The Impacts of Climatological Adjustment of Quantitative Precipitation Estimates on the Accuracy of Flash Flood Detection

Yu Zhang[a,*], Sean Reed[d], Jonathan J. Gourley[c], Brian Cosgrove[a], David Kitzmiller[a], Dong-Jun Seo[b], Robert Cifelli[e]

[a] *National Water Center, NOAA National Weather Service, Silver Spring, MD*
[b] *University of Texas at Arlington, Arlington, TX*
[c] *NOAA National Severe Storm Laboratory, Norman, OK*
[d] *Mid-Atlantic River Forecast Center, NOAA National Weather Service, State College, PA*
[e] *NOAA Earth System Research Laboratory, Boulder, CO*

Abstract

The multisensor Quantitative Precipitation Estimates (MQPEs) created by the US National Weather Service (NWS) are subject to a non-stationary bias. This paper quantifies the impacts of climatological adjustment of MQPEs alone, as well as the compound impacts of adjustment and model calibration, on the accuracy of simulated flood peak magnitude and that in detecting flood events. Our investigation is based on 19 watersheds in the mid-Atlantic region of US, which are grouped into small ($< 500km^2$) and large ($> 500km^2$) watersheds. NWS archival MQPEs over 1997-2013 for this region are adjusted to match concurrent gauge-based monthly precipitation accumulations. Then raw and adjusted MQPEs serve as inputs to the NWS distributed hydrologic model-threshold frequency framework (DHM-TF). Two experiments via DHM-TF are performed. The first one examines

*Corresponding author
Email address: yu.zhang@noaa.gov (Yu Zhang)

the impacts of adjustment alone through uncalibrated model simulations, whereas the second one focuses on the compound effects of adjustment and calibration on the detection of flood events. Uncalibrated model simulations show broad underestimation of flood peaks for small watersheds and overestimation those for large watersheds. Prior to calibration, adjustment alone tends to reduce the magnitude of simulated flood peaks for small and large basins alike, with 95% of all watersheds experienced decline over 2004-2013. A consequence is that a majority of small watersheds experience no improvement, or deterioration in bias (0% of basins experiencing improvement). By contrast, most (73%) of larger ones exhibit improved bias. Outcomes of the detection experiment show that the role of adjustment is not diminished by calibration for small watersheds, with only 25% of which exhibiting reduced bias after adjustment with calibrated parameters. Furthermore, it is shown that calibration is relatively effective in reducing false alarms (e.g., false alarm rate is down from 0.28 to 0.19 after calibration for small watersheds with calibrated parameters); but its impacts on detection rate are mixed. As an example, the detection rate of 2-Y events in fact declines for small watersheds after calibration is performed (from 0.4 to 0.28, and from 0.28 to 0.19 with raw and adjusted MQPE, respectively). These mixed outcomes underscore the complex interplays between errors in MQPEs, conditional bias in the reference gauge-based analysis, and structural deficiencies of the hydrologic model.

Keywords: flash flood, precipitation, hydrologic model, detection

1. Introduction

Accurate detection and prediction of flash floods are of great importance to reducing flood-related life losses and property damages, and yet these are also among the most challenging aspects of hydrologic prediction due to the short response nature of the flooding events (Sene, 2012). Since the advent of weather radar, near real-time radar-based and radar-gauge blended quantitative precipitation estimates (QPEs) have been routinely used for flash flood monitoring and prediction in the world(Cosgrove et al., 2012; Sene, 2012; Berne and Krajewski, 2013). In the United States, most of the warnings are issued based on coupling of high resolution QPEs and Quantitative Precipitation Forecast with Flash Flood Guidance(Gourley et al., 2012), while an emerging paradigm of distributed Model-Threshold Frequency (DHM-TF; Reed et al., 2007) has been gradually adopted. DHM-TF is based on a grid-based, distributed hydrologic model, and is therefore able to account for upstream inflow in calculating flood risk; it relies on historical streamflow simulations to define the thresholds for flooding and flood intensity levels, and thereby circumvents the difficulty in empirically establishing these thresholds at smaller reaches with no, or limited flow records. DHM-TF has been shown by Gourley et al. (2012) and Cosgrove et al. (2012) to outperform FFG in a number of experimental settings.

