Relies to the comments of Anonymous Referee #3

Authors' replies are in blue color and revised sentences are in italics.

> I general I think this manuscript lacks a coherent structure. Firstly, the introduction is mainly focussing on seasonal and long range forecasting, whereas the work actually reported is mostly concerned with classification of runoff time series. Critically, there is little or no discussion of other existing classification schemes, especially why they might not be adequate, and as a consequence it is not clear what scientific knowledge gaps is being addressed here.

We agree with the reviewer's general sentiment here. The original motivation of this research was to understand temporal variability of global streamflow in order to improve global-scale flood forecasting framework. However, the major findings of this article are not directly connected to flood forecasting, thus we have revised/restructured the focus of the abstract, introduction and conclusion to highlight the identification of flood seasons and more focus on existing methodologies.

> Secondly, I don't think there is much scientific merit in the comparison between the observed and simulated runoff series. Especially in section 4.1 where it is reported that the observed and simulated FS only share the same three months at (only?) 40% of the considered time series. Importantly, there is no discussion of what the authors would suggest is a lower limit of acceptable performance. It would have been more interesting if the mismatch between the observed and simulated series had somehow been used in a more quantitative assessment of the reliability of the model predictions.

The reviewer makes a good point, especially considering the relatively low value for identical FSs, however the original motivation of this work was not to explicitly determine the FS for hydrologic purposes alone, but rather for prediction purposes that can lead to better management. From that perspective, even if the "wrong" flood season is identified, streamflow in that season could still be predicted and used for decision-making (e.g reservoir, etc.) That said, it is of course ideal to identify the peak flood season if possible. It should be noted that the modeled PM is within one month of the observed PM > 80% of the time (i.e. within the FS), which is quite respectable in our opinion. And as the FS is defined as 3 months, the FS should contain the PM at least that % of the time. The Percent Annual Maximum Flow (PAMF) metric has been introduced to better gauge this. We have added categories (high [80%-100%], low [60%-80%], and poor [<50%]) of the PAMF to better provide the reader with a sense of PAMF performance, with some caveats. We have not provided a lower limit of acceptable performance as there are factors such as regulated streamflow and low and constant flow that may also influence PAMF value. These are all now explicitly detailed in the manuscript. One sentence explicitly addressing this review comment we have changed is P.4605, line 25-27:

Overall, however, more than 80% of both stations and sub-basins having similar PMs (± 1 month) supports that the global water balance model performs appropriately well in defining flood seasons globally at locations where observations are available.

> As it is, it seems like the performance has been accepted as it is in order to enable the production of some global map, but the usefulness (or reliability) of these maps is not really discussed. In my opinion, this makes the outcome of the study seem too open-

> ended with no firm conclusion, which is also partly down to the lack of a clear hypothesis in the beginning (i.e. identification of a knowledge gap).

Thanks for the comment. Because the PAMF is a good metric for the reliability of flood seasons in terms of containing annual maximum flows, we have not provide other usefulness/reliability. As authors' previous response, we have added categories of PAMF values to better provide a sense of performance for major FS. Additionally, we have further analyzed minor flood seasons and subsequently provided significance values (joint PAMF) for minor FS (please see updated figures below). The performance of minor FS has also been explained in the revision according to the categories of joint PAMF with well-known climate characteristics on there.

> Finally, I think the presentation of the methodology could be made more refined. In the current version it reads, I think, too much like a working paper where the individual sections are reported in the order that the authors encountered and fixed problems. Maybe group together 3.1, 3.3 and 5 to first present a coherent methodology and then apply it to the two datasets?

We agree with the reviewer's comments. We merged sections 3.1 and 3.2 and have reduced the "working paper" feel in the text.

> Specific comments:
>
> Section 2.2: Was the PCR-GLOBWB model calibrated against observed streamflow data?

The modeled streamflow used in this study was simulated by PCR-GLOBWB model without calibration against observations, however the model's performance has been validated by previous studies. We have provided the following sentence after P.4599, line 12:

The PCR-GLOBWB model has not been calibrated, thus simulation results may be biased and uncertain at course spatial resolution, however it has the ability to provide long time-series of streamflow globally, which has is sufficient to estimate long-term flow characteristics with spatial consistence (Winsemius et al., 2013). Additionally, this model has been validated in previous studies in terms of streamflow (Van Beek et al., 2011), terrestrial water storage (Wada et al., 2011) and extreme discharges (Ward et al., 2013), indicating model performance.

