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

We would like to thank Lukas Gudmundsson for his constructive comments on our manuscript. This document outlines our responses to his comments and the improvements made to the manuscript.

Response to 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.

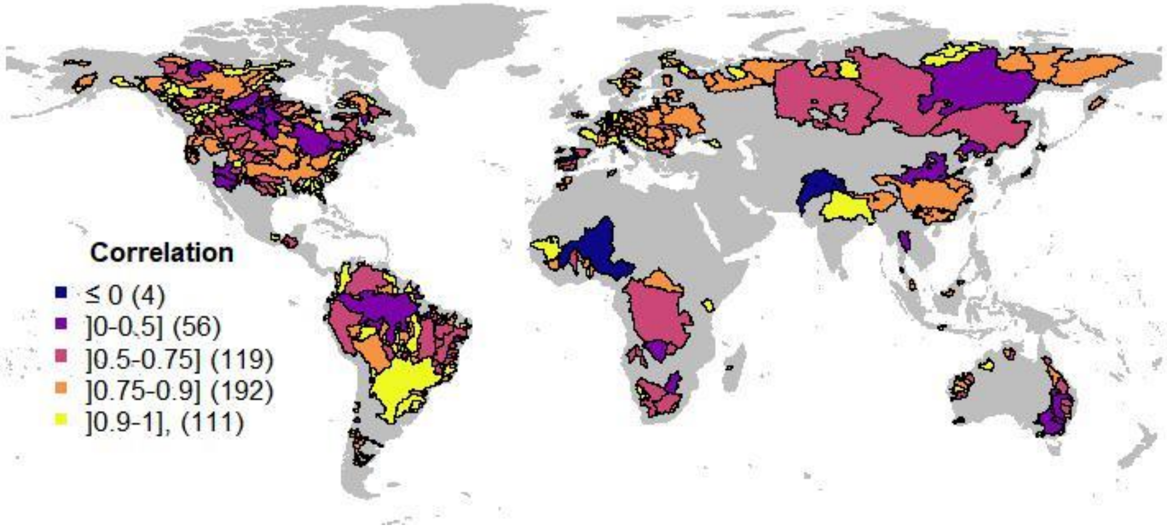


Figure S2: Temporal correlation of LORA with the observed streamflow over the gauged basins. Basins are colour coded by correlation range and their numbers are given in brackets.

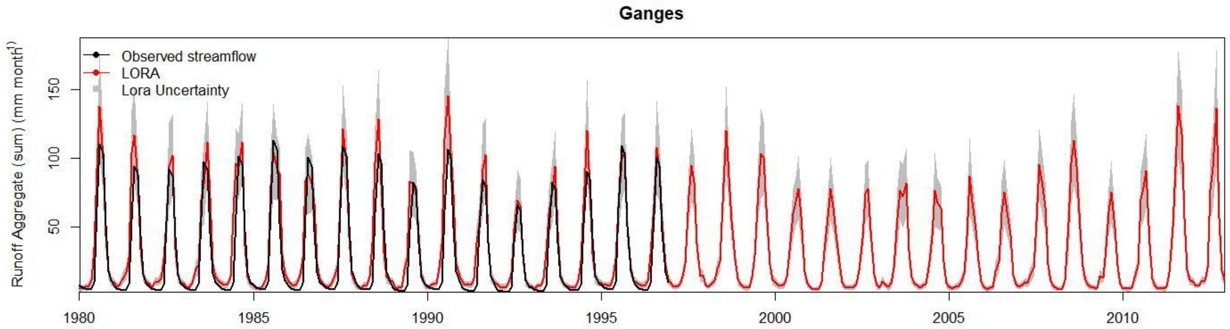


Figure S3: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Ganges basin (in mm month^{-1}). This basin was shown in yellow in Fig. S2, indicating that LORA exhibits a high temporal correlation (≥ 0.9) with the observation.