Note that since DHM-TF establishes the thresholds on the basis of flow simulations, it requires high-resolution, accurate *historical* QPEs in addition to real-time QPEs and reliable hydrologic model representations. Historical QPEs can be subject to a number of deficiencies. In the US, the widely used multisensor QPEs (MQPEs) based on blending radar and gauge observations are known to exhibit a time varying bias (Zhang et al., 2011a). This trending bias has clear implications for hydrologic prediction. Zhang et al. (2011a) demonstrated that the water balance based on uncalibrated runs of a distributed hydrologic model exhibits a conspicuous upward trend between 1998 and the early-mid 2000. Zhang et al. (2011a) further experimented with re-adjusting the MQPEs using monthly gauge-based precipitation analysis. Though the authors found that this adjustment greatly reduced the trending bias in simulated water balance, they also suggested that the adjustment may be detrimental to resolving the magnitude of rainfall and flood peaks.

Bias and inaccuracy of both real-time and climatological QPE products, and the associated impacts on flood and flash flood prediction have both been active research areas (Smith et al., 1996; Young et al., 1999, 2000; Hardegree et al., 2008; National Research Council, 2005; Oudin et al., 2006; Kitzmiller et al., 2011; Looper et al., 2012), so is calibration of hydrologic model (Duan et al., 1993; Gupta et al., 1998; Winsemius et al., 2009; Westerberg et al., 2011; Singh and Bàrdossy, 2012). Yet, to date, few studies have addressed the linkage between climatological adjustments and the accuracy of flash flood detection and prediction, though a few did examine the impacts of uncertainties in forcings and parameters. Oudin et al. (2006), for example, illustrated that some of the impacts of random and systematic errors in precipitation can be compensated by model calibration. The authors, however, did not explore climatological adjustment as a means to suppress the random and systematic errors. Zhang et al. (2011a)'s analysis on climatological adjustment focused on simulated water balance rather than on detection of flash flood events, and the authors did not address the relative effects of model calibration and adjustment. Strauch et al. (2012) attempted to account for the uncertainty in precipitation and parameters simultaneously by calibrating the model against an ensemble of precipitation inputs. Looper et al. (2012) assessed the compound effects of adjustment and model calibration. Neither of the latter two studies, however, delve into the mechanistic causes of precipitation errors and bias, nor did they address the impacts of calibration and adjustment on flood detection per se. The present study is intended to fill this gap by investigating isolated and compound impacts of climatological adjustment, both prior to and after model calibration, on the detection of flash floods over 19 watersheds in the eastern US. In this work, a long-term radar-gauge MQPE data set is adjusted using monthly gauge-based analysis, and both the original and adjusted MQPEs serve as inputs for calibrating a distributed hydrologic model. The streamflow simulation series from model with *a priori* and calibrated parameters are then used as the basis of the detection experiment. The work also complements a body of literature attempting to disentangle the impacts of structural and input errors on uncertainty in model prediction (e.g., Renard et al., 2010; Sun and Bertrand-Krajewski, 2013) by examining the differential impacts of calibration in the presence of non-stationary rainfall bias.

The remainder of the paper is organized as follows. Section 2 describes the data and methods. Section 3 summarizes the observations. Section 4 discusses the results, and Section 5 summarizes the key conclusions.

2. Data and Methodology

2.1. Study watersheds

Selected for this study are 19 watersheds located within the service area of Mid-Atlantic River Forecast Center (Fig. 1; Table 1), whose drainage areas range from 84 to 2116 km^2. These watersheds are divided into two groups: a) *small* watersheds - those with drainage area below 500 km^2 and b) *large* watersheds, with drainage area above 500 km^2. The threshold of 500 km^2 was chosen as it roughly divides the watersheds with short response time and therefore prone to flash floods from those of much longer response time: synthetic unit hydrographs generated using a distributed hydrologic model (to be described later) indicate that all except one (**WASHB**) small basins in the former group are associated with time to peak (T_p) less than 6 hours, whereas only one in the latter group does. The large watersheds are included in the analysis, as short-fused floods can also take place with an opportune combination of the spatio-temporal configuration of storm systems and antecedent soil moisture conditions (Zhang et al., 2003).

