	Report 1 – Referee no. 2	Response
R2.1	I would like to thank the authors for their efforts to improve the manuscript. The paper still lacks clarity regarding methodology and the did not address comments raised. See below:	-
R2.2	Although the authors claim that they have used the best-fit curve at each site, their description of methodology and presented results show the GLO was used for all sites. The last paragraph in section 4.1 and Figure 4 caption clearly states that GLO was used for all sites.	Lines 170-180 explain how the three distributions were fitted to each individual site, and the highest CvM p-value was then used to select the appropriate distribution for further analysis. Line 227 corrected to 'best-fit curves' and additional text on line 180 to remove doubt. Figure 4 caption does not state that GLO was used for all sites, rather 'GLO curves for **data pooled from all sites…**', '**from bins of catchment area**…' etc. which is quite different from saying that GLO was used for all sites. On figure 4, only 4(a) shows data from single sites, and these include GLO, Weibull and LP-III fits. We have edited the Figure 4 caption to try to avoid any confusion.
R2.3	The methodology section lacks clarity. For now it only describes the methodology to fit the curves and predicting high magnitude floods from catchment properties. The methodology section should clearly describe the steps the authors have taken to perform their analysis and evaluation of the results.	It is difficult to know what the reviewer is looking for here. The analysis includes linear and multiple regressions that are introduced where appropriate along with notes of, for example, transformations applied to variables. Adding a section to the end of section 4 that says that these equations will be used would add length, introduce repetition and would not help readers. We contend that regression analysis is sufficiently routine not to require a primer in the methodology section. We have, however, moved section 5.3.1 to 4.2, agreeing with the reviewer that this is a more appropriate location for this descriptive information.
R2.4	What is the purpose of showing fse if it not applicable to the majority of sites?	Although not applicable to the majority of sites, fse is the only way to understand the magnitude of errors within the data set. Having some error estimates provides an indication of the likely magnitude of errors at all sites, although we appreciate that errors are greater for sites with shorter records. An additional sentence has been added at the end of section 5.1 to confirm this.
R2.5	Caption of figures and tables are too long with unnecessary information that readers can easily understand by themselves. For instance, there is no need to describe what colors and symbols mean if the figure has a legend.	We have followed the journal's style guides and previous practice. Specifically, we have used comprehensive captions to ensure accessibility of the paper to readers with visual impairment who may be reliant on text translation. This also benefits readers, most commonly in the Global South, who have slow internet connections that may inhibit graphical display.

R2.6	There is no need to show number of excluded sites with records of at least 7 years in Table 2. It has been clearly made at this stage that they are excluded and analysis considers only 513 sites.	This is true, but for the avoidance of doubt and to enable Table 2 to be used in isolation we prefer to retain this information in the table.
R2.7	Line 205- 209. The ratios between flow estimated from different curves could be better presented as a Table.	We have thought about this issue at some length. The paper includes 8 tables already and we are reluctant to add further length by including another one. The information in the text is dense but can be followed easily by interested readers. Those readers who are less concerned with these details can move on to the next paragraph.
R2.8	Table 4 is better fitted in section 4.2.	Agreed and actioned – see previous comment.

	Report 2 – Referee no. 3	**Response**
	Major Comments	
R3.1	Although the authors mentioned uncertainty many times in the manuscript, they did not provide any quantification of the uncertainty in their results. This is a major issue that needs to be addressed.	We strongly disagree with this comment. There are many places where uncertainty is considered and demonstrated – for example residuals plots in Figures 5 and 6, residuals mapping in Figure 7, R^2 and se information in Tables 6,7, and related discussion in the text, and the comparative results in section 6.2.2.
R3.2	Also, in general, flood frequency analysis is not a proper method to estimate flood magnitude when you have limited data. Fitting a curve to 7–10 data points is not a reliable method to estimate flood magnitude.	The paper makes clear that the Philippines is a data poor environment and that we are trying to produce usable equations for flood magnitude estimation when data are sparse and will continue to be sparse. The paper makes clear that our approach is to use all available data and to pool this to maximise the information that we can derive from what is available, while being realistic about remaining uncertainties. This is not an uncommon approach, although in more data-rich settings there are better methods available as we note – see lines 55-61, and 62-71 for Philippines-specific commentary. We attempt in the final paragraph of the paper to assess the way forward for flood estimation in the Philippines, and other data-poor settings and note that a large majority of the global population reside in data-poor countries and regions.
R3.3	**Abstract**	This is a very helpful comment on the abstract, and we have re-written the

