Response to Referee 2:

In summary I found this an interesting paper and a useful contribution to the incorporation of non-stationarity in water infrastructure planning. As noted in the comments below, I think a restructuring of the paper is needed to improve the contribution and make it more useful to water planners and managers.

Response: We thank the reviewer for acknowledging the contribution of our work and we appreciate their careful review. We have provided detailed responses to every comment below.

Most importantly, the thrust of the paper should be the demonstration of a method – with essentially no calibration or quantitative validation of results for the Truckee river, this case study is not useful in itself.

Response: We agree with the reviewer that one of the primary purposes of this manuscript is to demonstrate non-stationary flood frequency analysis and to make these tools more accessible to the hydrologic community. However, we respectfully disagree with the assertion that this study is not useful in itself. While it is true that we are working in a data limited basin, we would argue that this is true of many if not most basins across the globe and it does not change the fact that water managers need to make decisions. Despite data limitations, we still made every effort to both calibrate and validate our model. As noted in the comments below we have expanded the manuscript to make these aspects of our work more quantitative and transparent.

1) The title "Climate change and non-stationary flood risk for the Upper Truckee River Basin" does not reflect the contribution of the paper, which is really a demonstration of a methodology. This is noted in p. 5082, line 18: "This paper provides an end-to end demonstration of nonstationary GEV analysis coupled with contemporary down-scaled climate projections (specifically, downscaled climate projections from the Coupled Model Intercomparison Project Phase-5 (CMIP-5)), to quantify how the risk profile of existing infrastructure, designed on the basis of a specified flood event, evolves with time over its design life." As noted in my comments below, the Truckee seems to be more of a demonstration data set. This should not be interpreted as a paper providing significant planning information for managers of the Truckee system.

Response: As noted above, we are of the opinion that this paper has two primary contributions. The intent of this work is to both demonstrate the method and provide site specific results for the Truckee basin. To make this point more clearly we have revised the sentence in question to read:

> ***"While the methodology used for this analysis is previously established, this paper provides the first end-to-end demonstration of non-stationary GEV analysis coupled with contemporary downscaled climate projections (specifically, downscaled climate projections from the Coupled Model Intercomparison Project Phase-5 (CMIP-5)), to quantify how the flood risk profiles may evolve in the Truckee river basin over the next century. The intent of this work is 1) to investigate potential flood risk changes over time in the Truckee basin and 2) to demonstrate the applicability of non-stationary techniques in a regional flood analysis to make these tools more accessible to the hydrologic community." (Revised Manuscript, Page 6, Lines 20-28)***

Also we would like to note that this study was commissioned by water managers and our results have been communicated directly with decision makers. While we acknowledge that there is limited data,

we respectfully disagree that this work does not provide planning information. We hope the changes we have made in response to the Reviewer comments will help make this point more clearly.

2) p. 5084, line 15, it states "we simulate unregulated flows from 1950 to 1999 using the Variable Infiltration Capacity (VIC) model and validate results using the available unregulated flow estimates." There does not appear to have been any calibration done as part of this effort. Is the validation done on an uncalibrated model? Some basic hydrology validation statistics would be helpful (NS, RMSE, ...) is assessing the streamflow simulation. The qualitative interpretation like "in close agreement..." (p. 5092, l 9) and "in good agreement" (p. 5092, l 14 and l 19) needs to be quantitative. That would provide support for the claim "This demonstrates that the model behavior is a reasonable match to the natural system."

Response: The VIC model we used for this analysis was developed and calibrated as part of the Bureau of Reclamation West Wide Risk assessment (Gangopadhyay e al., 2011). We agree with the reviewer that calibration was not adequately discussed in the original manuscript and we have added the following text to the revised version:

> ***"The VIC model used for this analysis was developed and calibrated as part of the Bureau of Reclamation's West Wide Climate Risk Assessment (WWCRA). The WWCRA VIC model encompasses the western US. Streamflows were evaluated at 152 locations primarily from the USGS Hydroclimatic Data Network [Slack et al., 1993] and 43 additional locations of importance to Reclamations water management activities. Among the evaluated locations are several in the Truckee basin including the Truckee River at Farad. For details on model calibration and development we refer the reader to Reclamation [2011] and Gangopadhyay et al. [2011]. While we do not discuss model calibration further here, in the subsequent sections we provide additional model verification for flood simulation in the UTRB." (Revised Manuscript, Page 9, Lines 7-16)***

Also we added the following clarification to section 4.1:

> ***"A suite of models were fit to the logarithms of block (cool season, November-April) maxima flows (simulated by the calibrated VIC model)" (Revised Manuscript, Page 14, Lines 11-12)***