For each basin, flood events were identified from the hourly time series collected by the United States Geological Survey (USGS) using the 2-Y Averaged Recurrence Interval (ARI) values as thresholds; the former of these is widely considered a rough indicator of the over-bank flow (Reed et al., 2007). In this study, these ARI values are established based on the annual maximum hourly peak discharge using the standard procedure outlined in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982; Reed et al., 2007). For years where annual peaks were underrepresented due to missing observations, estimates of instantaneous peak discharge rate from USGS are used instead. Flood producing mechanisms vary depending on watershed size and location. Smaller watersheds are more susceptible to flooding driven by summertime convective systems (Zhang et al., 2001), whereas a substantial number of major floods in both groups of watersheds were due to tropical and extratropical cyclones. Snowmelt and earlier spring frontal systems are potent flooding drivers for large but rarely for small watersheds. In this study, the focus is given to only events between April and October to avoid the complications of snow-melt events where flood response may be driven jointly by temperature and precipitation.

2.2. Multisensor Precipitation Estimates

The primary forcing for this study is the National Weather Service (NWS) Multisensor QPE (MQPE) products retrieved from the Mid-Atlantic River Forecast Center (MARFC) for 1997 to 2013. These products were created by blending radar-only QPE from the NEXRAD Precipitation Processing System (PPS, Fulton et al., 1998) and gauge reports. The products over the earlier (1997-2001) and later (2001-2013) periods were created using the Stage III and the Multisensor Precipitation Estimator (MPE) package, respectively (Seo et al., 2011; Zhang et al., 2011a). The MPE multisensor blending algorithm is similar to that of Delrieu et al. (2014). Since 2000, several River Forecasting Centers (RFCs) started ingesting 24-h accumulations from Cooperative Observer (COOP) gauge reports into MPE, either by inserting disaggregated COOP reports into MPE or by adjusting the 24-h MQPE accumulations to match the COOP reports.

A number of studies have pointed to a negative bias in the earlier Stage III and MPE products, i.e., precipitation amounts based on these products are systematically lower than corresponding gauge observations. This bias can be attributed in part to the presence of a truncation error (TE, Fulton et al., 2003) in the earlier version of the NEXRAD PPS. Zhang et al. (2011a) showed that the bias gradually improved between the late 1990's and early 2000's, most likely due to a combination of later-day ingest of COOP station reports, expanded gauge data set, better quality assurance, and the correction of the TE. Zhang et al. (2011a) also demonstrated that this time-varying bias can be alleviated by post-adjustment using the PRISM monthly product.

As in Zhang et al. (2011a), the MARFC MQPEs underwent PRISM-based post-adjustments that essentially revised the MQPE hourly amount at each Hydrologic Rainfall Analysis Project (HRAP; Reed and Maidment, 1999) pixel by a constant multiplicative factor so that the monthly accumulation for that pixel matches that of PRISM. This method, despite its simplicity, has been shown to substantially improve the negative bias in streamflow simulations.

2.3. Hydrologic Model and Simulation Experiments

This study employs the NWS Research Distributed Hydrologic Model (RDHM; Smith et al., 2012), a flexible modeling system that consists of a number of modules for simulating a full range of hydrologic processes. Key ingredients of RDHM include the Sacramento Soil Moisture Accounting (SAC-SMA, Burnash, 1995) for water balance and runoff computation, SNOW-17 for estimating snowmelt and ablation, and the 1-D kinematic wave routing module. Fig.2 shows a sketch of the SAC-SMA framework with model states. In brief, SAC-SMA divides a soil column into a thin upper zone and a thicker lower zone. Water in each zone is partitioned into *free water* that drains by gravity and *tension water* held by capillary head of soil matrix. The free water storage of the lower zone is further subdivided into *supplemental* and *primary storages*, corresponding to faster and slower draining groundwater flows, respectively. *Percolation* is allowed from the upper to the lower zone, and its rate is controlled by parameter ZPERC. Both the lower zone primary and supplemental storages contribute to baseflow, and the rate of depletion associated with each storage is controlled by parameters LZPK and LZSK, respectively. Upper zone free water contributes to interflow, whose rate is determined partially by a parameter UZK.