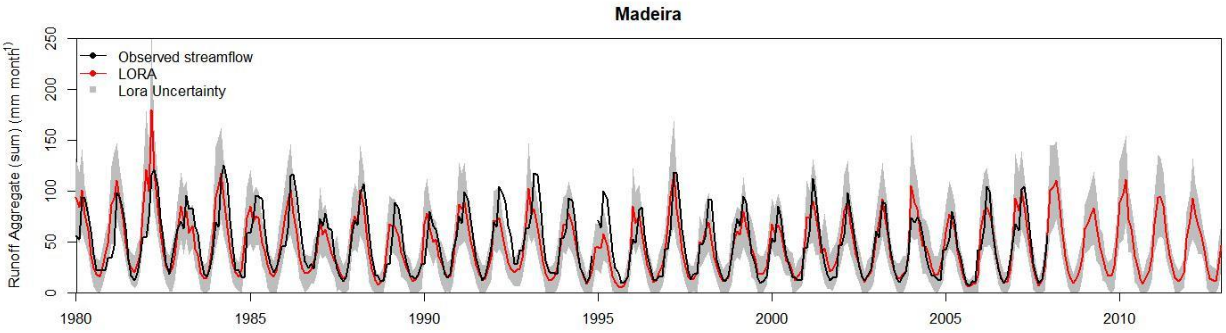


Figure S4: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Madeira basin, i.e. a sub-basin of the Amazonas (in mm month^{-1}). This basin was shown in orange in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.75 – 0.9[with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

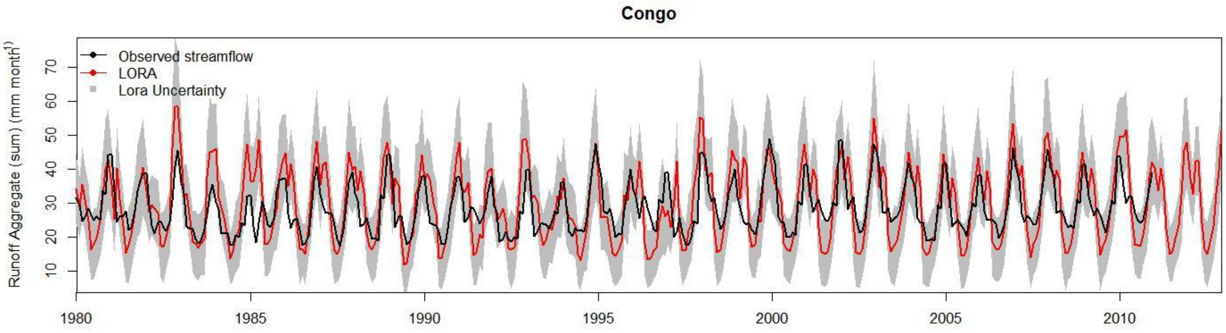


Figure S5: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Congo basin (in mm month^{-1}). This basin was shown in violet in Fig. S2, indicating that LORA exhibits a temporal correlation in the range [0.5 – 0.75[with the observation.

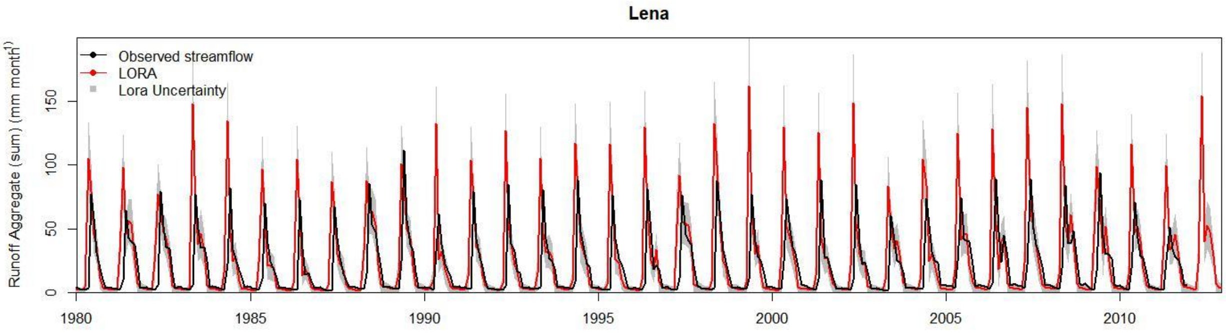


Figure S6: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over Lena basin (in mm month^{-1}). This basin was shown in purple in Fig. S2, indicating that LORA exhibits low temporal correlation (<0.5) with the observation.