We agree with the reviewer that quantitative validation statistics would be ideal. However, given our small sample size (six floods) we do not have enough points to calculate metrics which would be statistically significant. We have added additional specifics throughout this section in the following places:

- ***"generally good agreement" (p. 5091 l17) Replaced with "*****the maximum percent difference between the natural logarithm of simulated and observed flows is 12%."*** ***(Revised Manuscript, Page 15, Lines 20-21)***
- ***"matches very closely" (p. 5091) added "(percent difference in the natural logarithm of flows are 0.5% and 1.2% respectively)" (Revised Manuscript, Page 15, Lines 26-27)***
- ***"in close agreement..." (p. 5092, l 9) added "(the difference between the natural logarithm of simulated and observed flows is the smallest of any event at 0.5%)" (Revised Manuscript, Page 16, Lines 10-11)***
- ***"in good agreement" (p. 5092, l 14) replaced with "the VIC simulated flow falls within the interquartile range of the GEV model" (Revised Manuscript, Page 16, Lines 16-17)***

- ***"we note good agreement" (p. 5092, l 19) we rewrote this paragraph to read: "In general, Figures 3 and 4 show that the VIC simulated flows match closely with the observed floods (based on percent difference in the natural logarithm of flows) and that the interquartile range of the GEV distributions encompass the observed and simulated flows in most instances. Figure 3 does illustrate some of the complications in matching individual events, however based on analysis of the driving forces behind each individual event we are able to document the sources of these discrepancies. Based on this analysis we conclude that the model behaviour is a reasonable match with the natural system." (Revised Manuscript, Page 16-17, Lines 28-32 & 1-2)***
- ***"This demonstrates that the model behavior is a reasonable match to the natural system." See the revision above.***

3) p. 5085, line 20, 234 projections are analyzed, which lumps together extremely aggressive mitigation futures (like RCP 2.6) and more business as usual scenarios (RCP 8.5). It would seem that, for planning purposes, these should be separated. Only one pathway into the future will actually be experienced, and the variability among GCM projections should reflect that. It would make more sense to present each RCP separately, as this allows a consideration of the variation due to following different pathways from the variation in how the atmosphere might respond to the changed atmospheric conditions. These are very different sources of variability.

Response: This is an interesting question and we agree with the reviewer that between the GCM projections there are multiple sources of uncertainty. In our original analysis we did compare results based on emissions scenarios. The figure below plots the risk of a 1 day flow exceeding the design flood of 37,600 cfs in 10 years for three time periods with results grouped by the emissions pathway. As you can see here in all cases the differences between emissions pathways are smaller than the changes over time. Based on this result we decided not to group our results by emissions pathways in the final paper. Still we agree that this is an important point and we discussed it on page 5096 lines 5-11:

> ***"Although the figures are not shown here, results were also grouped by RCPs to analyze connections between greenhouse gas emission rates and changes in flood risk. We observed no clear trend in flood risk based on the different RCPs. This indicates that the variability between GCM model form and initial conditions likely overwhelms the influence of greenhouse gas emissions when comparing between scenarios. In other words, the variability between projections within any RCP scenario is larger than the difference between RCP scenarios."***