For this study, RDHM was implemented on an approximately 2km grid mesh with all the aforementioned modules incorporated. Each module requires an initial set of parameters, or *a priori* parameters. The *a priori* parameters for SAC-SMA were derived based on Nature Resources Conservation Service Soil Survey Geographic (SSURGO) Database (Anderson et al., 2006; Zhang et al., 2011b, 2012), and National Land Cover Dataset (NLCD). The SNOW-17 parameters were computed based on physiographic grid data sets and climatic wind data. The routing parameters were derived from USGS cross-section survey and discharge measurements. The parameters to be calibrated comprise of 9 SAC-SMA parameters and two routing parameters (Table 2).

In addition to precipitation, RDHM requires temperature and potential evapotranspiration (PET) as forcings. This study uses 6-h gridded surface temperature from NCEP reanalysis, and monthly climatic PET that is invariant across years; the latter is first disaggregated onto daily scale by linearly interpolating the values assuming that each month value belongs to 16th day of the month, and the daily values are then equally divided among the 24 hours to produce hourly PET values.

The study comprises two sets of simulations experiments. The first set relies solely on uncalibrated model runs, and the foci are on the effects of adjustment on the accuracy of annual flood peaks. The second set involves split-sample calibration-validation experiments intended to illuminate the joint impacts of readjustment and model calibration on the accuracy of flood frequency distribution and flood detection. Layouts of the experiments are summarized below.

Uncalibrated Model Runs

The first experiment relies on the uncalibrated RDHM (with *a priori* parameters) run over the entire period (1997-2013) using a) raw and b) adjusted MQPEs as forcing. To reduce the influence of uncertain initial conditions, the first year (1997) is treated as the spin-up period to reduce the errors due to uncertain initial conditions, and the associated simulations are not used in subsequent evaluations. Following the approach of DHM-TF, the simulated hourly streamflow for the remaining period is postprocessed to yield the annual maximum series, which is then used to construct flood frequency distribution (FFD) via the Bulletin-17B procedure. The FFDs based on simulations using raw and adjusted MQPEs are then compared with those based on concurrent streamflow observations to gauge the impacts of adjustment on the accuracy of FF. Subsequently, the estimated flood peaks corresponding to the ARI of 2 years are used to delineate the flooding events.

Calibrated Model Runs

The second experiment is a calibration-validation experiment in which the entire period is split into the calibration (1998-2007) and validation (2008-2013) sub-periods. Figs. 3a and b illustrate the time periods and process involved in the uncalibrated and calibrated simulations. Calibration involves adjusting 11 parameters using the RDHM automated calibration module that implements the sequential line search (SLS) algorithm (Kuzmin et al., 2008). SLS is a local searching algorithm that has been shown by Kuzmin et al. (2008) to be more efficient, and sometimes as robust as the Shuffled Complex Evolution (SCE; Duan et al., 1993), a global searching algorithm.

The value of each parameter is adjusted in a spatially uniform fashion using a scalar multiplier whose initial value is set to unity. SLS seeks to minimize the so-called multi-scale objective function (MSOF) by incrementing a particular element of the vector of scalar multipliers at a time until a minimum MSOF is attained. The MSOF is a composite metric that weighs errors at different temporal resolutions. Its formal definition is given below:

$$MSOF = \left(\sum_{k=1}^{n} \frac{\sigma_1^2}{\sigma_k^2} \sum_{i=1}^{m_k} \left(q_{o,k,i} - q_{s,k,i}(X) \right)^2 \right)^{1/2} \tag{1}$$

where n is the number of time scales, σ_1 and σ_k are the standard error at the base time resolution (normally hourly), and resolution k. $q_{0,k,i}$ and $q_{s,k,i}$ are observed and simulated discharge at time interval i and resolution k, respectively. In this study, three time resolutions, i.e., hourly, 24-hourly and 240-hourly are used.

