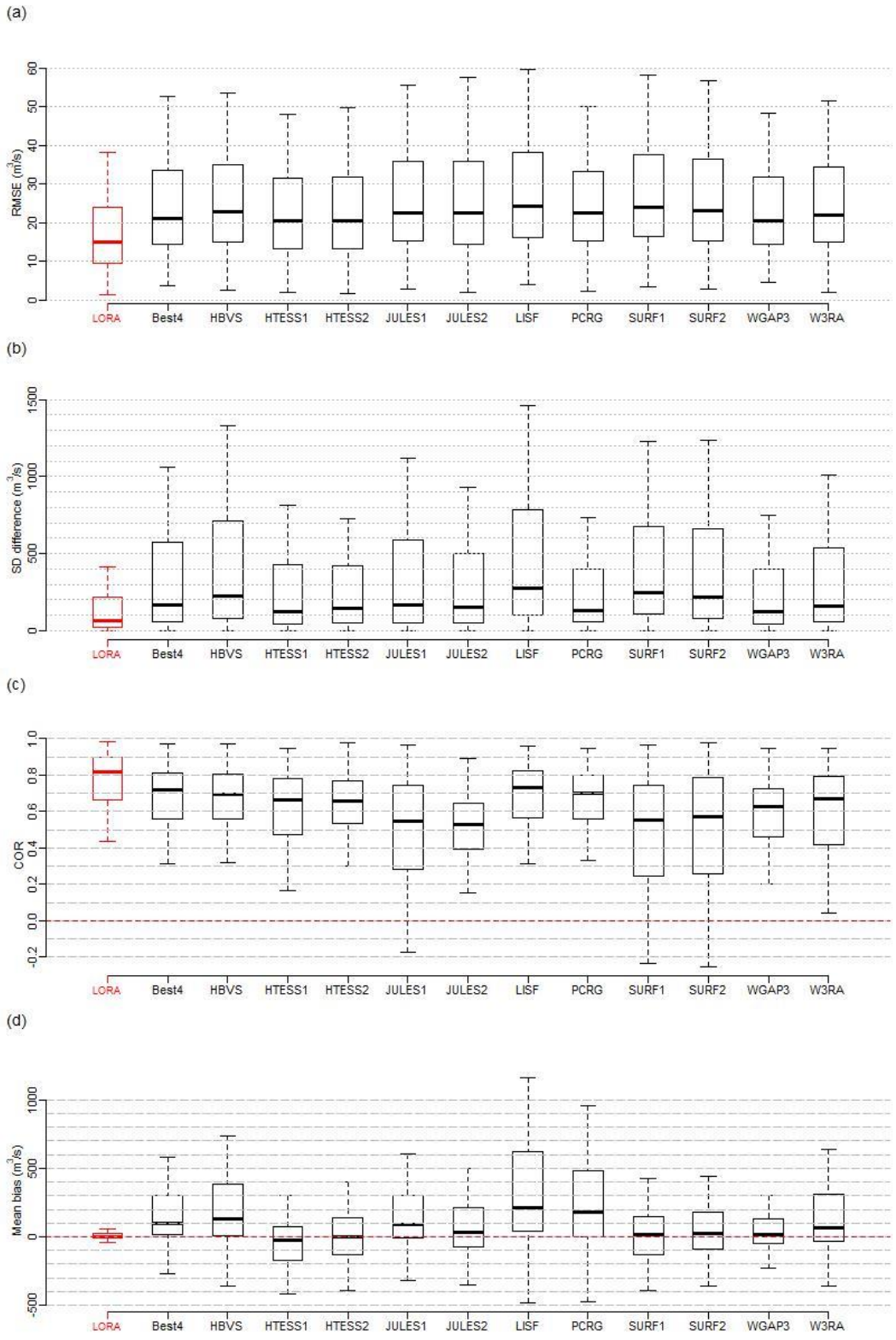


Figure 6: Four statistics, (a) RMSE, (b) SD difference, (c) COR and (d) Mean bias, calculated for LORA, Best4 (i.e. the simple average of runoff estimates from LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) and each runoff product involved in this study at the gauged basins. See Table 1 for dataset abbreviations.

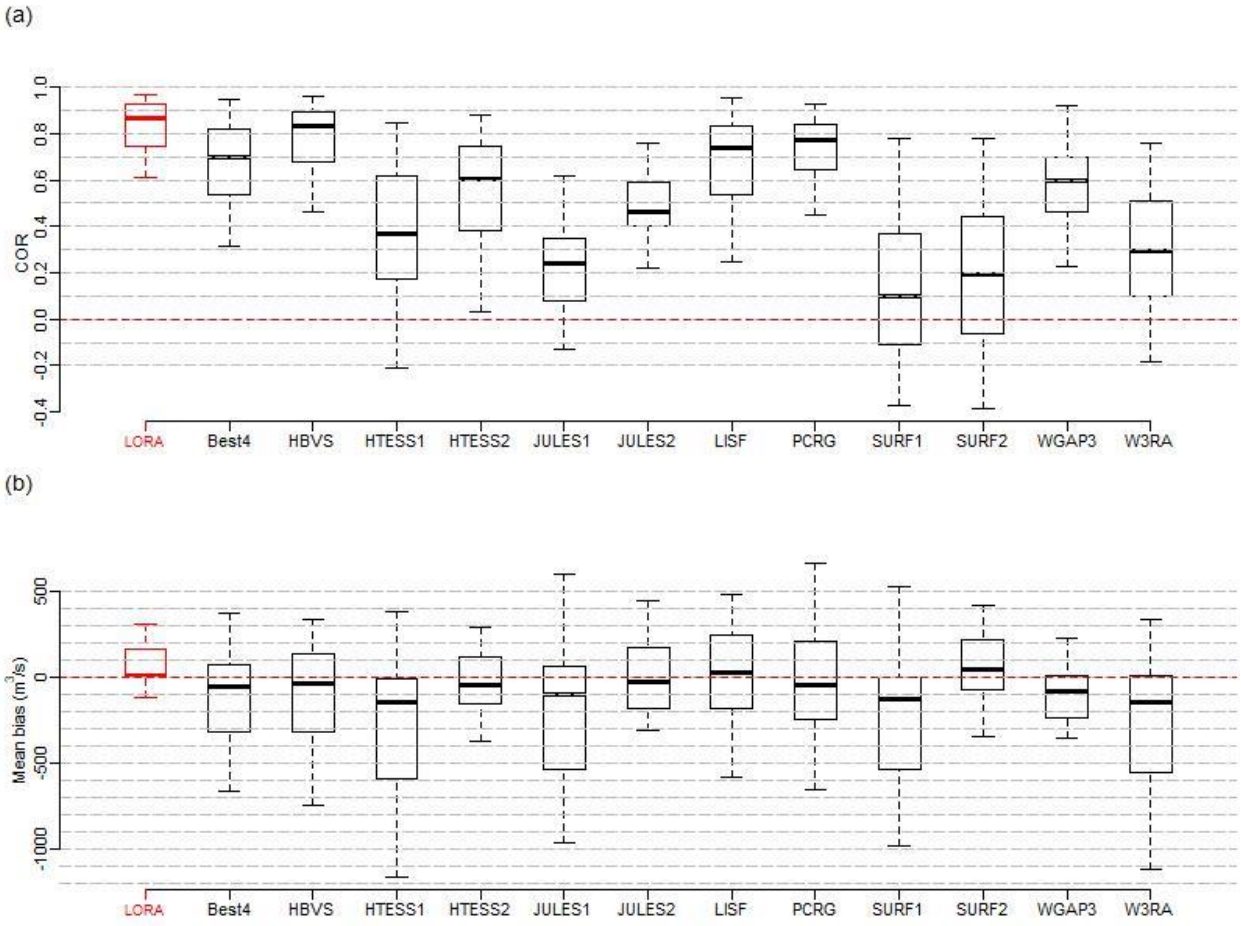


Figure 7: Two statistics, (a) COR and (b) Mean bias, calculated for LORA, Best4 (i.e. the simple average of runoff estimates from LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) and each runoff product involved in this study at the gauged basins located at the high latitudes (>60°). See Table 1 for dataset abbreviations.

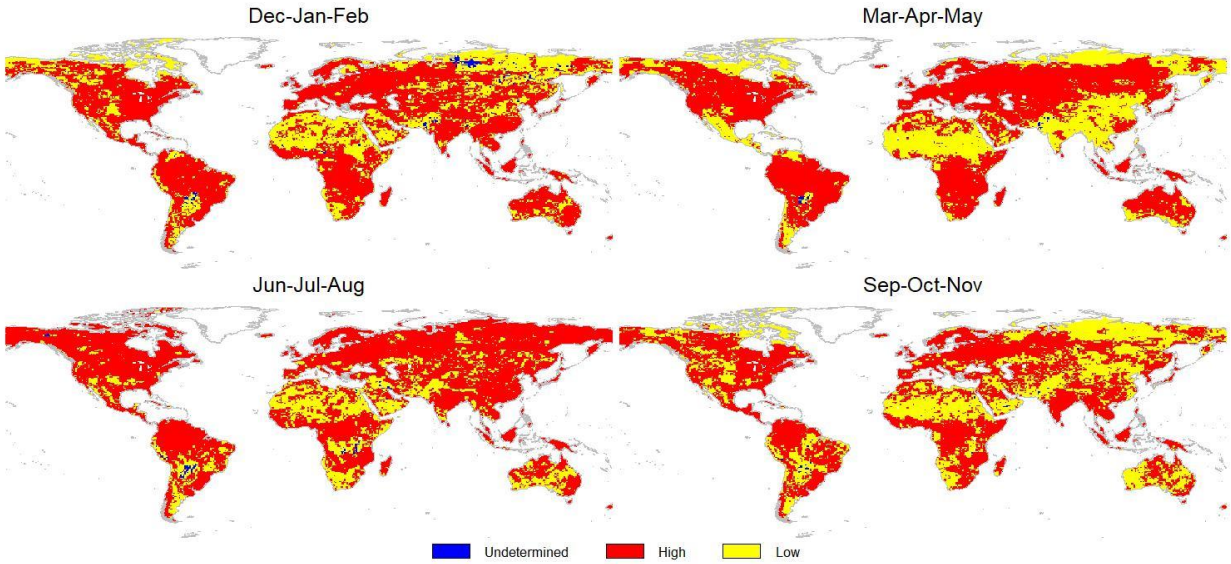


Figure 8: Seasonal reliability, defined as high ($\frac{men\ runoff\ uncertainty}{mean\ runoff} < 1$, in red), low ($\frac{men\ runoff\ uncertainty}{mean\ runoff} \geq 1$, in yellow) and undetermined (mean runoff = 0, in blue).