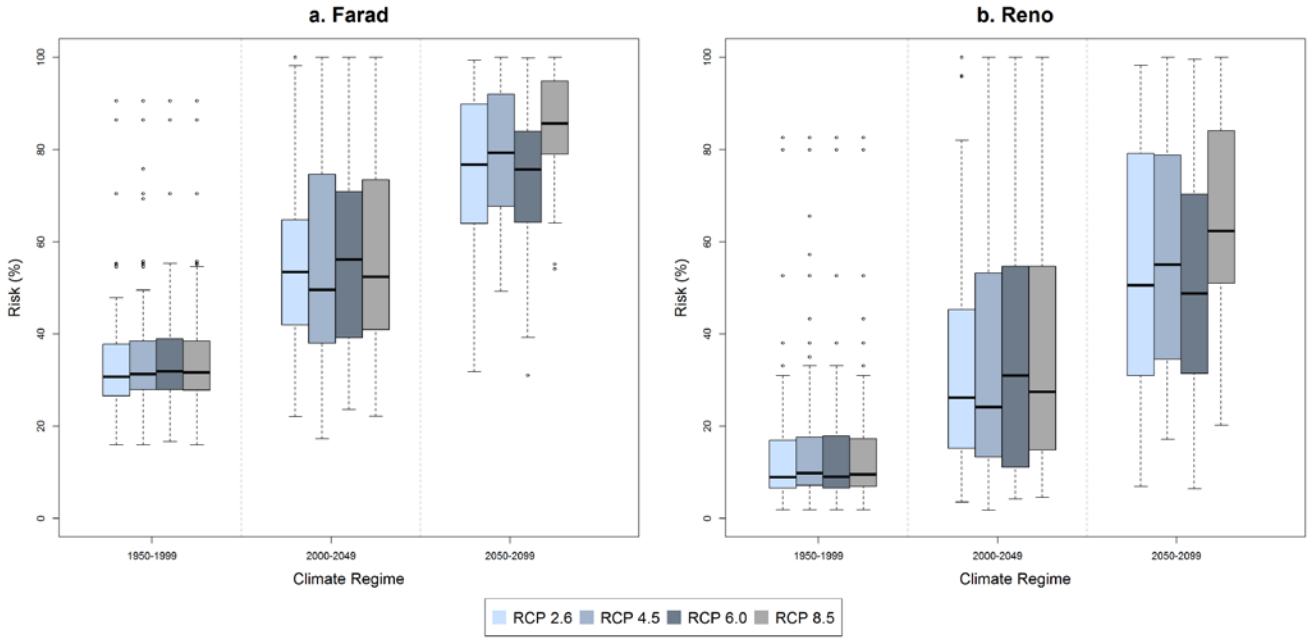

Figure 1: Risk of 1 day flood exceeding 37,600 cfs in a ten year period grouped by emissions pathways.

A second point is that this includes multiple contributions from some GCMs and single contributions from other GCMs. Plenty of research demonstrates that different runs of a single GCM are less independent than runs of different GCMs, and lumping them all together inappropriately weights models that happen to have submitted many runs as part of CMIP5. To demonstrate the method for this paper, there is no need to use 234 projections – a more carefully selected set of a dozen or two would seem to suffice, and also provide a better demonstration of appropriate use of climate model output.

Response: While we understand the reviewers point, we would like to note that there has been much scientific debate on this topic in recent years. As of yet, there is no consensus on the "best practice" for selecting climate projections and likely there may not be one given the fundamental non-linear dynamics of the earth system. The IPCC presents every GCM scenario as equally likely and there is significant debate about whether it is appropriate to select projections based on historical simulation skill because this approach is based on the assumption that historical simulation skill is correlated to realistic simulation of sensitivity to increased greenhouse gasses. As noted in the West Wide Climate Risk Assessment (Reclamation, 2011):

> ***"To date, there is still limited evidence to support such a philosophical bridge (Reichler and Kim 2008; Santer et al. 2009; Pierce et al. 2009). It also has been shown that when such skill assessments are based on many climate metrics (e.g., Tebaldi et al. 2005; Mote and Salathé 2010), the clarity of "better" versus "worse" climate models is less obvious than when the assessment is based on few metrics (Brekke et al. 2008; Reichler and Kim 2008; Gleckler et al. 2008). Even when the historical skill assessment results have been used to rank and cull climate models, thereby conditioning the assessments of future climate uncertainty (Brekke et al. 2008) or detection and attribution of causes for trends in historical atmospheric water vapor over large spatial scales (Santer et al. 2009), the effect of model culling on***

assessments has been minor. These latter results suggest that other factors, beyond historical skill, are driving impact assessments from projected climate conditions within an ensemble, including emissions pathway and a GCM's —natural variability."

Therefore, we decided to use all 234 projections because we do not have a rational way to "carefully select" a smaller subset of projections. Again, there is presently no one objective way to subset projections.

4) Section 2.3, last paragraph, it is mentioned that the Bureau of Reclamation has developed an archive of downscaled data, but then the downscaling is described as if it were done again for this effort. Were the projections obtained from a published archive? If so, state that, provide an appropriate citation and acknowledgement (I see on the BoR site there is a standard citation and acknowledgement).

Response: The BCSD CMIP-5 projections are currently available on the BoR archive. However, at the time of this analysis the complete set of BCSD projections was not developed yet and so we extended the CMIP-5 analysis for our domain as part of this project. Still, we agree with the reviewer that it would be helpful to point readers to the most recent references for this data source. Therefore we have added the following clarification to the revised manuscript:

"For this analysis we extended the existing hydrology archive to cover the UTRB domain for all 234 BCSD CMIP-5 climate projections following the steps detailed below. A subset of the CMIP-5 hydrology projections is publically accessible through the "Downscaled CMIP3 and CMIP5 Climate and Hydrology Projections" archive at http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/. Additional documentation on the archive and the methodology is provided in Reclamation [2014]" (Revised Manuscript, Page 10, Lines 4-9)

5) p. 5092, the lack of qualitative model validation appears again here, such as "the VIC simulated and observed floods are in close agreement and the discrepancy with the GEV model is explained by the flood timing described above." What constitutes 'close agreement' and at what point would they be considered not in agreement? And the discrepancy is not explained by the timing, but is apparently consistent with it, which is much weaker. Later instances in this section show things like "the GEV model is in good agreement with the VIC simulated flow", and ultimately "This demonstrates that the model behavior is a reasonable match to the natural system." These general observations are not helpful in determining significant correspondence of modeled and simulated values.