After the model is calibrated for each basin, the 2-Y ARI values are again calculated from the simulated streamflow. Then, the calibrated model is run for the entire 16-year period, and the ARI values determined over the *calibration period* are used as thresholds to detect flood events. As this study focuses on linkage of precipitation and flood events, we chose to sample only warm season (April through October) flood events so to avoid the complications surrounding the interpretation of events driven by snowmelt.

The detection method is summarized as follows. For each basin, a collection of windows with *observed flow* exceeding a threshold, i.e., the 2-Y ARI, are first established. Then, simulated discharge over each of these windows is extracted. If simulated discharge for a window exceeds the corresponding threshold established using simulations, a successful detection is declared for the event. False alarms are calculated in a parallel way, with the events defined using simulated discharge. A false alarm is declared when the observed discharge does not exceed the prescribed threshold whereas the simulated discharge does. The accuracy of model simulations is gauged by probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and ranked correlation (Kendall's Tau). Let X_i denote the number of flooding events successfully detected for basin i, Y_i the number of flooding events that occurred but were not detected, and Z_i the number of false alarms (events reported by model but not present in observed series). The POD, FAR and CSI for basin i are given below:

$$POD_i = \frac{X_i}{X_i + Y_i} \tag{2}$$

$$FAR_i = \frac{Z_i}{X_i + Z_i} \tag{3}$$

$$CSI_i = \frac{X_i}{X_i + Y_i + Z_i} \tag{4}$$

As the number of flooding events can be limited given the short time period, we also use the multi-basin aggregate POD, FAR and CSI, hereinafter denoted by $\overline{POD}$,$\overline{FAR}$,$\overline{CSI}$. Each quantity is derived by aggregating all the flooding events over each basin of a given group. For example $\overline{CSI}$ is defined as:

$$\overline{CSI} = \frac{\sum_i X_i}{\sum_i (X_i + Y_i + Z_i)} \tag{5}$$

The definition of Kendall's Tau is given below:

$$Tau = \frac{N_c - N_d}{1/2n(n-1)} \tag{6}$$

where N_c and N_d are the number of concordant and discordant pairs, respectively.

Note that, although PRISM climatology is unlikely to be available at real-time for bias-adjustment, this experiment helps gauge the relative merit of forgoing the spatial details brought by radar and relying on the latter exclusively as a tool of disaggregating daily gauge data (e.g., the forcing from the North America Land Assimilation System; Cosgrove et al., 2003).

3. Results

This section first presents the impacts of adjustment on hourly mean areal precipitation. Then, the results of uncalibrated model simulations will be summarized, with attention given to the comparative accuracy of annual peak statistics based on simulated streamflow before and after climatological adjustments, and the associated accuracy of detection over the period of 1998-2013. The second subsection explores the compound impacts of climatological adjustment and model calibration on the accuracy of detection through the calibration-validating experiment.

3.1. Outcome of Precipitation Adjustment

For each basin, the ratio of mean areal precipitation (MAP) after and prior to adjustment was computed for each month between 1997 and 2013. The monthly series of multi-basin mean of this ratio are shown in Figs. 4a and b, for the small and large basin groups, respectively. For both groups, a downward progression in the ratio is evident; the adjustment factor is overwhelmingly positive for the pre-TE correction period; it progressively declines toward neutral around the time when TE was corrected (Dec. 2003), and becomes mostly negative onwards. To assess the significance of these trends, Mann-Kendall(MK)'s test (Mann, 1945; Kendall, 1975) was applied to the ratio time series of the pre-TE period. MK test is a non-parametric test that is based on comparing pairs of data points in a time series and tracking the number of increases, decrease and ties. It yields the statistic S that varies in [-1,1], with -1/1 indicates that the series exhibits perfect monotonic downward/upward trend. MK test for the series yields S of -0.377/-0.412 for the small/large basins. The associated P value are well below 0.05 (4 x 10^{-7} and 3 x 10^{-8}, for small and large basin groups, respectively). This confirms that the trends are statistically significant. For the post-TE period, minor declines are observed but the trend is not statistically significant for either group (P value beyond 0.05). The downward trends over the earlier period are unsurprising: Zhang et al. (2011a) pointed out that the negative bias in the hourly Stage III and MPE product as induced by TE gradually diminished due to a combination of increased number of real-time gauge used in MPE and the introduction of manual quality using daily cooperative observation (COOP) network.