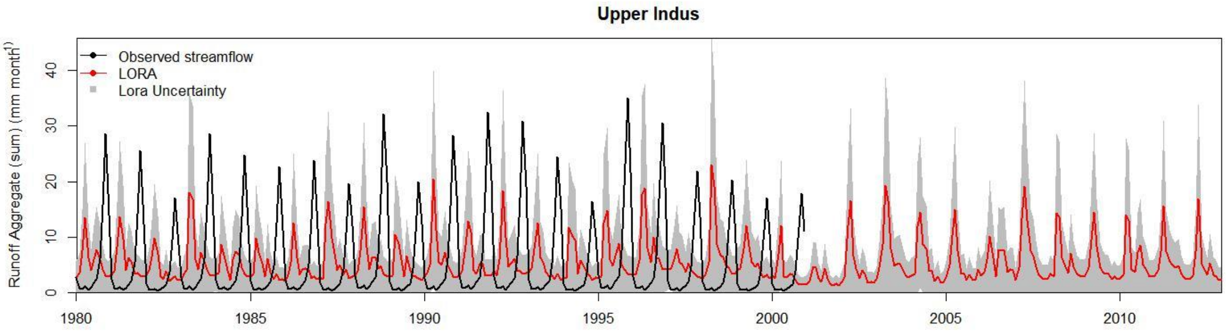


Figure S7: observed streamflow (in black), LORA Runoff aggregate (in red), and its uncertainty range (grey) over the Indus basin (in mm month^{-1}). This basin was shown in dark blue in Fig. S2, indicating that LORA exhibits a negative temporal correlation with the observation.

(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.

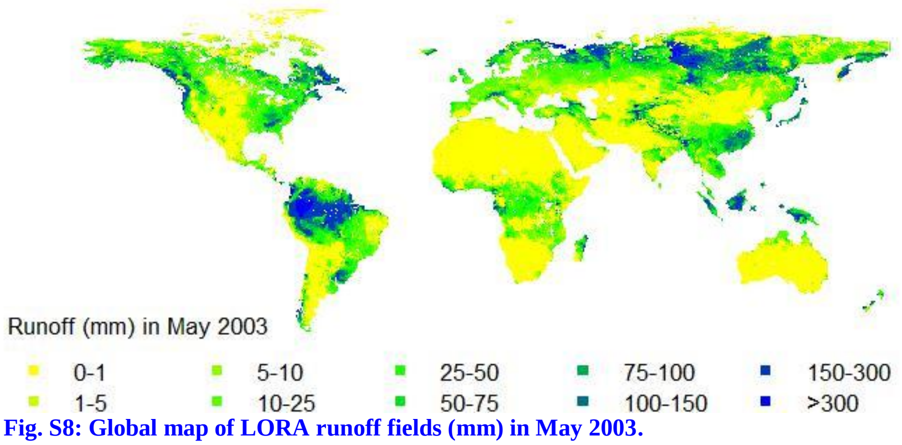


Fig. S8: Global map of LORA runoff fields (mm) in May 2003.

(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* $\alpha = 1 - Kmin(w_k)$ *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* $\tilde{x}_k^j = {Ragg_{\mu}}^{j} + \beta(\bar{x}^j + \alpha(x_k^j - \bar{x}^j) - {Ragg_{\mu}}^{j})$ *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*
>
> $\sigma_q^{2j} = \sum_{k=1}^{K}\widetilde{w}_k(\tilde{x}_k^j - {Ragg_{\mu}}^{j})^2$ *and ensures that the equation* $\frac{1}{J}\sum_{j=1}^{J}\sigma_q^{2j} = s_q^2$ *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.* $Ragg_{\mu}{}^{j}$*, 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.*

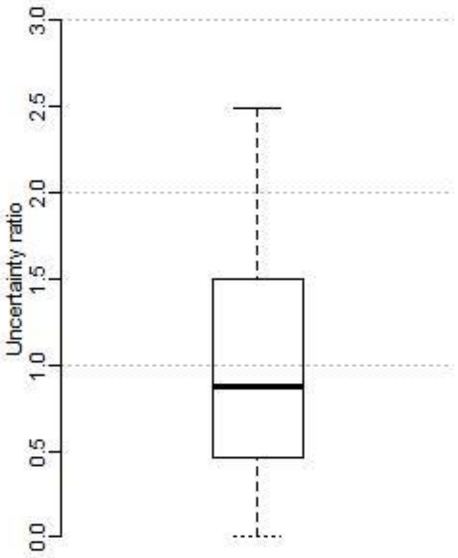


Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample to (2) the uncertainties of the spatial aggregate of the weighted computed in-sample.

