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

We would like to thank the anonymous reviewer for their constructive comments on our manuscript. This document outlines our responses to their comments and the improvements made to the manuscript.

Response to overall comments and questions

(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.* $A = \begin{pmatrix} c_{1,1} & \cdots & c_{1,k} \\ \vdots & \ddots & \vdots \\ c_{k,1} & \ldots & c_{k,k} \end{pmatrix}$. *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:

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

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.

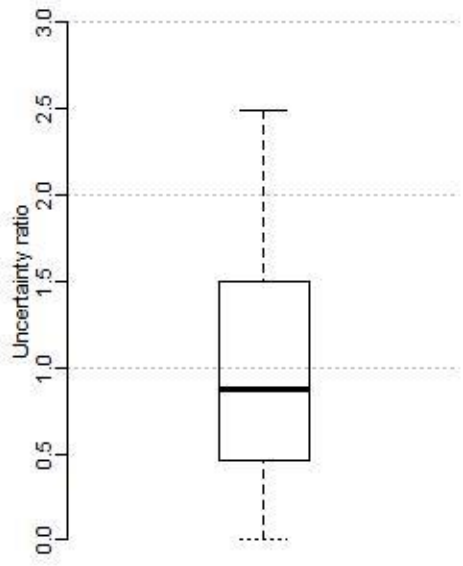

Figure 5: Box and whisker plots displaying the ratio of (1) the uncertainties of the spatial aggregate of the weighted product computed out-of-sample to (2) the uncertainties of the spatial aggregate of the weighted computed in-sample.

(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).*