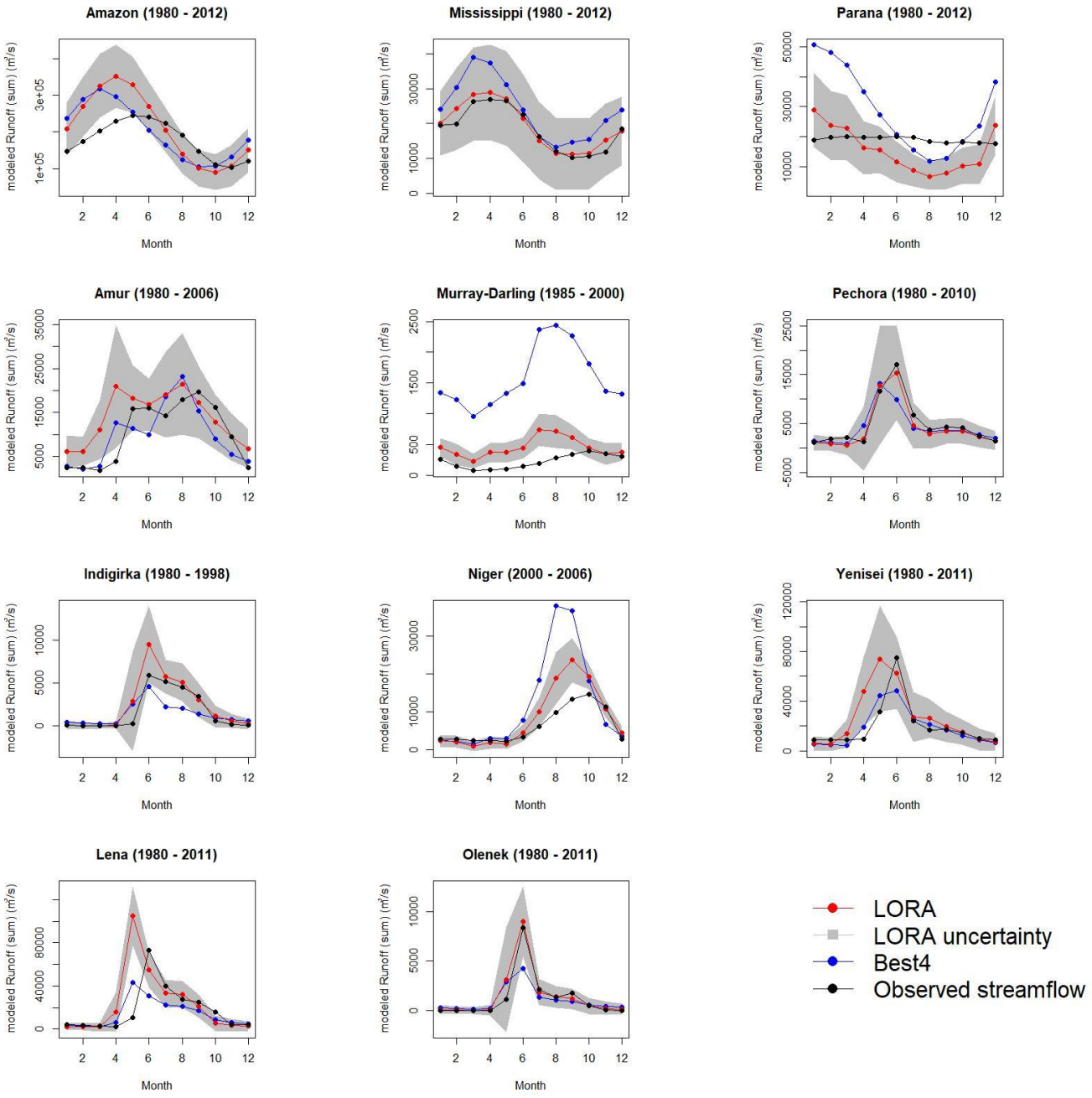


Figure 9: Seasonal cycle of Runoff aggregates from LORA and Best4 compared with the observed streamflow over 11 major basins. Runoff aggregates and the observed streamflow were averaged for each month across the period of availability of observation. The shaded regions show the aggregated uncertainty derived for LORA.

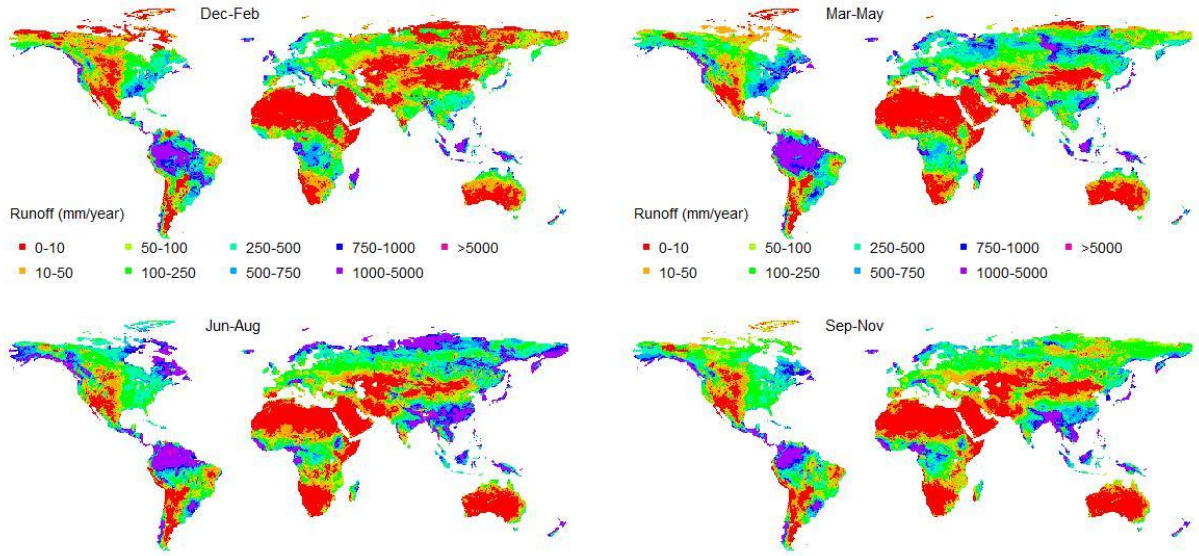


Figure 10: Mean seasonal runoff calculated for the period 1980 – 2012

Supplement of :

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

Table S1: Example of weights (w) and bias ratios (r) computed for the participating products over a range of river basins.

	HBVS		HTESS1		HTESS2		JULES1		JULES2		LISF		PCRG		SURF2		SURF1		WGAP3		W3RA	
	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r	w	r
Amur			-1.22	-0.78	0.14	-0.18					0.46	0.11	1.75	0.09			0.51	-1.35	-0.71	-0.10	0.08	-0.46
Copper	0.33	-0.32	-0.35	-0.42	0.59	0.02	-0.47	-0.35	-0.33	0.20	0.14	-0.18	0.89	-0.19	0.32	0.42	0.84	-0.01	0.07	0.03	-1.02	-0.20
Indigirka			-0.35	-0.80	0.03	0.02					1.23	-0.27	0.42	-0.10			0.79	-0.95	-0.02	-0.04	-1.10	-1.15
Mississippi			0.33	-0.14	0.02	-0.45					-0.09	0.39	0.31	0.28			-0.13	-0.06	0.25	0.02	0.30	0.24
Murray-Darling			1.01	0.75	0.08	0.74					-0.12	0.91	0.01	0.93			0.08	0.81	0.34	0.40	-0.41	0.84
Olenek			-0.35	-0.76	0.08	-0.12					2.07	-0.16	-0.95	-0.18			0.28	-0.70	0.10	-0.03	-0.22	-0.83
Parana			0.26	0.01	0.16	-0.19					-0.88	0.34	0.18	0.48			0.20	-0.12	1.29	0.06	-0.21	0.36
Pechora			-0.38	-0.34	0.33	-0.15					0.84	-0.13	0.82	-0.21			-0.01	-0.42	-0.47	-0.01	-0.14	-0.41
Yenisei			-0.71	-0.71	-0.10	-0.21					2.20	-0.09	-0.96	-0.09			0.52	-0.65	0.39	-0.01	-0.33	-0.72