The net effects of adjustment on moderate-heavy precipitation are characterized by the 99% quantiles of positive MAP (Fig. 5). For the earlier period, adjustment has a clear tendency to elevate the 99% quantile for all small basins and a majority of large basins (Fig. 5a). though the differences are slightly less conspicuous for the latter. For the post-TE correction era (Fig. 5b), adjustment still exhibits a slight tendency to increase the 99% quantile, though the differences are rather minor. The increase in the pre-TE period is consistent with the earlier observation of prevailing positivity of adjustment factors, which, as discussed earlier, is the consequence of the negative bias of the earlier era (Fig. 4). For the later period, the impacts of adjustment on moderate-heavy precipitation range from being neutral to slightly positive.

3.2. Uncalibrated Model Runs

Fig. 6 shows the long-term adjustment factor for MAP and the bias ratio in cumulative runoff for each basin using raw and adjusted MQPE over the entire period. The adjustment factor is the ratio of multi-year total MAP from PRISM to that based on MQPE, and the bias ratio is the ratio of cumulative simulated streamflow to the observed value. The adjustment factor is positive for a majority of watersheds (i.e., bias ratio above unity; Fig. 6a), where runoff bias using raw MQPEs is negatively biased (i.e., bias ratio below unity; Fig. 6b). Runoff bias is much improved for most of the watersheds, when the model is forced by adjusted MQPEs, though it remains overall negative. Variations among basins tend to be large, but no clear distinctions are seen between the small and large basins.

The median annual peak discharge from the two sets of simulations is shown in Fig. 7 along with the ratio to observed values. To discern the impact of the earlier bias in MQPE, the medians were computed both using the entire length of data (1998-2013; Figs. 7a and b) and using only the post-TE period (2004-2013; Figs. 7a and b). Table 4 summarizes the percentage of events where simulated median annual peaks, and percentage bias have declined after adjustment for the entire period (1998-2013) and for the post-TE period (2004-2013), where percentage bias is defined as the difference between simulated and observed discharge scaled by the latter, i.e., $100(1 - Q_{sim}/Q_{obs})$. Notable observations are summarized below.

First, as shown in Figs. 7a and b, bias in annual peak is strongly dependent on the size of drainage: all small watersheds exhibit a negative bias in the simulated median annual peaks, whereas bias is positive for a majority of larger ones. Second, when the entire period is concerned, adjustment tends to suppress simulated median peaks for large watersheds, while its impacts on small watersheds are mixed (Figs. 7a and b; Table 4). Decline in median peak is observed in 82% (9 out of 11) of large watersheds, but only in 38% (3 out of 8) small watersheds. The magnitude of the reduction is quite conspicuous for several larger watersheds. Adjustment in general helps mitigate the positive percentage bias in median annual peaks for the large watersheds, with 82% exhibiting reduction. Its impacts, however, are again mixed for the small watersheds, with 38% of them exhibiting reduction in percentage bias (Table 4. Note that the overall suppression of peaks contrasts with, but does not contradict, the increased and unchanged quantiles of heavy precipitation shown in Fig. 5. It will be shown in the later portion of the paper that adjustment indeed reduced the monthly MAP for a majority of months where flood occurred despite the fact it in general increased the quantiles of heavy precipitation.

For the period following TE-correction (Table 4), the most prominent feature is perhaps the overwhelming reduction in the median peaks: all but one watersheds show reduced value after adjustment, with the median of reduction nearly 30%. For the small watersheds, bias in fact turns worse after adjustment, with only 25% of watersheds showing reduction in percentage bias (Table 4). Similarly, only a minority of larger watersheds experienced decline in percentage bias (36%, or 4 out of 11). Though post-TE era MQPE appears to be bias-neutral relative to PRISM (Fig. 4), there is a tendency for adjustment to reduce median annual peaks for small and large watersheds alike over this period.