Response: We appreciate the comment and have made changes throughout section 4.2. to make our analysis more precise (please refer to the changes listed in comment 2). Also, in response to this comment we have replaced the phrase 'is explained by' with 'is consistent with'. We hope that the review will find our revised section 4.2 to be more clear.

6) p. 5094, Figure 6 is presented, which is interesting. another way to cast this would be in a manner similar to that of Mailhot and Duchesne (J. Wat. Res. Plann. Mgmt., 2010, doi 10.1061/_ASCE_WR.1943-5452.0000023) Figure 3, which aims to provide planners with a design return period for today that would be needed to provide protection at the level of a historic return period (in a stationary climate).

Response: We thank the reviewer for pointing out this paper. We focused our analysis on the risk to existing infrastructure. Hence we present our results in terms of the probability of exceeding the flow rate that was used in previous design. However we agree that Mailhot and Duchesne provide an interesting approach for taking non-stationarity into account with new design. We have added a reference to their work in the beginning of the Introduction as follows:

> ***"This discrepancy has not gone unnoticed within the scientific community and there is a growing body of research investigating, (1) trends in observed floods [e.g. Franks, 2002; Vogel et al., 2011], (2) ways to incorporate non-stationarity into frequency distributions [e.g. Katz and Neveau, 2002; Raff et al., 2005] and (3) methodologies to interpret risk and approach design within a non-stationary framework [e.g. Mailhot and Duchesne, 2010; Rootzen and Katz, 2013; Salas and Obeysekara, 2014]." (Revised Manuscript, Page 3, Lines 3-8)***

Minor comments:

SI units should be used throughout, not square miles, feet , etc.

Response: We have changed all of the units to SI.

p. 5088, line 16, is the 0.05 alpha? And what significance test is being referred to here?

Response: Yes the threshold is referring to the alpha. For the significance test we are using a chi-squared distribution. These points have been clarified in the revised manuscript as follows:

> ***"For this analysis the best model is selected using pairwise comparisons of NLLH scores following the methods of Salas and Obeysekera [2014] and others. Models are compared using the deviance statistic (D) which is equal to twice the difference in NLLH scores. Deviance statistics are then tested for significance based on a chi-squared distribution with the degrees of freedom set equal to the difference in the number of parameters (K) between models. P-values less than 0.05 indicate a statistically significant (alpha of 0.05) improvement in model performance." (Revised Manuscript, Page 12, Lines 19-25)***

References cited in Response:

Brekke, L.D., M.D. Dettinger, E.P. Maurer, and M. Anderson. 2008. "Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments," *Climatic Change*, 89(3–4), 371–394.

Gleckler, P.J., K.E. Taylor, and C. Doutriaux. 2008. "Performance metrics for climate models," *Journal of Geophysical Research* 113(D06104).

Mote, P.W., and E.P. Salathé. 2010. "Future climate in the Pacific Northwest," *Climatic Change*, DOI: 10.1007/s10584-010-9848.

Pierce, D.W., T.P. Barnett, B.D. Santer, and P.J. Gleckler. 2009. "Selecting global climate models for regional climate change studies," *Proc National Academy Sciences*, v106 (21), 8441–8446.

Reclamation. 2011. "West-Wide Climate Risk Assessments: Bias-Corrected and Spatially Downscaled Surface Water Projections," prepared by Bureau of Reclamation. U.S. Department of the Interior, Draft Technical Memorandum 86-68210-2011-01, March 2011, 138 pp.

Reichler, T., and J. Kim. 2008. "How Well Do Coupled Models Simulate Today's Climate?" *Bulletin of the American Meteorological Society*, 89(3) pp. 303–311.

Santer, B.D., K.E. Taylor, P.J. Gleckler, C. Bonfils, T.P. Barnett, D.W. Pierce, T.M.L. Wigley, C. Mears, F.J. Wentz, W. Bruggemann, N.P. Gillett, S.A. Klein, S. Solomon, P.A. Stott, and M.F. Wehner. 2009. Incorporating model quality information in climate change detection and attribution studies, *Proc National Academy Sciences*, v 106 (35), 14778–14783.

Tebaldi, C., R.L. Smith, D. Nychka, and L.O. Mearns. 2005. —Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multi-model Ensembles, *Journal of Climate*, 18 pp. 1524–1540.