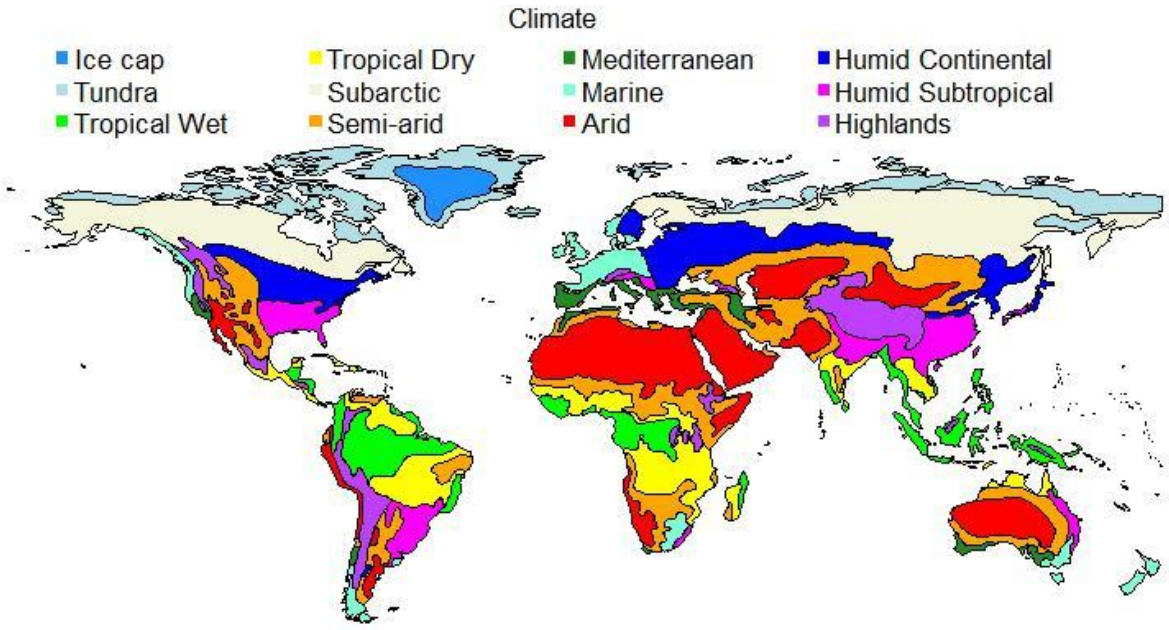


Figure S1: Climate map used in this study (available from ArcGIS online). It is a simplified climate zones map consisting of 12 broad climate classes.

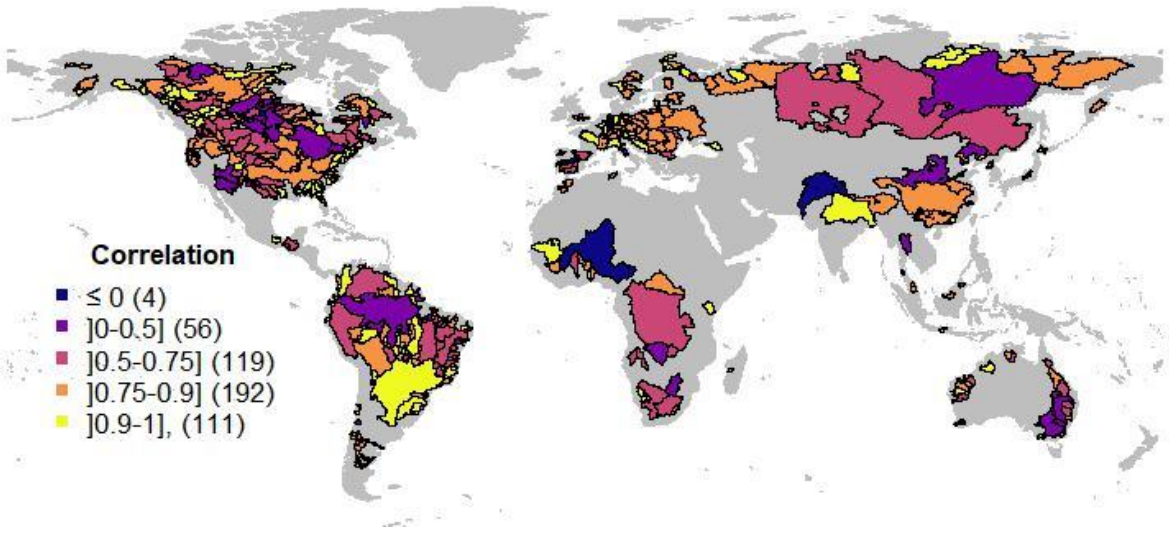


Figure S2: Temporal correlation of LORA with the observed streamflow over the gauged basins. Basins are colour coded by correlation range and their numbers are given in brackets.

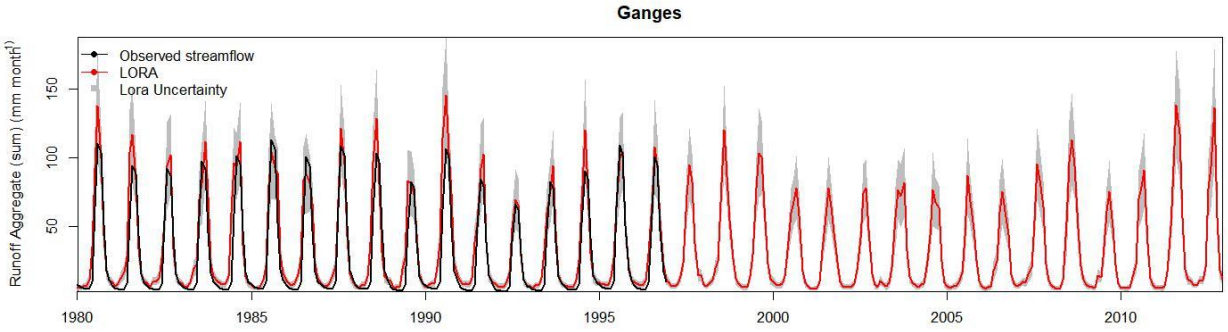


Figure S3: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Ganges basin (in mm month-1). This basin was shown in yellow in Fig. S2, indicating that LORA exhibits a high temporal correlation (≥ 0.9) with the observation.

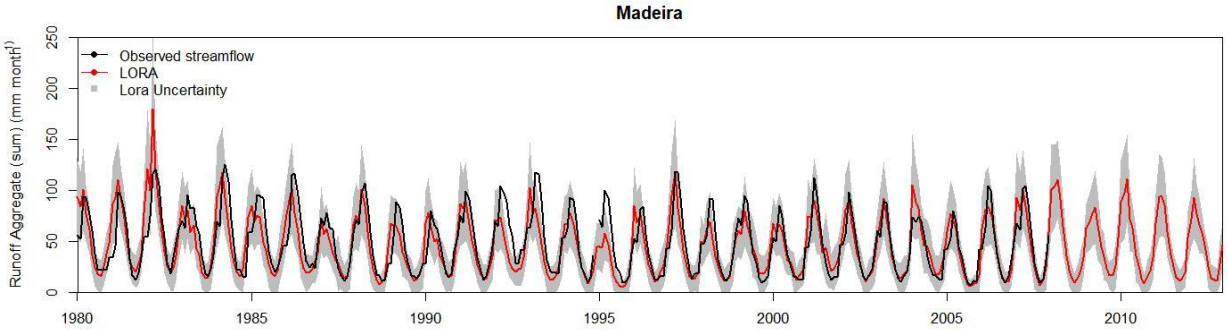


Figure S4: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Madeira basin, i.e. a sub-basin of the Amazon (in mm month-1). This basin was shown in orange in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.75 – 0.9[with the observation.

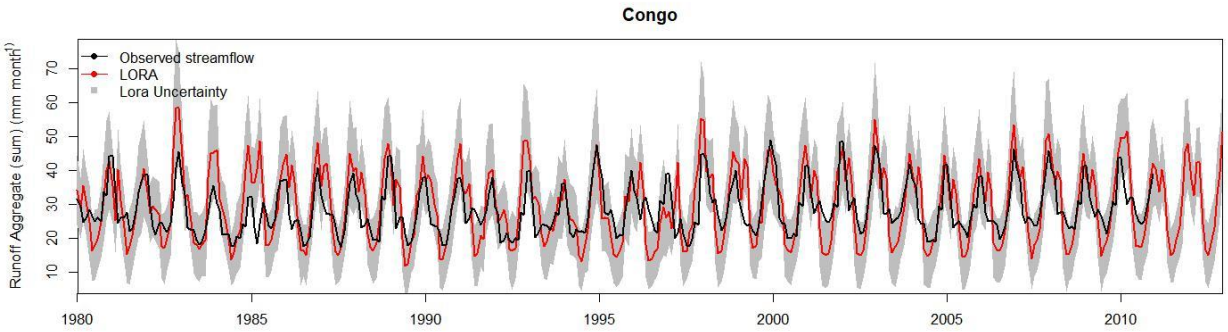


Figure S5: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Congo basin (in mm month-1). This basin was shown in violet in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.5 – 0.75[with the observation.

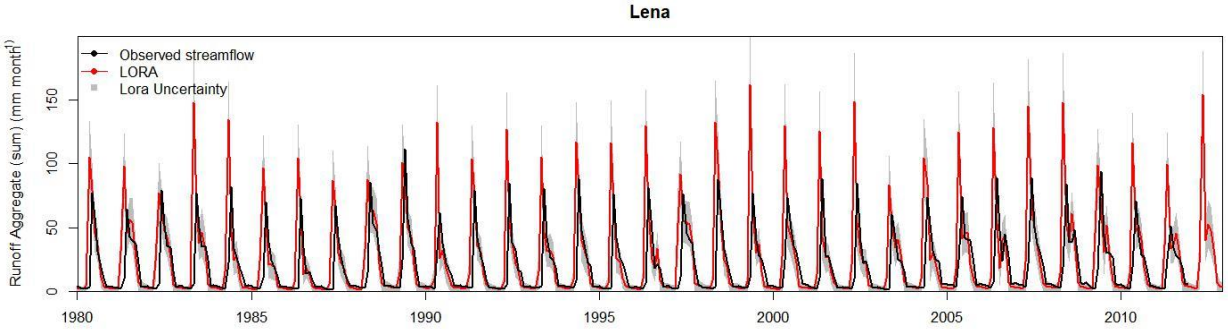


Figure S6: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Lena basin (in mm month-1). This basin was shown in purple in Fig. S2, indicating that LORA exhibits low temporal correlation (≤0.5) with the observation.