> Page 4600, line 17-18: I think the POT model was proposed somewhat earlier than this - see e.g. Shane and Lynn (1964) or Todorovic and Zelenhasic (1970)

Thanks for the references on POT. We have added these as appropriate into the manuscript.

> Page 4600, line 25: What is meant by 'bi or multi-model flood conditions'?

The bi- or multi-modal flood conditions imply that there are two or multiple peak flows occurring annually. Because of incoherence in the context, P.4600, line 24:25 has been removed.

> Page 4601, line3-4: Is this really a volume-based threshold? Seems to me it only considers a particular threshold based on daily runoff data. What part does volume play in this?

We apologize for the confusion here. For clarification, we have changed P.4600, line 20 – P.4601, line 7 to:

In contrast to the AM method, this characteristic of threshold can capture multiple large independent floods within a year, including the annual maximum flow, but may also miss the annual maximum flow in years in which streamflow is less than the pre-defined threshold (Cunderlik and Ouarda, 2009; Cunderlik et al., 2004a; Ouarda et al., 1993.) Thus, deciding the proper threshold level is important.

Therefore, to define the FS, and specifically the PM, both volume and magnitude aspects need to be considered (Javelle et al. 2003). To do this, we adopt a volume-based threshold technique. This technique is similar to a streamflow volume-based method in terms of capturing the Julian day by which a fixed percentage of the annual streamflow volume has occurred (Burn, 2008), however it also applies this fixed percentage across the entire streamflow record and records points where streamflow volume surpasses it, drawing from the prescribed threshold concept in the POT method. Here we select streamflow surpassing the top 5% of the flow duration curve (FDC) across all years (1958-2000) as the threshold for considering a high streamflow level, as commonly adopted in threshold approaches (Burn, 2008; Mishra et al., 2011.)

> Page 4602, line 15-: The high degree of correlation is to be expected as these different criteria are extracted from the same dataset using only minor variations in threshold levels. However, I don't understand the statement that this should somehow indicate successful success in capturing volume and magnitude. Please clarify (see also comment above).

We apologize for the misunderstanding. The objective of the volume-based threshold technique is to consider both volume and magnitude of streamflow to define the PM. Thus, if the modeled and observed PMs have "similar" correlation with indices favoring streamflow magnitude (Q_AM and Q_7days) and with indices favoring streamflow volume (Q_15 and Q_30) simultaneously, it supports that the volume-based threshold method is likely best for defining the FS. For clarification, we have changed P.4602, line 11-17 to:

Compared to the full length of PM (30 days), the flow-based classification techniques with a shorter time component (1-7 days) favor identifying flood magnitude while the techniques with longer time components (15-30 days) favor identifying flood volume. The volume-based threshold method is an attempt to bridge these two criteria.

Cross-correlations of PM between the volume-based threshold technique and other classification techniques are quite similar (0.87-0.90; Table 1), preliminarily indicating some success in capturing both magnitude and volume.

> Page 4603, line8: The statement that seasonality if often used to delineate catchments is backed-up by three (out of four) references to the same (excellent) research group. However, I don't that is enough to suggest that it is often used. Also, how did these publications define seasonality?

These references (e.g. Burn, 1997; Cunderlik et al., 2004a) define seasonality using various techniques including directional methods, POT and fixed percentage of streamflow volume, however, the key point of regionalizing/grouping catchments is to differentiate seasonality within a basin. For clarification, we have changed P.4603, line 5-11 to:

Previous studies have investigated flood seasonality as it relates to basin characteristics; for example, basins are often delineated/regionalized and grouped according to *similarity/dissimilarity of flood streamflow seasonality (Burn, 1997; Cunderlik et al., 2004a), or conversely, flood seasonality is occasionally used to assess hydrological homogeneity of a group of regions (Cunderlik and Burn, 2002a; Cunderlik et al., 2004b), thus evaluating at the sub-basin scale is warranted.*

Page 4607, line7-9: Why are seven references needed to state a well-known fact?

The reviewer makes a good point. We have selected one reference (Seleshi and Zanke, 2004) instead.

Page 4608, line 7: what unit does 'cms' refer to?

It refers to cubic meters per second (m^3/s). For clarification, we have changed the P.4608, line 7 to:

The low-flow classification is defined for annual average streamflow less than 1 m^3/sec.

Page 4608, line 25: why does (50%) refer to?

It refers to results for sub-basins. For clarifying, we have changed P.4608, line 24 – P.4609, line 2 to:

As a result, 40% of stations and 50% of sub-basins have identical peak months and 81% of stations and 89% of sub-basins are within 1 month, indicating strong agreement between model and observed flood seasons.