	While the abstract effectively conveys the general research objective and findings, in my opinion it may need some revisions to improve clarity and precision. When you start to mention R^2 and then express the added value of including the new variables, the sentence is not clear (L25-27). I suggest you revise it. It would benefit from a clearer statement regarding the limitations of the low R^2 values and the implications for design uncertainty.	abstract to improve its clarity and to remove some potential areas of mis-understanding.
R3.4	**Introduction** The introduction provides an overview of the study. It briefly describes the importance of catchment area and mean annual rainfall as predictors of flood magnitude. The authors tried to highlight the impact of pooling data from available sources to improve flood estimation when the data are limited in time and space. I suggest merging the two middle paragraphs of the introduction to make it more concise and clear. Also, the hypothesis and research questions of the study are not clearly stated in the introduction. It would be better to state them explicitly.	We have decided not to merge paragraphs as the current structure appears, to us, to separate distinct aspects of the background to the study. We note the comment regarding research questions and have amended the final paragraph of the introduction accordingly.
R3.5	**Methodology** In the section \`Data sources\`, the authors provide a detailed description of the data sources and the process of data collection. Different sources introduce distinct uncertainties and biases into the analysis. For example, in Figure 1, some sources (red and blue dots) are more concentrated in regions such as the north of the country, while in the west and south—where there is lower rainfall and lower contribution from tropical cyclones—we have no or only one source of data. This may introduce bias in the analysis. The authors should discuss this issue in the manuscript.	This is a valid comment, but we have to work with the available data. The maps of residuals in the paper do not show systematic regional patterns, nor do residuals from regressions show systematic effects from different data sources (Figure 5 a-c; Figure S5c). These results give us confidence in the comparability of the data from different sources, and we have checked the methodology used in collecting the data and it is consistent (stage-discharge curves, often with contextual notes about issues regarding reliability of measurement of the highest flows). In section 6.3, we discuss if there are any issues surrounding the data and their amalgamation. One additional sentence "*Comparison of results from different data sources (e.g. Figure 5(a-c)) shows no statistically significant differences between results from analysis for each of the data sets, so supporting our*

		amalgamation of the data from different sources for aggregated analysis." has been added to 6.3 to re-state our confidence in the data sources.
R3.6	More details regarding the screening criteria for data quality and the rationale behind the selected catchment properties would improve transparency. For example, three sources of data are used in the analysis; while they were recorded differently, in different periods of time, and likely with different measurement techniques, the method of merging these data should be discussed in the manuscript. It is highly likely that the quality of measurements before the 1980s is lower.	We think that the paper contains sufficient information regarding the data selection, and also refer to the previous comment and response. Although measurement methods were different, the principles of stage-discharge gauging have been largely unchanged for two centuries and we have confidence in all of the data. The books containing data for the 1910-1920s include excellent and detailed information on rating curves and site characteristics. For example, we have been able to use rating data to infer periods of river bed aggradation or degradation at some locations. Hence, the quality of measurements is likely to have been higher for the earlier data, although we do not have rating curves from more recent data collection periods to assess this further. Note that the analysis here uses only annual maxima, rather than full flow records which would be more affected by gaps in the data record and other quality issues.
R3.7	**Analysis Methods** I am curious to know whether you ever tried to employ two peaks per year or any POT analysis to identify the peaks in the data, instead of only using the annual maximums. This approach would give you more freedom not only to select the highest peak in the year but also the second highest independent peak in the year. This could help you better understand the flood frequency in the region, as the second peak may occur in another season and allow you to better capture your basin's behavior. Then, you could continue to determine Q_med of the new series of peaks.	This is an interesting suggestion. It would have been possible, although some of the flow records do not contain sufficient information to identify all POT events (for example, at some sites we only have peak flow available for each month of the year and at others only the annual maxima are provided). Note also that in this tropical environment there is strong and generally consistent flow seasonality. If the paper is accepted, all of our data will be openly available for others to use and to undertake additional analyses.