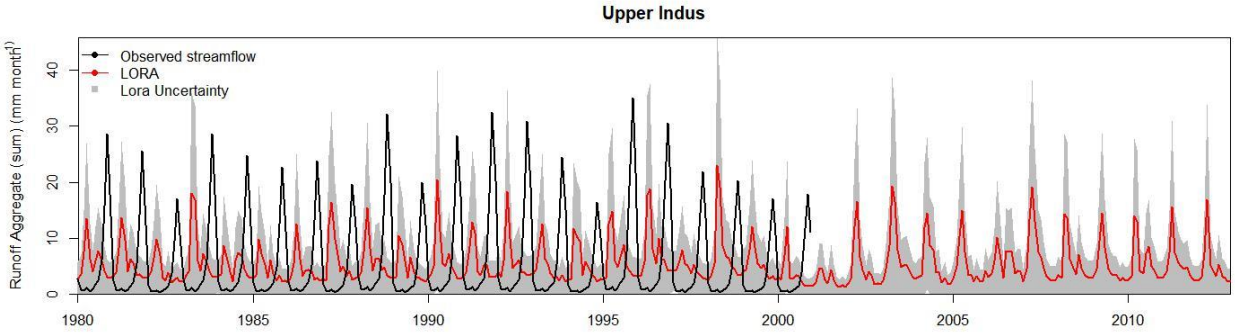


Figure S7: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the upper Indus basin (in mm month-1). This basin was shown in dark blue in Fig. S2, indicating that LORA exhibits a negative temporal correlation with the observation.

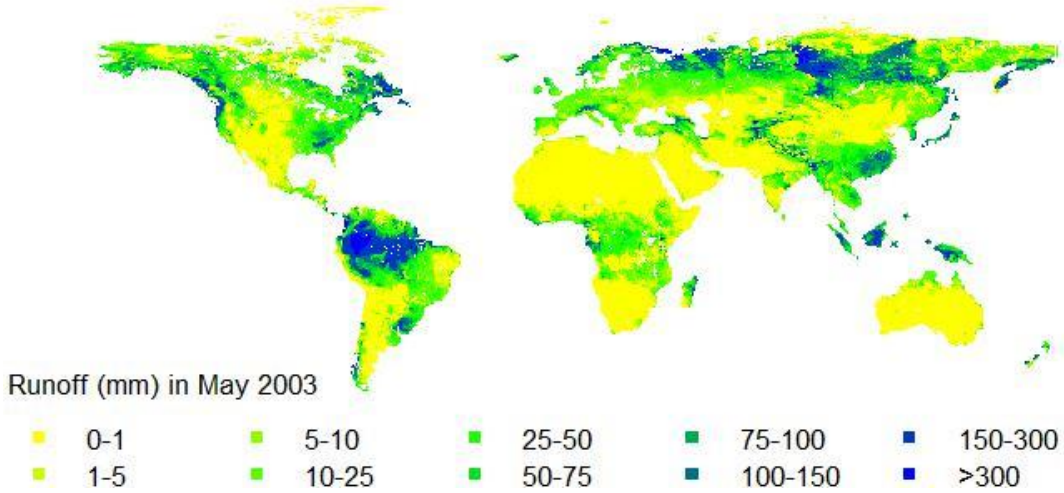


Fig. S8: Global map of LORA runoff fields (mm) in May 2003

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 0.5° 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.

1 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). Characterizing its dynamics and magnitudes is a major research aim of hydrology and hydrometeorology and a critical importance to improve our understanding of the current conditions of the large-scale water cycle and predict its future states.

More accurate estimates also provide additional constraint for climate model evaluation. ~~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; ~~(~~Falcone et al., 2010~~)~~ 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;Bierkens, 2015; 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 ~~(~~(Fekete et al., 2002~~); 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 but not limited to 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).

Field Code Changed

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.

2 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.

We detail in sections 2.3 and 2.4 our methods to derive the weighted runoff product for the global land. A flowchart summarizing the process is provided in Fig. 2~~, and Fig. 3 summarizes the steps carried out to derive the weighted runoff product for the global land.~~.

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: $\mu_q^j = \sum_{k=1}^{K} w_k(x_k^j - b_k)$ where $j \in [1, J]$ are the time steps and $k \in [1, K]$ represent the participating models, x_k^j (i.e., integrated runoff $Ragg_k^j$ over the basin areas in small basins and modelled streamflow at a gauge location q_k^j in large basins) is the value of the participating dataset in $m3 s-1$ at the j^{th} time step of the kth 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 $\sum_{j=1}^{J}(\mu_q^j - Q^j)^2$ subject to the constraint that $\sum_{k=1}^{K} w_k = 1$, where Q^j is the observed streamflow at the j^{th} time step. This minimization problem can be solved using the method of Lagrange multipliers by finding a minima for

$$F(w, \lambda) = \frac{1}{2}\left[\frac{1}{(J-1)}\sum_{j=1}^{J}(\mu_q^j - Q^j)^2\right] - \lambda\left(\left(\sum_{k=1}^{K} w_k\right) - 1\right).$$

Formatted ...
Formatted ...
Formatted ...
Formatted: Font: 10 pt, Not Italic, Font color: Auto, Complex Script Font: Not Italic
Formatted ...
Formatted ...
Formatted ...
Formatted: Font: +Headings CS (Times New Roman), Not Italic
Formatted: Centered

The solution to the minimization of $F(w,\lambda)$ can be expressed as $w = \frac{A^{-1}1}{1^T A^{-1} 1}$, where $1^T = \overbrace{[1,1,\ldots,1]}^{k\ \text{elements}}$ and A is the k × k error covariance matrix of the participating datasets (after bias correction), i.e. $A=\begin{pmatrix} c_{1,1} & \cdots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \ldots & c_{k,k} \end{pmatrix}$. A is symmetric and the term $c_{a,b}$ is the covariance of the a[th] and b[th] bias corrected dataset after subtracting the observed dataset, while each diagonal term $c_{k,k}$ is the error variance of dataset k. We note here that the solution presented here is based on the performance of the participating products (diagonal terms of A) and the dependence of their errors (accounted for by the non-diagonal terms of A). 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:

$\mu_r^j = \sum_{k=1}^{K} w_k (r_k^j - b'_k)$ ~~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: $\mu_q^j = \sum_{k=1}^{K} w_k (x_k^j - b_k)$ where $j \in [1,J]$ are the time steps and $k \in [1,K]$ represent the participating models, x_k^j (i.e., $Ragg_k^j$ in small basins and q_k^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 $\sum_{j=1}^{J} (\mu_q^j - Q^j)^2$, 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:~~

~~$\mu_r^j = \sum_{k=1}^{K} w_k (r_k^j - b'_k)$~~

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

To calculate the runoff bias b'_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 $Ragg_k^j$ instead of the modeled streamflow q_k^j Eq. (2).

$$\left[\frac{q_k^j - Q^j}{q_k^j} = \frac{b'_k}{r_k^j}\right]_{basin} \quad (1)$$

$$\left[\frac{Ragg_k^j - Q^j}{Ragg_k^j} = \frac{b'_k}{r_k^j}\right]_{basin} \quad (2)$$

We note that there no empirical evidence in the literature that the assumptions presented in Eq. 1 and Eq. 2 are valid. However, given that these assumptions constitute a part of our overall approach that we tested and proved its success later in this paper, the validity of these assumptions is very likely to hold true.

To avoid over-fitting when applying the weighting approach, we limited the number of participating models so that the ratio of number of records (i.e. total number of available monthly observations within the period of study) to number of models does not fall below ten~~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. Table S1 shows examples of weights and bias ratios calculated for the participating models over a range of river basins. It shows that HBVS, JULES1, JULES2 and SURF2 didn't participate in the weighting over the large basins (i.e. Amur, Indigirka, Mississippi, Murray-Darling, Olenek, Parana, Pechora and Yenisei) since these models don't have estimates for streamflow which are needed to construct the weights over large basins. For the smaller Copper River basin, however, runoff estimates from all models participated in deriving weighted runoff estimates. Table S1 also shows that in many cases, models were assigned negative weights. While this might not be expected in typical performance-based weighting, it is possible when weighting is based on error covariance as well as their performance differences in this formulation. We show below how the weights can be modified to non-negative weights.

~~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, R_{agg_μ}, then quantified s_r^2, the error variance of R_{agg_μ} with respect to the observed streamflow Q over time and space. We then transformed the constituent modelled estimates so that their variance about R_{agg_μ} at a given time step σ_r^{2j}, averaged over all time steps where we have available streamflow data for the current basin, is equal to 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 $\sqrt{\sigma_r^{2j}}$ 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).~~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. For any given gauged basin, we first calculated the spatial aggregate of our weighted runoff estimate Ragg_μ, then quantified ${s_q}^2$, the error variance of Ragg_μ with respect to the observed streamflow Q over time as:

$$ {s_q}^2 = \frac{\sum_{j=1}^{J} (\mathrm{Ragg}_\mu{}^j - Q^j)^2}{J-1} $$

Then, we wished to guarantee that the variance of the constituent modelled estimate $\sigma_q^{2^j}$ about $\text{Ragg}_\mu{}^j$ at a given time step, averaged over all time steps where we have available streamflow data, is equal to ${s_q}^2$, such as ${s_q}^2 = \frac{1}{J}\sum_{j=1}^{J}\sigma_q^{2^j}$.