Page 4609, line 7-15: This reads more like the motivation for the study than a conclusion of the work undertaken. I think this belongs in an introduction.

We agree with reviewer's comments. We have moved it to the introduction.

References

Shane, R. M., & Lynn, W. R. (1964). Mathematical model for flood risk evaluation. Journal of the Hydraulics Division, 90(6), 1-20.

Todorovic, P., and E. Zelenhasic (1970), A Stochastic Model for Flood Analysis, Water Resour. Res., 6(6), 1641–1648

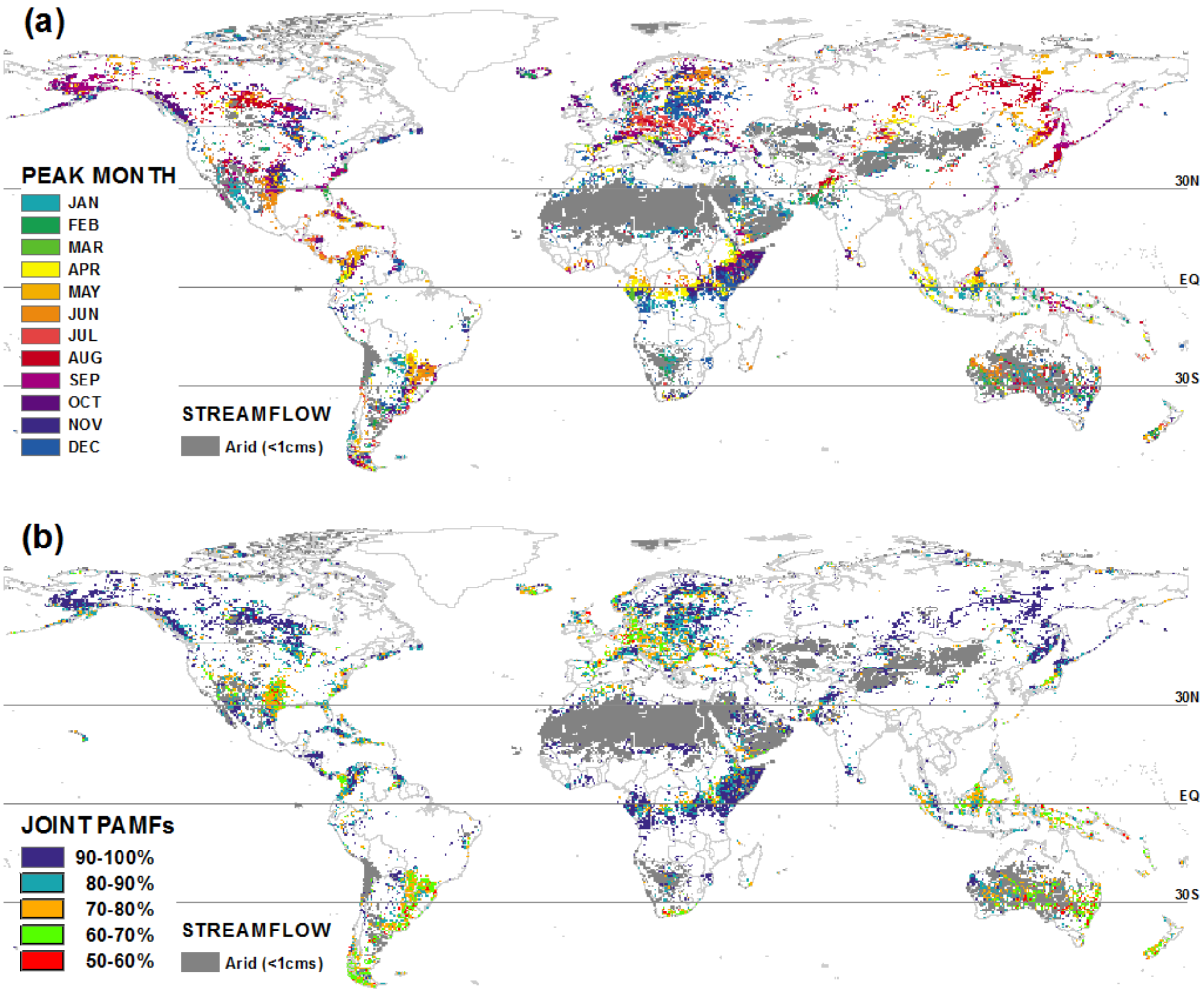

Figure 12. (a) Minor Peak Month (PM) for flooding as defined at detected grid cells and (b) joint PAMFs of major and minor PMs at corresponding cells.