R3.8	The manuscript provides a thorough description of the curve fitting using L-moments and the subsequent regression analyses. Yet, the discussion on the potential biases arising from combining data of varying quality and the choice of best-fit distributions (with respect to low R^2 values) deserves further elaboration. Moreover, since the study aims to estimate extreme floods, linear regression may not be the best approach. The authors should consider using a more robust method, such as quantile regression, to account for the non-linear relationship between the predictors and the response variable.	There are several issues raised here, and we concur with the overall premise of needing to use the most appropriate methods for the data that are available. As noted above, we have had to combine data from different sources and to rely on relatively short records in order to produce a data set with national coverage. Quantile regression would have required longer and more complete data records than are available for nearly all of our sites, so severely limiting our analysis. We have previously applied quantile regression methods (Franco-Villoria et al., 2019, DOI: 10.1002/env.2522) and appreciate the potential of this approach where suitable data are available. Further, the analysis methods used are standard (eg Kjeldsen, 2013) and we adopt this methodology to ensure consistency with previous work. The paper comments on some of the background analysis that we undertook to assess the data (Figure S6 shows cross-correlations that show the nature of relationships between all variables utilised). We are considering undertaking further analysis of these data, potentially using GAM methods, but consider this to be a separate project from the current work. The potential value of design equations that use established methods in a data-sparse country should not be under-valued, and we consider our approach to be the most robust and reliable way to develop these equations at this time.
R3.9	What is the set threshold of low CvM p-values used to exclude data from the analysis in L183?	It is best not to interpret CvM p-values against a critical alpha value, but to compare the CvM statistics between distributions (Asquith, 2020). The median p-value of best-fit curves was 0.93 and this has been noted in the text.
R3.10	**Results** - The correlation approach in Table 4 does not lead to a new conclusion. The fact that a larger catchment area leads to a higher	This comment is understood and has been addressed in response to other reviewer's comments by moving Table 4 into section 4.2 where it is presented as background information.

	correlation is not a new finding. It is the same with the DPLBAR variable, the length of the streamflow network, and the mean annual rainfall. Therefore, your addition in Table 5 should be highlighted. I suggest restructuring the results section to emphasize the new findings of the study.	
R3.11	Perhaps testing and illustrating your approach on only the new dataset as a test case would be a good idea to show the robustness of your approach. This will also help in understanding the uncertainty in the results.	We would agree with this suggestion IF the aim of the paper was to test a new method. However, our aim is to produce reliable and robust design equations for the Philippines and so this calibrate and test approach is less appropriate. We note that several of the plots differentiate the different datasets. If there was bias related to the data sources, this would be apparent in these plots.
R3.12	What would be the expected best R^2 value by adding the new variables? It would be better to have a benchmark to compare the results. What is the ideal R^2 value for flood frequency analysis in the region? Is the benchmark 0.92 in Papua New Guinea? You could randomly generate some synthetic data and try to estimate the flood frequency analysis to see the ideal R^2 value.	Previous global analysis (eg Meigh et al., 1997) has reported R^2 values from 0.61 (Kerala) to 0.92 (PNG). Equations that go beyond catchment area and one rainfall variable can improve R^2 values slightly (eg in Indonesia improvement from 0.881 to 0.889 by adding both catchment slope and lake area terms).
R3.13	As you mentioned, land use change is a major factor in flood frequency analysis, and you employed almost current land use data in the analysis. This is a significant challenge and limitation of the study.	Yes, we acknowledge this and make some comments to this effect in the paper. At the catchment scale, the influence of changes over time may be less than at smaller scales. Historical land-use data do not exist for the Philippines, but we note that none of the catchments in the study is extensively urbanised. Similar challenges would be encountered in almost any tropical country.
R3.14	The abbreviations in this study are not mathematically scientific, such as AREA or RMED. It would be better to use the full names of the variables in the text and use better letters for the variables. For example, A for area, and R_m for RMED, and so on.	We have followed convention from the UK Flood Estimation Handbook (FEH) in naming variables and consider that it is appropriate to retain these variable names to facilitate easy comparison with a wide range of previous studies.
R3.15	Since the results are mainly presented on Q10 and they are not significantly appropriate for flood control and design, it would be better to include a discussion on the results and the limitations of the study.	We note Q2 (close to the geomorphologically effective bankfull flood) and Q10 results in the tables in the paper. Q10 was selected for presentation rather than Q100 as the relatively short time series available for