The contrasting bias in the simulated annual peaks for small and large watersheds may be due to a combination of factors. It is plausible that the positive and negative model biases are reflecting differing structural and parametric deficiencies of models at different watershed scales. Meanwhile, the fact that adjustment greatly reduced the positive bias in the simulated median annual peaks for several large watersheds can be an indication that MQPE tends to overrepresent the rainfall amounts of flood-producing storms.

3.3. Calibrated Model Runs and Detection Experiments

Model calibration over 1998-2007 using raw and adjusted MQPEs yielded two sets of scalar multipliers. Table 3 summarizes the multi-basin means of calibrated scalar multiplier for each parameter. Since calibration was done individually using the raw and adjusted MQPE as input, there are two sets of multiplier values, and these are further stratified by small and large basins. Note that the differences between the resultant multipliers using raw and adjusted MQPEs are relatively minor: the largest difference is observed in in LZSK (depletion rate of lower zone supplemental water storage), and ZPERC (shape parameter of the percolation curve). The multipliers for small and large basins contrast sharply. For example, calibration slightly reduces ZPERC for small watersheds, whereas it increases ZPERC for large watersheds, regardless of whether adjustment is performed. Lower ZPERC implies reduced percolation rate and increase in faster runoff originating from the upper zone. This is consistent with the need of compensating for the negative bias in peak discharge for small watersheds and positive bias for larger ones. Similarly, small watersheds exhibit increases in routing parameter QMCHN whereas large ones exhibit declines. As higher QMCHN leads to accelerated flood peaks and magnified peak magnitude, this contrasting outcome is again a result of the differing bias behaviors of uncalibrated model for larger and smaller basins.

Each parameter set is subsequently used to generate streamflow simulations for 2008-2013. As in the uncalibrated run, the annual peaks based on the calibrated model simulation for 1998-2007 were used to establish the FFDs. The 2-Y quantiles based on these FFDs then serve as threshold in the detection experiment. To simplify descriptions, each of the four groups of simulation results is assigned a unique label: a) uncalibrated model simulations with raw MQPE - UX; b) uncalibrated model simulations with adjusted MQPE - UA; c) simulations with raw MQPE using model calibrated with raw MQPE - CX; and d) simulations with adjusted MQPE using model calibrated with adjusted MQPE - CA.

Fig. 8 compares the median annual peaks from CX and CA versus those based on observations for both the entire period (1998-2013) and the post-TE era (2004-2013). Table. 4 provides the percentage of watersheds showing reduction in median peaks and those showing improved bias with adjustment. The most notable observation in Fig. 8a and b is that the contrasting bias behavior of small and large basins, i.e., negative/positive bias for the small/large, has diminished after calibration. Calibration did not, however, entirely eliminated the bias - bias appears to be consistently, albeit slightly, negative for a majority of small and large basins alike. The impacts of adjustment are not visually conspicuous, but for a majority of watersheds the median peaks show decline, and fewer watersheds experience reduction in percentage bias in comparison to the uncalibrated case (Table. 4). Features for the later period (2004-2013) are largely similar, except that slightly more watersheds experienced decline in median peaks.

To assess the effects of model calibration on the FFD, the multi-basin averages of Log Pearson type III (LP3) parameters derived from each simulation group are used to construct the "representative" FFDs for that basin group. These are compared with observation-based ones in Figs. 9. For the small watersheds (Fig. 9a), FFDs from all four groups of simulations are below that based on observations. Among these, FFDs from uncalibrated model runs (UX and UA) show consistent underestimation of quantiles at short ARI. At longer ARI, the UX-based FFD in fact shows the closest resemblance to the observed whereas UA-based curve is much flatter and well below the observed. Calibration helps mitigate this underestimation only at shorter ARI (below 5-Y). At longer ARI, it in fact worsens the quantiles based on unadjusted MQPEs. For the large basins(Fig. 9b), quantiles from uncalibrated model runs are appreciably higher than the observed though those from UA are broadly lower, pointing to beneficial impacts of adjustment. Calibration reduces the quantiles but introduces a negative bias at longer ARI. Among the four groups, CX offers the closest approximation of the curve at longer ARI, though it suffers a negative bias throughout ARIs.