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

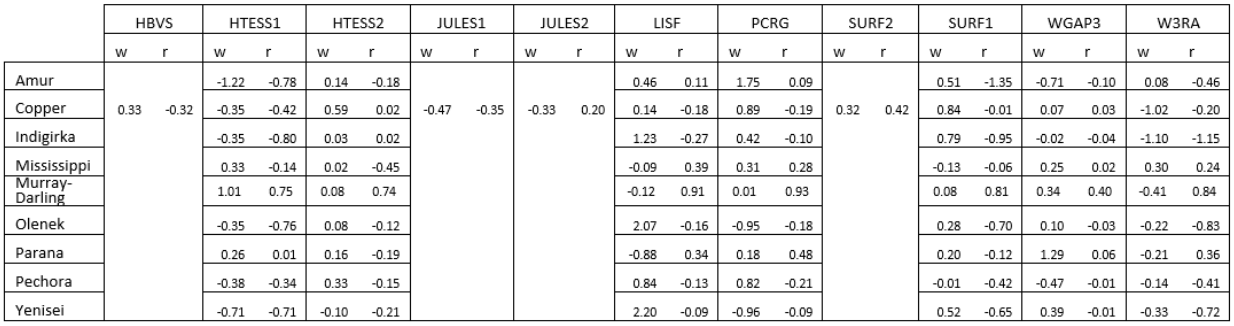
Good point. When runoff is smaller than its associated uncertainty, the uncertainty bounds will certainly produce negative runoff. This is an artefact that arises from using a generic process to determine an uncertainty range. A negative runoff is obviously not physical and requires some interpretation 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	0.33	-0.32	-1.22	-0.78	0.14	-0.18	-0.47	-0.35	-0.33	0.20	0.46	0.11	1.75	0.09	0.32	0.42	0.51	-1.35	-0.71	-0.10	0.08	-0.46
Copper			-0.35	-0.42	0.59	0.02					0.14	-0.18	0.89	-0.19			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.*

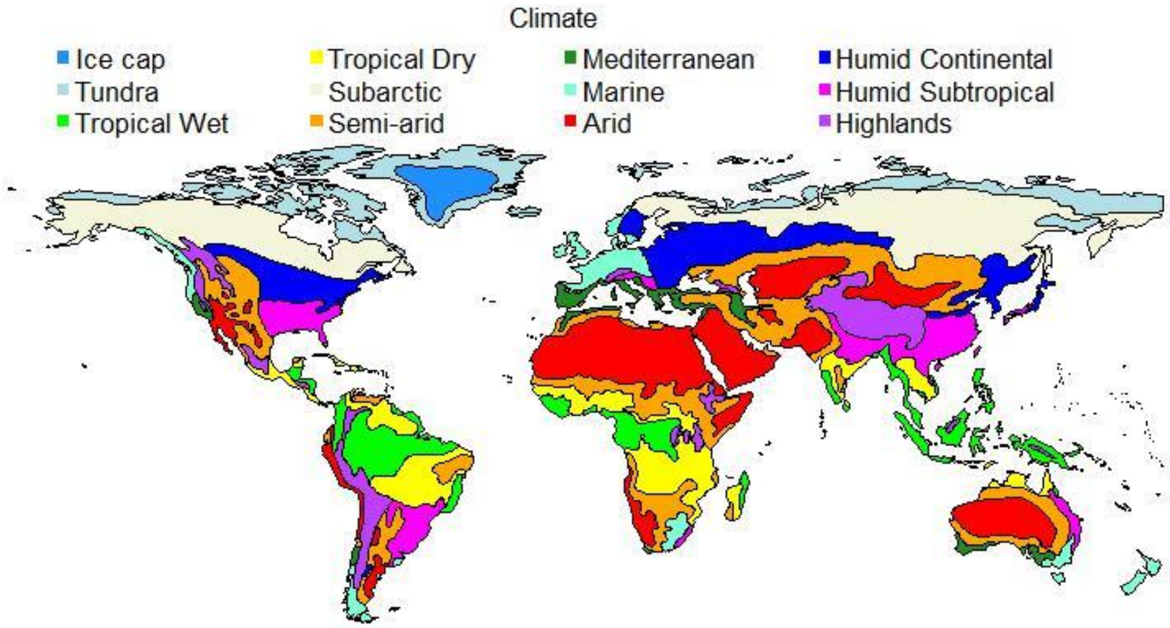


Figure S1: Climate map used in this study (available from ArcGIS online). It is a simplified climate zones map consisting of 12 broad climate classes

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.