		analysis make estimation of Q100 less reliable. We have added a note to the conclusions to make the point about the limitation of Q10 for design purposes.
R3.16	**Discussion** - It would be valuable to discuss the limitations (e.g., stationarity assumptions, data quality issues, and land use change) more explicitly and to outline potential paths for future improvement, such as incorporating non-stationary models or enhancing continuous monitoring.	Section 6.4 does address the limitations of the study – stationarity and alternative approaches are referenced in the final paragraph of 6.4. Section 6.4 also contains suggestions for grouping of catchments for analysis, noting that this may not involve grouping adjacent catchments. We consider that there is sufficient acknowledgment and discussion of limitations and potential enhancements to the study through the paper.
R3.17	Tropical cyclones were not part of your investigation; however, they play a role in the discussion.	Rainfall contributions from cyclones are introduced in Figure 1, and they are mentioned in the discussion as an issue that may require further consideration to account for possible changes to precipitation patterns due to climate change. A sentence has been added to the end of section 6.4 for clarity.
R3.18	Climate change and spatiotemporal variability in the region are not discussed in the manuscript at all, despite the merged data varying over time.	Within the limitations of what is an already lengthy manuscript, we do report the most recent (Tolentino et al., 2016) assessment of future hydroclimatic change. There are few studies of climate change in the Philippines over the past century and a lack of data to make reliable statements regarding past changes.
R3.19	The discussion section is generally long. I suggest revising it to be brief, more concise, and clear. However, the current form is good for readers to understand the results and limitations of the study.	This comment is somewhat contradictory. Several of this reviewer's comments ask for more discussion of the context and limitations in the study, so it is difficult to see how we could shorten the discussion without over-simplification. Some minor changes have been made in response to this, and the other, reviewer's suggestions that we hope clarify our reasoning.
R3.20	The comparison with HEC-HMS modeling lends additional credibility, though the discussion might be expanded to explain the practical implications of the observed	We agree that we could expand on this further, but note the reviewer's previous comment suggesting making the discussion shorter. We have tried to use the HEC-HMS comparison fairly, and