Since the variance of the existing constituent products do not, in general, satisfy this equation. We transformed them so that it does. This involved first modifying the set of weights w to a new set $\tilde{w}$ such that

$\tilde{w} = \frac{w^T + (\alpha - 1)\frac{1^T}{K}}{\alpha}$, where $\alpha = 1 - K min(w_k)$ and $min(w_k)$ is the smallest negative weight (and α is set 1 if all w_k are non-negative). This ensures that all the modified weights $\tilde{w}_k$ are positive. We then transform the individual estimates x_k^j to $\tilde{x}_k^j$ where $\tilde{x}_k^j = \text{Ragg}_\mu{}^j + \beta(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - \text{Ragg}_\mu{}^j)$ and $\beta = \sqrt{\frac{s_q^2}{\frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{K}\tilde{w}_k(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - \text{Ragg}_\mu{}^j)^2}}$.

The weighted variance estimate of the transformed ensemble can be defined as $\sigma_q^{2^j} = \sum_{k=1}^{K}\tilde{w}_k(\tilde{x}_k^j - \text{Ragg}_\mu{}^j)^2$ and ensures that the equation $\frac{1}{J}\sum_{j=1}^{J}\sigma_q^{2^j} = s_q^2$ holds true. Furthermore, $\sqrt{\sigma_q^{2^j}}$ is the temporally varying estimate of uncertainty standard deviation of the transformed ensemble that (a) is varying in time, and (b) accurately reflects our ability to reproduce the observed streamflow.

We refer the reader to Bishop and Abramowitz (2013) for proofs.

In order to estimate $\sqrt{\sigma_r^{2^j}}$, the uncertainty of the runoff attributes μ_r^j at each point in time and space, we first transformed the runoff fields r_k^j to $\tilde{r}_k^j$ by applying the same transformation parameters α and β such that $\tilde{r}_k^j = \mu_r^j + \beta(\bar{r}^j + \alpha(r_k^j - \bar{r}^j) - \mu_r^j)$. We then calculated the error variance $\sigma_r^{2^j} = \sum_{k=1}^{K}\tilde{w}_k(\tilde{r}_k^j - \mu_r^j)^2$.

Finally, we used $\sqrt{\sigma_r^{2^j}}$ as the spatially and temporally varying estimate of runoff uncertainty standard deviation, which we will refer to below simply as 'uncertainty'. It provides a much more defensible uncertainty estimate than simply calculating the standard deviation of the involved products.

We note that for a given basin, $\sqrt{\sigma_q^{2^j}}$ represents the uncertainty of the modelled streamflow i.e. $\text{Ragg}_\mu{}^j$, while $\sqrt{\sigma_r^{2^j}}$ represents the uncertainty of modelled runoff at each grid cell across the basin. This means that at every time step, there is one value for $\sqrt{\sigma_q^{2^j}}$ per basin, and one value for $\sqrt{\sigma_r^2}$ per grid across the basin.

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. This resulted in one set of weights and bias ratios obtained jointly from the three donor basins-. 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:

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

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. $Z_{p,a}$ and $Z_{p,b}$ are the values of the characteristic p at donor and receptor basins, respectively. 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 covered with ~~permanent~~ ice during most of the year (defined by climate zones Tundra, Subarctic and Ice cap using a simplified climate zones map (Fig. S1) 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~~ three most similar donors for every receptor basin. The dissimilarity technique has been previously applied to find ten donors for one receptor. Given that all the selected donors must have very close similarity indices, we found by trial and error that increasing the number of donor basins might introduce donor basins that have a significantly different similarity index, and that setting the number of donor basins to three seemed most appropriate.

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 climate zones map; Esri Education Team, 2014) defined above. We used this particular climate

Formatted: Font: (Default) +Headings CS (Times New Roman), Not Bold, Font color: Auto, Complex Script Font: +Headings CS (Times New Roman), 10 pt, Italic

Formatted: Font: (Default) +Headings CS (Times New Roman), Not Bold, Font color: Auto, Complex Script Font: +Headings CS (Times New Roman), 10 pt, Italic, English (United States)

Formatted: Font: Italic, Complex Script Font: Not Italic

Formatted: Indent: Before: 0 cm, Line spacing: 1.5 lines

Formatted: Subscript

Formatted: Font: Italic, Complex Script Font: Not Italic

Formatted: Font: Italic, Complex Script Font: Not Italic

Formatted: Font: Italic, Complex Script Font: Not Italic

Formatted: Font: Italic, Complex Script Font: Not Italic

map because it comprises only 12 broad climate groups (compared to more than 30 in other climate maps e.g. Köppen–Geiger). This reduced the divisions made to large heterogenous basins, while ensuring that the resultant basin zones of individual basins have very distinct climate characteristics. Figure 3 shows the spatial coverage of the donor basins, receptor basins and non-donor and non-receptor basins. ~~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 (WP_{in}) derived by using observed streamflow for this basin to weight models; (c) an out-of-sample weighted product (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 WP_{in}, 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= $\sigma_{\text{Ragg}} - \sigma_{\text{observed streamflow}}$. 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.

The uncertainty estimates computed at the gauged basins represent well the deviation of the spatial aggregate of our weighted product ($Ragg_\mu$) from the observed streamflow, since the in-sample uncertainty estimates are calculated from the variance of the transformed ensemble, which by design equals MSE of $Ragg_\mu$ against the observation (i.e. error variance of $Ragg_\mu$). To test if the uncertainty estimates perform well out-of-sample (i.e. at the ungauged basins), we performed another out of sample test. In this test, we took a gauged basin, but instead of constraining the weighting using observed streamflow from this basin,

we constructed model weights by using the three most similar donor basins. We could then calculate MSE of $Ragg_{\mu}$ against observation from the three donor basins, and we denoted this by MSE_{in}, which represents the uncertainty estimates calculated in-sample, since the observational data used in this case is the same dataset that was used to train the weighting. We also calculated the MSE of the aggregated weighted product against the actual observation of the gauged basin and we denoted this by MSE_{out}. MSE_{out} represents the uncertainty estimates computed out-of-sample, since the comparison was performed against observational data that has not been used to train the weighting. We repeated the out-of-sample test for all the gauged basins. We displayed the results of the out-sample-test by showing the ratios of MSE_{in} to MSE_{out}. If the approach is succeeding, we expect that this ratio is around one, indicating that the values of MSE_{in} and MSE_{out} are close to each other. We used a box and whisker plots to show the results.

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. 4~~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. Furthermore, the boxplot in Fig. 5 shows that, overall, when the uncertainty estimates are computed out-of-sample they are very similar to what they would have been if they were computed in-sample. This demonstrates that the dissimilarity technique can be effectively used to derive not only the weighting product but also its associated uncertainties at the ungauged basin.

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. 6~~5~~ (a-d) display the results.

The RMSE plot in Fig. 6 ~~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. 6~~5~~ (b)) across more than half of the basins. The Mean bias plot in Fig. 6 ~~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. 6~~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~~6~~ (c-d) but this time for the snow-dominated basins only. We display the results in Fig. 7~~6~~.

The temporal correlation plot in Fig. 7~~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. 6~~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 7~~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. 6~~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. 6~~5~~ (a and b). The results displayed in Fig. 6~~5~~ and Fig. 7~~6~~ are discussed further below.

We calculated the seasonal relative uncertainty expressed as the ratio of the seasonal average uncertainty to seasonal mean runoff (i.e. $\frac{\text{mean runoff uncertainty}}{\text{mean runoff}}$) over the period 1980 – 2012.~~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. 8~~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~~8~~ 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 9~~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, w~~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 as well as adjusted VIC estimates after enforcing the physical constraints of the water budget 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). We also showed the average annual volume of water that discharges from these basins computed from LORA and the observational data.

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. Other global estimates of total runoff are also available such as GLDAS and Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger et al., 2016), however we haven't compared LORA with these datasets because they either have a short common period with LORA, or a coarser resolution (i.e. 1º) and showed a significant disagreement with observation when interpolated to a 0.5° grid.

Finally, we provide in Fig. S8 an example of runoff fields in an individual month (e.g. May, 2003).

Formatted: Font: (Default) +Headings CS (Times New Roman), 10 pt, Font color: Auto, Complex Script Font: +Headings CS (Times New Roman), 10 pt, English (United States)

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. 8 that the runoff values computed over dry climates tend to be less reliable than those in other regimes. This is perhaps due to biases in the WFDEI precipitation forcing that are propagated and intensified in the simulated runoff (Beck et al., 2017a). Another possible reason is the reduced proficiency of models in representing runoff dynamics in arid climates where runoff tends to be highly non-linearly related to rainfall and often evaporates locally without reaching a river system (Ye et al., 1997). ~~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 .~~ 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. ~~4~~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. ~~6~~5 (c) and Fig. ~~7~~6 (a), and also explains the highest weights attributed to LISFLOOD in the majority of snow-dominated basins (Table S1). Because of this, and because LISFLOOD tends to overestimate runoff across half of the snow-dominated basins (as shown in Fig.~~7~~6 (b)) LORA exhibits a positive bias across half of the snow-dominated basins (Fig. ~~7~~6 (b)) and particularly in Lena, Amur and Yenisei basins (Fig. ~~9~~8)~~-~~.

Further, we provide in Fig. S2 the spatial distribution of correlation results from Fig. 6 (c). The basins are colour-coded by their temporal correlation with the observed streamflow and the number of basins in each category is given. Basins in yellow are those where LORA is highly correlated with the observation while dark blue basins are those where LORA exhibits a negative correlation with the observation. It can be noted from Fig. 6(c) that occurrence of negative correlation is extremely unusual which explains why these were considered outliers and were not shown in the box and whisker plot. Likely, low correlation basins are unusual and constitute less than 12% of the number of basins (excluding basins with negative correlation). Also, the median value is above 0.8, which is higher than any constituent estimates. We selected a basin from each correlation range and examined the timeseries of LORA and the observed streamflow more closely (Fig. S3-S7), in particular illustrating the uncertainty estimate of LORA. In Ganges, LORA captures well the observed time-series dynamic with a tendency to over-estimate streamflow peak in August (Fig. S3). Over Madeira basin, LORA is able to represent reasonably well most of the climatic variability found in the observation (Fig. S4). In Congo, the catchment has an irregular time-series dynamic, LORA is in principle able to capture a large part of the climatic variability in the observation (Fig. S5). In Lena, the observation shows a peak in June and a second less significant peak in September (Fig. S6). Both peaks are

captured by LORA during most of the time series with a tendency to underestimate the late summer peak and overestimate the early summer peak. In the upper Indus, LORA does not capture the magnitudes of observed streamflow and shows a reversed seasonal cycle which explains why it exhibits negative correlation with the observation (Fig.S7). Zhang et al. (2018) found disagreement between simulated runoff from three LSMs and observed streamflow over Indus basin which they expected to be due to errors in the observational data from GRDB dataset.

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. 87). 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 109 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 Amazonn 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. 9~~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. Also, given that there are no direct observations for runoff, uncertainties were computed from the discrepancy between the modelled runoff aggregates and observed streamflow. This ignored the lag time between LORA integrated runoff and observed streamflow at the mouth of the river and induced biases that possibly led to overestimated uncertainty over large gauged 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 Data availability

LORA v1.0 can be downloaded from geonetwork.nci.org.au and its DOI is http://dx.doi.org/10.25914/5b612e993d8ea

87 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.

98 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, Department of Physical Geography, Utrecht University, Utrecht, The Netherlands. [online] Available from: http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf, 2008.
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.
Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resour. Res., 51, 4923–4947, doi:10.1002/2015WR017173, 2015.
Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41, 885–900, doi:10.1007/s00382-012-1610-y, 2013.
Burek, P., van der Knijff, J., de Roo, A.: LISFLOOD, distributed water balance and flood simulation model revised user manual, Joint Research Centre of the European Commission., 2013.
Dai, A.: Historical and Future Changes in Streamflow and Continental Runoff: A Review, Terr. Water Cycle Clim. Chang. Nat. Human-Induced Impacts, Geophys. Monogr., 221, 17–37, doi:10.1002/9781118971772.ch2, 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, r a Heal. Ctry. Natl. Res. Flagsh., 3(5), 100 pp, 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.
Fekete, B. M., Vörösmarty, C. J. and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cycles, 16(3), 15-1-15–10, doi:10.1029/1999GB001254, 2002.
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, Hydrol. Earth Syst. Sci, 22(2), 1317–1336, doi:https://doi.org/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, Hydrol. 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. Atmos., 99(D7), 14415--14428, 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., 48(1), 1–12, doi:10.2166/nh.2016.166, 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., 17, 3707–3720, doi:10.5194/hess-17-3707-2013, 2013.
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, Report to the National Land and Water Resources Audi., 2000.
Rantz, S. E.: Measurement and computation of stream flow. Volume 2: Computation of discharge, US Geol. Surv. water-supply Pap., 2(2175), 631 pp, 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, 9(2), 389–413, doi:https://doi.org/10.5194/essd-9-389-2017, 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), 1–5, 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.
Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M. and Jakeman, A. J.: Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resour. Res., 33(1), 153–166, 1997.
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.
Huntzinger, D.N., C.R. Schwalm, Y. Wei, R.B. Cook, A.M. Michalak, K. Schaefer, A.R. Jacobson, M.A. Arain, P. Ciais, J.B. Fisher, D.J. Hayes, M. Huang, S. Huang, A. Ito, A.K. Jain, H. Lei, C. Lu, F. Maignan, J. Mao, N. Parazoo, C. Peng, S. Peng, B. Poulter, D.M. Ricciuto, H. Tian, Xiaoying Shi, W. Wang, N. Zeng, F. Zhao, Q. Zhu, J. Yang, and B. Tao. 2016. NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (version 1) in Standard Format. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1225.
~~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, 2008)
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 yearly volume of LORA runoff aggregates (i.e. flow in Km^3) and observed annual flow (Km^3) over the basins and 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. Observed annual flow is given only if data from all contributing stations is available over a whole year over for at least 17 years out of 33 years covered in this study.~~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. Observed annual flow is given only if data from all contributing stations is available over a whole year over for at least 17 years out of 33 years covered in this study.~~

Basin	VIC mm/year	VIC adjusted for water budget closure mm/year	LORA (Runoff) mm/year	LORA (uncertainty) mm/year	LORA yearly flow ± uncertainty Km^3	Observed yearly flow Km^3	Dominant climate
Amazon	1048	1029	1151	3605~~7~~	6763 ± 2115~~7148~~	-	Tropical wet
Amur	135	129	219	115	428 ± 225~~428~~	325	Humid continental and semi-arid
Columbia	318	293	333	101	218 ± 66~~218~~	209	Semi-arid and highlands
Congo	407	404	358	147	1292 ± 532~~1292~~	1240	Tropical wet and tropical dry

Danube	272	265	260	125	199 ± 95~~199~~	205	Marine Humid, continental and humid subtropical
Indigirka	132	120	228	171	78 ± 59~~78~~	53	Subarctic
Lena	142	134	301	137	731 ± 332~~731~~	557	Subarctic
Mackenzie	189	173	191	110	323 ± 186~~323~~	294	Subarctic
Mississippi	220	215	212	123	616 ± 359~~615~~	581	Humid continental and humid subtropical
Murray-Darling	42	41	15	6	12 ± 5~~12~~	-	Arid and semi-arid
Niger	198	194	106	41	239 ± 87~~239~~	170	Arid, semi-arid and tropical dry
Olenek	114	106	230	208	48 ± 43~~48~~	40	Subarctic
Parana	278	279	189	97	471 ± 247~~471~~	600	Marine and humid subtropical
Pechora	342	308	420	420	131 ± 131~~131~~	153	Tundra and subarctic
Yenisei	217	195	324	203	828 ± 520~~828~~	612	Subarctic
Yukon	149	139	229	102	188 ± 83~~188~~	214	Subarctic

~~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~~

Formatted: Centered, Line spacing: single
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: English (United States)
Formatted: English (United States), Pattern: Clear
Formatted: English (United States)
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Centered
Formatted: Left

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

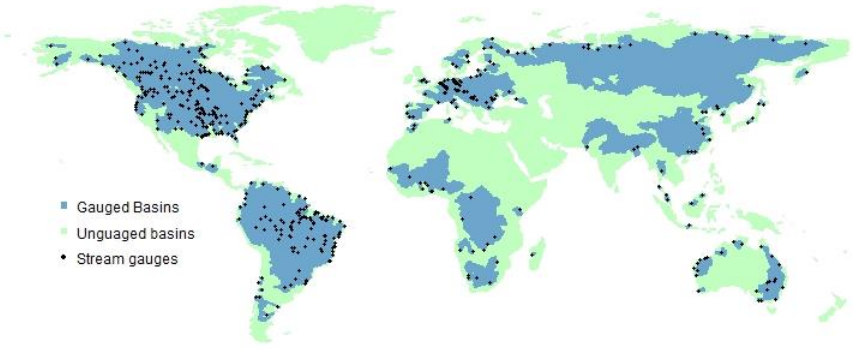
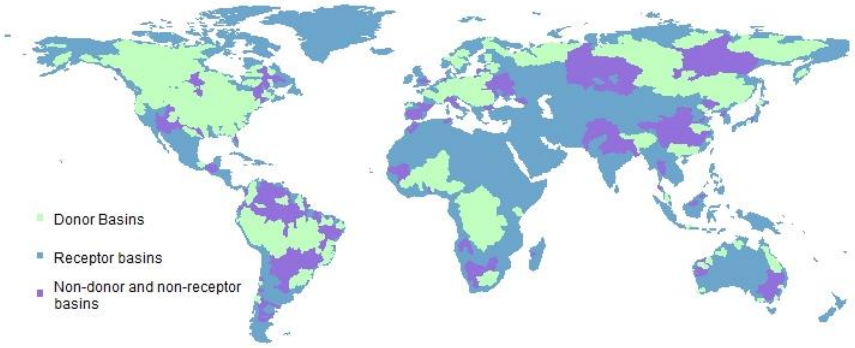


Figure 1: Spatial coverage of gauged and ungauged river basins and location of stream gauges.

Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left
Formatted: Left
Formatted: Centered
Formatted: Left



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

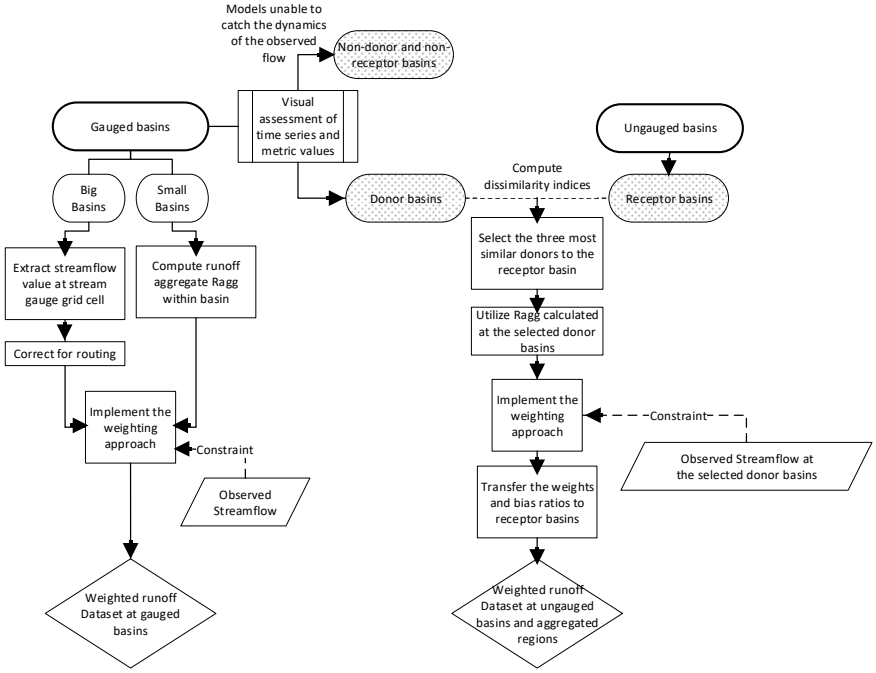


Figure 23. Flow-chart summarizing the steps carried out to derive the weighted runoff product for the global land surface.