	discrepancies between instantaneous peak flows and daily mean flow estimates.	not to over-state its value as the HEC-HMS modelling relied on several assumptions that we are not able to evaluate.
R3.21	**Conclusion** The conclusion is well-structured and effectively summarizes the key findings of the study.	Noted and appreciated.
	Minor comments	
R3.22	**Abstract:** L18: Split the long sentence `However, the global …` into two sentences for clarity. The current sentence contains four commas.	Edited to improve clarity.
R3.23	L23: What does `national and regional scales` mean? Are they two different scales?	Abstract clarified 'both national and regional'. Section 3 explains the basis for regionalization.
R3.24	L25: The term `GIS-derived` is not needed here. You can simply say `geospatial catchment characteristics`.	Done
R3.25	L30: There is a redundancy with the term `predictive equation` in the same sentence.	Done
R3.26	Introduction: L43: The sentence `The resulting equations …` is not well connected to the previous sentence. The starting lines are quite good, but there is a gap between the first and second parts of the first paragraph.	A connecting sentence has been added: *"Understanding flood magnitude and frequency is crucial for designing mitigation strategies, and this understanding relies on using empirical analyses to generate predictive models."*
R3.27	L79: A reference to Figure S1 is needed.	Done
R3.28	Please provide a map of the available length of time series in the Philippines. This will help in understanding the data availability in the country (L80). Although the time period is indicated in Table 1, it is not clear whether the records are continuous or if there are gaps in the data. Alternatively, you can provide some sentences in the text to explain this issue.	This is quite a complex issue – the opening paragraph of section 3 does address the nature of the records and explains how we have combined some records from adjacent measuring sites.
R3.29	How do you define short time series? Is it less than 35 years? (L80). It would be better to provide a definition for short time series or a reference for the definition.	We have changed to 3-20 years, as 20 years is often used as an arbitrary threshold for undertaking flood frequency analysis. 'Long' has been replaced by 'multi-decadal' where the length of data records is first introduced.
R3.30	L84: The `FEH` abbreviation has already been defined previously in L50.	This has been corrected.

R3.31	**Data sources:** Figure 1: In the caption, it is mentioned that `the four climate types that have been identified for the Philippines (Coronas, 1920)`. Since the climate types were identified in the 1920s, is there any more recent climate type identification for this region? Given global warming and climate change, the climate types may have changed or been better defined in recent years. Figure 1: Please replot panels b and c and use discrete colors instead of gradient colors. Also, Figure 1C does not support any of your results except for a sentence in the conclusion. It would be better to remove it from the manuscript or integrate its insight into your interpretation.	The Coronas (1920) classification continues to be used for the Philippines, and is extensively referenced in climate and hydrological publications. We are unaware of more recent re-evaluations of these climate types, and the familiarity of Coronas' classification to Philippines readers will aid their understanding and ability to interpret our results. We do not agree with the recommended re-plotting of 1b and 1c, as gradient colours better reflect the interpolation that has been undertaken to generate the maps. Figure 1c provides context here – as noted, it is referred to in the discussion. Had we omitted Fig 1c, we would have expected reviewers to ask about the importance of tropical cyclones!
R3.31	I suggest moving Table 1 to the supplementary material, as it is not necessary in the main text.	We are following journal guidelines that discourage long supplementary materials. As the paper focuses on integrating data sets, we consider that it is useful for readers to see these data sets described at the outset.
R3.32	**Analysis Methods:** To achieve more consistency in the manuscript text, I suggest adding Q5 and Q50 in Figure 2, and so on, in your text.	This could be done, but would add length and complexity to the paper. The full data set will be made available if the paper is published enabling users to compute Q5, Q20, Q50 or other return periods as they wish.
R3.33	Since Table 2 does not show any relation between the size of the catchment, climate type, and the best-fit curve, is there any geographical pattern in the best-fit curve? For example, do catchments in the north of the country have the same best-fit curve? What if you plot the best-fit curve on the map of the country? Usually, subcatchments in the same basin may have the same best-fit curve since they are flow-connected.	This is an interesting issue that we looked at in detail when producing the plots for the paper. There are few consistent patterns, but we do observe some consistency within large catchments as suggested. Given the complexities introduced by variable lengths of record, a map of best-fit curve type is not especially helpful for this paper and would require extensive explanation to make sense of what is a complex pattern.
R3.34	L204: The phrase `(Figure S1) show this pattern` is unclear. I have not seen this pattern in Figure S1. Please revise the text. The mentioned figure is `Administrative regions of the Philippines`. Since the numbering starts from north to south, it would be better to reorder the legend of the	We have revised Figure S1 to rearrange the numbering in the legend and rechecked the text to correct reference of figures.