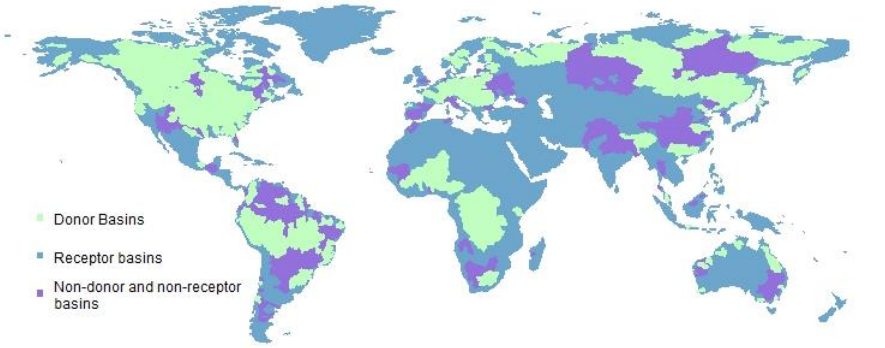
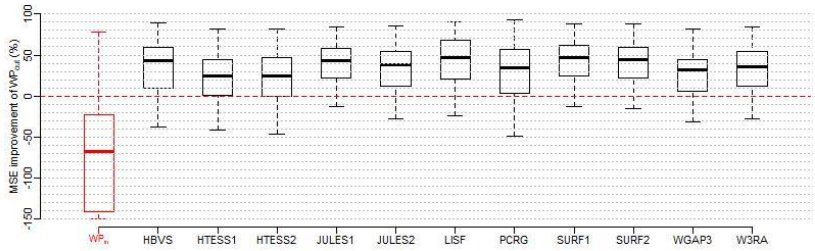
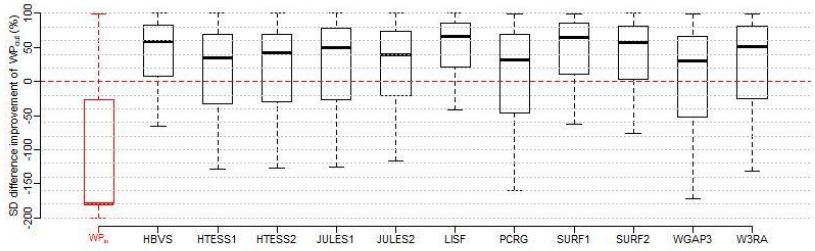


Figure 3: Spatial coverage of donor basins, receptor basins and non-donor and non-receptor basins.

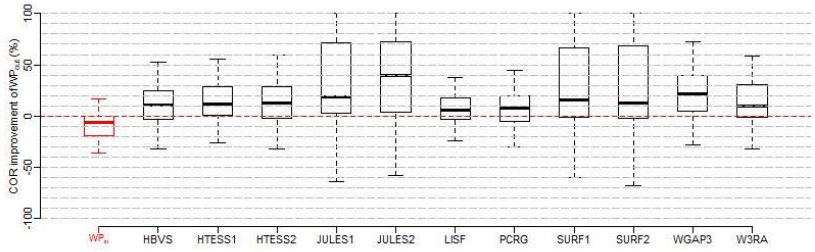
(a)


MSE improvement of VWP$_{ERA}$ (%)
100
50
0
-50
-100
-150
WP$_{ERA}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(b)


SD difference improvement of VWP$_{ERA}$ (%)
100
50
0
-50
-100
-150
-200
WP$_{ERA}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(c)


COR improvement of VWP$_{ERA}$ (%)
100
50
0
-50
-100
WP$_{ERA}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

(d)

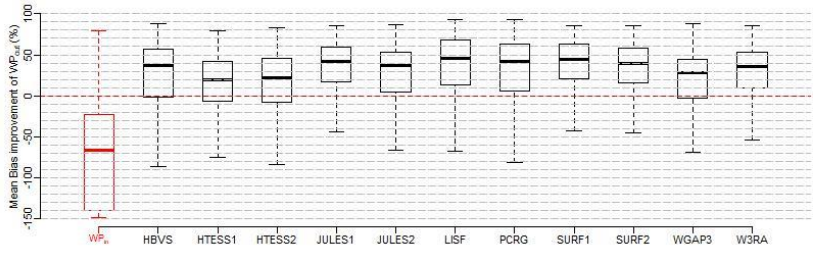

Mean Bias improvement of VWP$_{ERA}$ (%)
100
50
0
-50
-100
-150
WP$_{ERA}$
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

Figure 4: Box and whisker plots displaying the percentage improvement that the weighted product (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 (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.

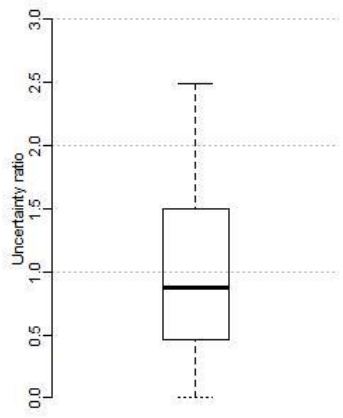
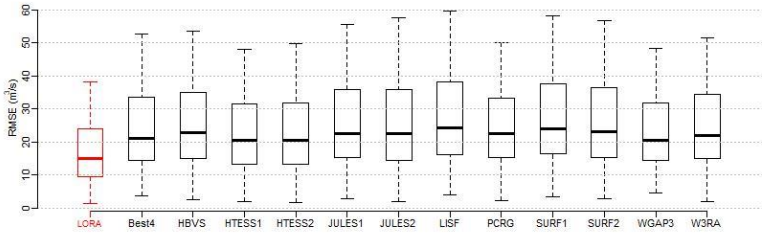


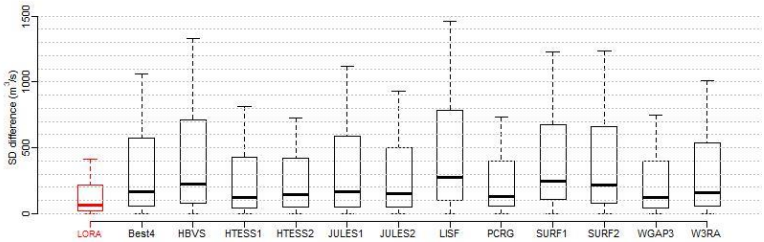
Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted computed in-sample to (2) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample

Formatted: English (United Kingdom)

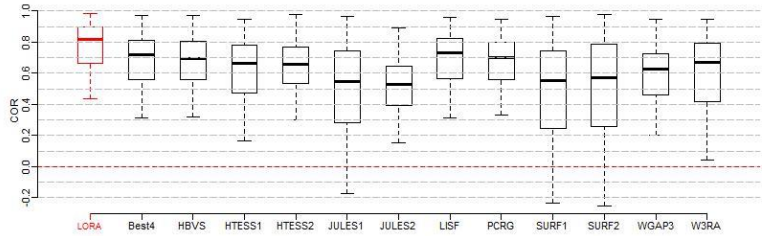
(a)


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)

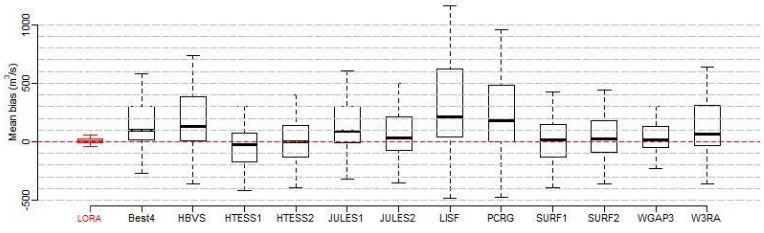

Mean bias (m³/s)
-500
0
500
1000
LORA
Best4
HBVS
HTESS1
HTESS2
JULES1
JULES2
LISF
PCRG
SURF1
SURF2
WGAP3
W3RA

Figure 65: 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.

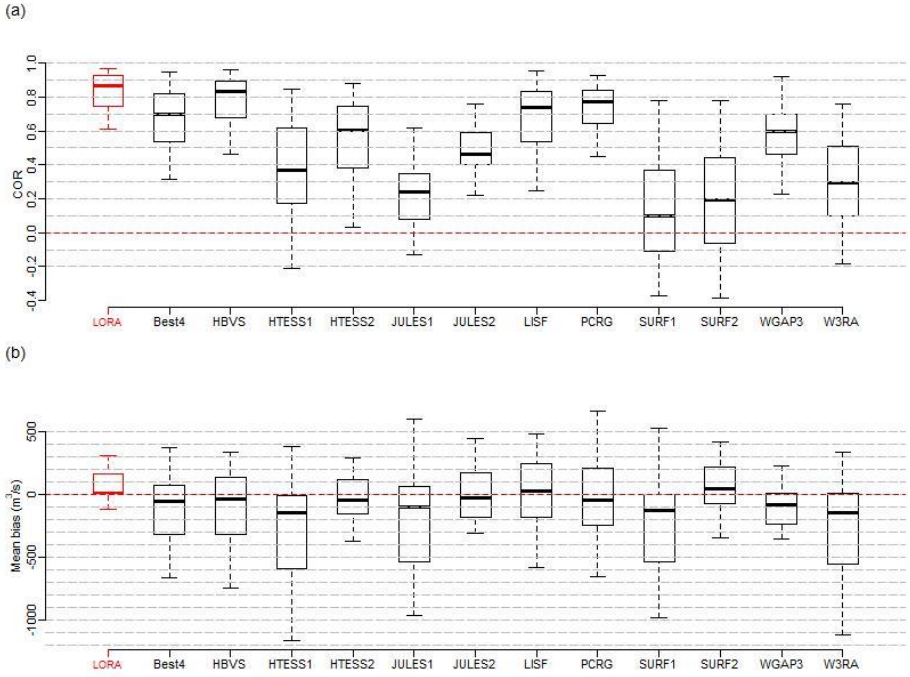


Figure 76: Two statistics, (a) COR and (b) Mean bias, calculated for LORA, Best4 (i.e. the simple average of runoff estimates from LISFLOOD, WaterGAP3, W3RA and HBV-SIMREG) and each runoff product involved in this study at the gauged basins located at the high latitudes (>60°). See Table 1 for dataset abbreviations.

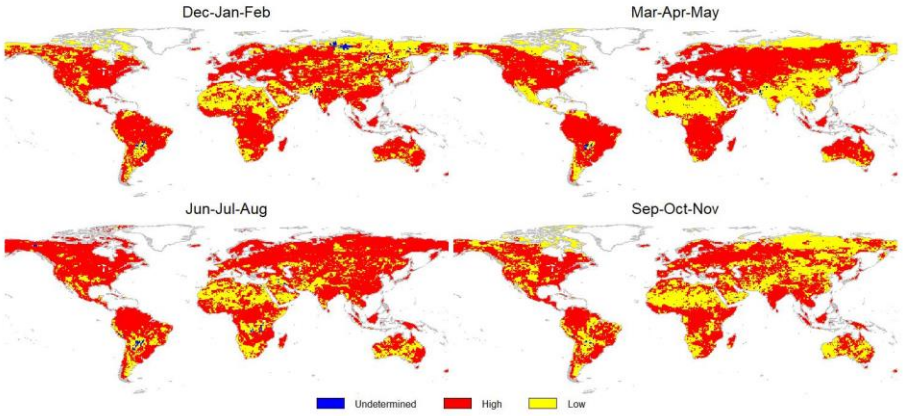
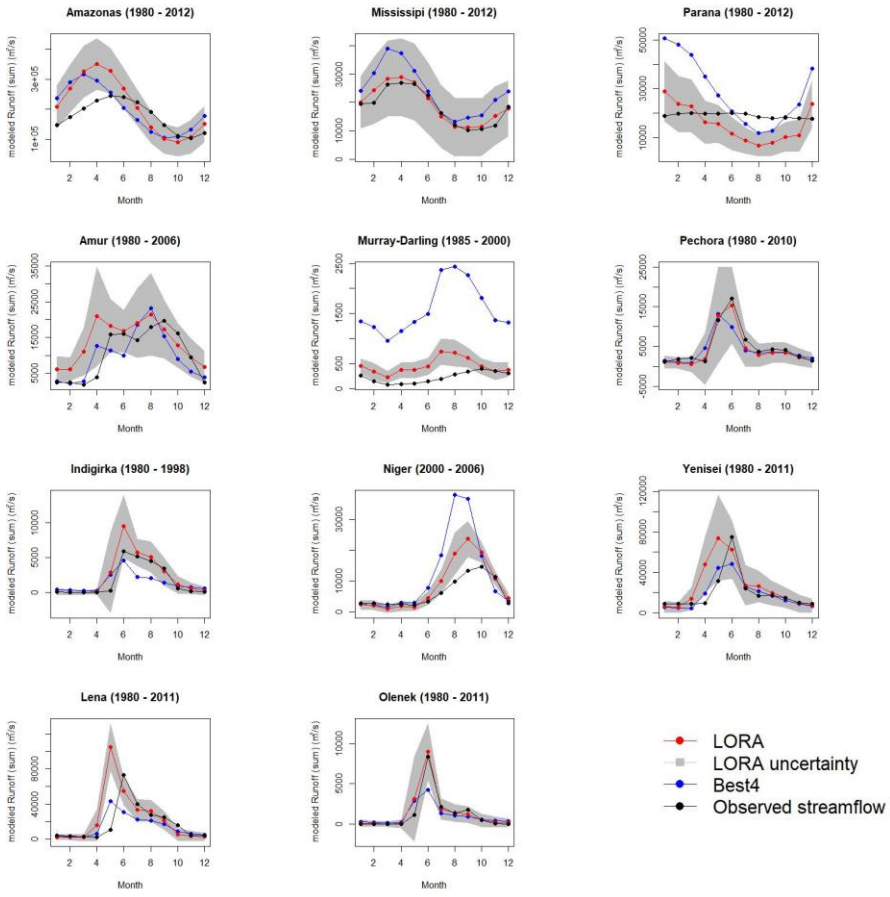


Figure 87: Seasonal reliability, defined as high ($\frac{men\ runoff\ uncertainty}{mean\ runoff} < 1$, in red), low ($\frac{men\ runoff\ uncertainty}{mean\ runoff} \geq 1$, in yellow) and undetermined (mean runoff = 0, in blue).


Amazonas (1980 - 2012)
Mississipi (1980 - 2012)
Parana (1980 - 2012)
Amur (1980 - 2006)
Murray-Darling (1985 - 2000)
Pechora (1980 - 2010)
Indigirka (1980 - 1998)
Niger (2000 - 2006)
Yenisei (1980 - 2011)
Lena (1980 - 2011)
Olenek (1980 - 2011)
modeled Runoff (sum) (m³/s)
Month
LORA
LORA uncertainty
Best4
Observed streamflow

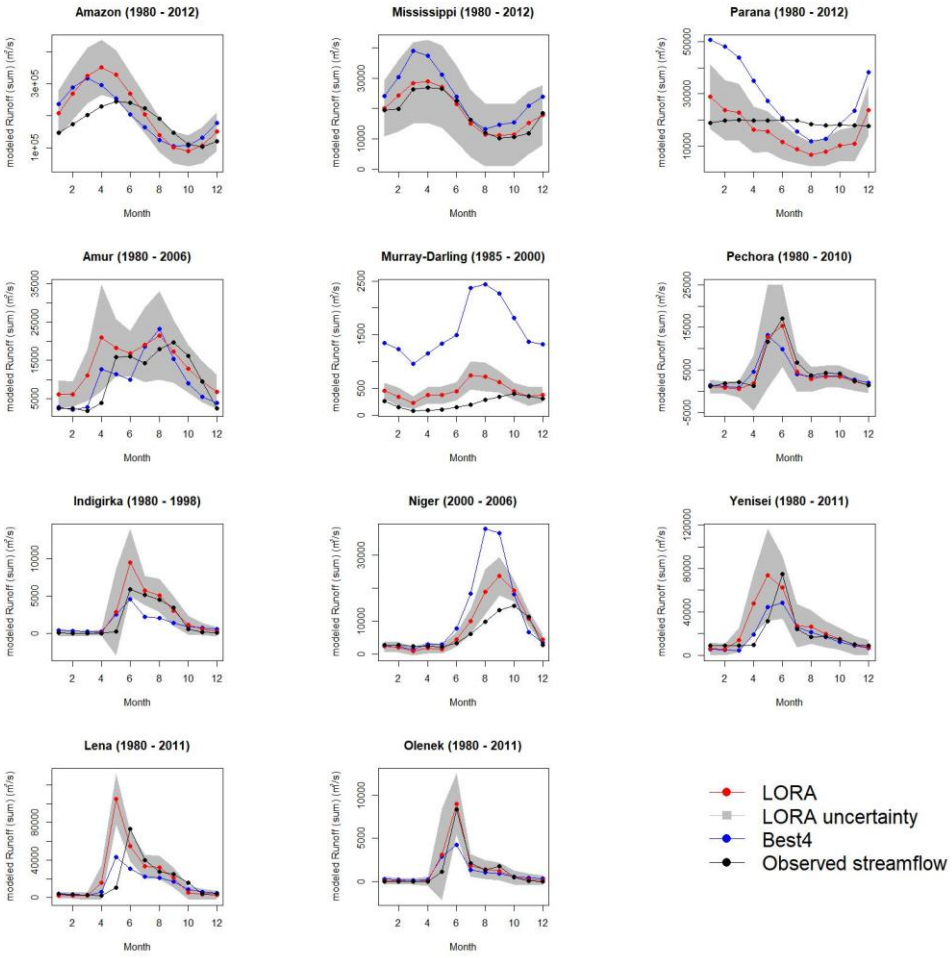


Figure 98: Seasonal cycle of Runoff aggregates from LORA and Best4 compared with the observed streamflow over 11 major basins. Runoff aggregates and the observed streamflow were averaged for each month across the period of availability of observation. The shaded regions shows the aggregated uncertainty derived for LORA.

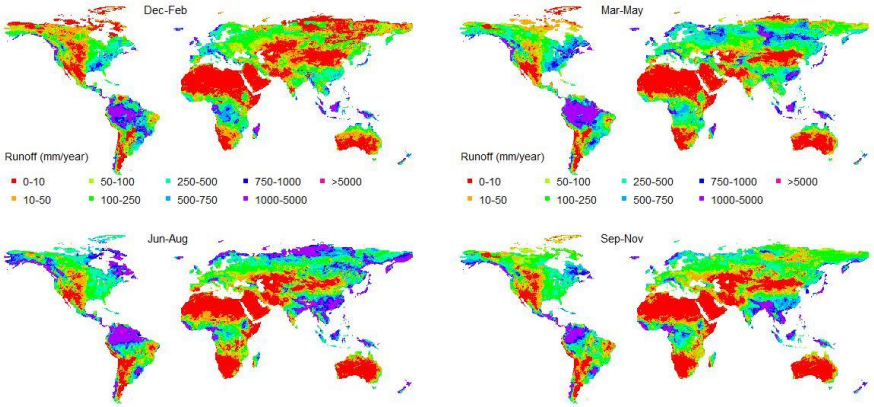


Figure 109: Mean seasonal runoff calculated for the period 1980 – 2012