	figure to follow the same order instead of alphabetical order.	
R3.35	L208: The reference to Figure S2 is incorrect. It is currently written as \`(Figures 2, S1)\`; it should be \`(Figure S2)\`, as seen in the supplementary material. Also, the figure itself is not well plotted.	We have corrected the numbering of the Figures.
R3.36	The quality of Figure S3 is too low. It is not readable. Please revise it. The current figure overlays the main curves on top of each other. The area of concentration should be zoomed in to see the differences between the curves on the right part of the x-axis.	The symbol size on Figure 3 has been reduced to aid clarity. The high-resolution figures that will be used for final production are clear enough on-screen.
R3.37	The same applies to Figures S4 and S5. However, they are slightly better. I think these figures can be highlighted for regionalization since they are important for understanding how the curves differ by region, climate type, and catchment size. The current format does not help in understanding the differences between the regions.	See previous comment. We have examined plots on a region-by-region basis and no patterns emerge from this. As the results in the paper show, the regions are not hydrologically distinct. As these regions are very widely used in the Philippines, as they are administrative regions, it is useful to include some regionalisation in the paper for the benefit of local users. However, climate type and catchment area are more informative as ways of grouping the data than the regions.
R3.38	L247: As far as I know, we have free global DEMs with 30 m resolution. So why did you resample the DEM?	It is better to resample a higher resolution DEM than to use data collected at lower resolution.
R3.39	**Results:** L287: The phrase \`This contrasts with Meigh's\` should be moved to the conclusion.	We have not done this as this sentence only makes sense within its current context – moving it to the conclusion would require 2 or 3 sentences of explanation that would be out of place in the conclusion.
R3.40	Table 4: Instead of \`NA\`, write \`-\` in the table.	We have explained NA in the caption to Table 4.
R3.41	Theoretically, your Q_med should be equivalent to your Q2 when you have a limited length of time series. If you look at Table 4, the columns Q_med and Q2 are almost identical. Also, the correlation is sensitive to the number of data points.	Noted and agreed.
R3.42	Table 5: The alignment of the table is not correct. Please revise it, and make it more readable.	Done (partly) and will require final alignment by the journal in production
R3.43	Figure 5 and subsequent figures: Please elaborate on "normally distributed" in the figure caption, especially for subplots b, c, e, and f.	Have added (Gaussian) in Figure 5 caption, and subsequent references to 'normally distributed' should all be interpreted as being Gaussian

R3.44	Set a fixed significance level for the p-values in the text. In section 5.3.3 it is 0.05, while previously it was 0.01.	We do not use a significance level to assess our calculated p-values against; rather, all of the p-values reported are direct computations of the probabilities of Type I errors. This provides readers with a ready way of assessing significance, rather than setting arbitrary pass/fail significance levels.
R3.45	Figure S7 must be revised. The current figure is not readable enough; it is a bit small, and the selected colors do not help the readers understand it. Also, since this figure has three parts, the main body of the manuscript does not support it well.	Some of the lines will be enhanced for the final production version. The colours have been chosen for consistency with other figures in the paper (which in turn are based on previous literature). We have tested the figures using the Coblis colour-blindness simulator and have selected combinations of colours and line styles that enhance the accessibility of the figures. Note that on-screen high resolution figures as will be used in final production are very clear.
R3.46	**Discussion:** Figure 8: It seems that the x-axis of panel b is not correct. Please revise it.	Figure 8(b) x-axis label is correct. This plot is checking for a catchment size effect (and bias) in the results.
R3.47	This study contributes to hydrological modeling by demonstrating how pooling individually short historical flood records—combined with high-resolution geospatial data—can produce nationally applicable flood estimation equations even in data-sparse tropical regions. The `Recommended design equations` section is a part that the authors may consider including in their analysis.\ I suggest authors consider the above points in their revision and I look forward to seeing the revised manuscript.	Noted, and Tables 6-8 provide design equations and an example of their use.