The individual and compound impacts of calibration and MQPE adjustments on the detection of flood events (i.e., events with peaks exceeding 2-Y ARI), are assessed on an multi-basin aggregate basis using aggregate POD, FAR CSI, and Tau in Figs. 10, and 11, for small and large watersheds, respectively. For the calibration period, a total of 50 events were identified in the observed flow series for small and large basins. For the validation period, the corresponding numbers are 39 and 47. For the small basins (Fig. 10), the following observations are evident. First, the impacts of adjustment can be beneficial or detrimental depending on the metrics and evaluation period. For the calibration period, adjustment alone leads to improved POD, FAR, and CSI (Fig. 10a, c and e), whereas for the validation period, it in fact reduces POD and CSI (Fig. 10b and f). Calibration, curiously, slightly worsens POD, FAR, or CSI over the calibration period (Fig. 10a, c and e), though Tau values are much improved (Fig. 10g). For the validation period, the gap in metrics related to adjustment widens slightly after calibration (Fig. 10b,d, f and h). For example, the deterioration in the composite measure CSI becomes more pronounced after calibration (Fig. 10f).

For the large basins, a distinct feature is that adjustment has clearly positive impacts on the evaluation statistics for both periods when the model is calibrated (Fig. 11a-h). By contrast, with uncalibrated model parameters, POD and CSI decline slightly after adjustment (Fig. 11a,b, e and f). Similar to small basins, the impacts of calibration are quite positive on Tau, but are muted to slightly negative on POD, FAR and CSI.

The incremental impacts of calibration vary widely among watersheds. Table 6 summarizes the *net percentage* of basins exhibiting improvements after adjustment before and after calibration for the *validation period*, where *net percentage* is defined as the difference between the percentage of basins showing improvements and that experiencing deterioration. At 2-Y ARI threshold level, it is evident that for both uncalibrated and calibrated simulations, a majority of small watersheds, and a slight minority of large watersheds exhibit deterioration in POD observed after adjustment. By contrast, a minority of small watersheds show reduction in false alarms in response to the adjustment, whereas a small majority of large watersheds do. To further quantify the impacts of adjustment, a one-side Mann-Whitney test is performed on the POD and FAR from pairs of unadjusted and adjusted results (i.e., UX vs. UA, and CX vs. CA), with the alternative hypotheses that adjustment worsens the POD and FAR. Prior to calibration, the reduction in POD and FAR after adjustment for small basins are deemed statistically insignificant (P=0.12, 0.38). After calibration, by contrast, the corresponding P values are at 0.03 and 0.02, respectively, indicating that the deterioration/improvement in POD and FAR due to adjustment in fact become statistically significant. For larger basins, changes in POD and FAR as induced by adjustment are statistically insignificant both before and after calibration.

3.4. Case Study

To explain the slight amplification of the impacts of adjustment following calibration, we examine the individual flood peaks over the small watershed **ROCKS** based on the simulations. **ROCKS** exhibits deterioration in POD and CSI with adjustment both before and after model calibration (Fig. 12). It is clear from Fig. 12 that calibration using adjusted MQPE led to much more dramatic increases in simulated peaks for the calibration period. Yet, the corresponding increase in the 2-Y quantile was even larger. As a consequence, three floods detected prior to calibration dropped below the elevated threshold. It is not immediately clear why calibration using adjusted, rather than raw MQPEs, yielded an increase in threshold. Our comparison of the calibrated parameters for **ROCKS** indicates that, in the earlier case, searching algorithm yielded a parameter combination that would allow the simulated peak to closely mimic the observed one for the largest event in the calibration period (25 June 2006), whereas it did not when raw MQPEs were used.

4. Discussions

Adjustment of radar and multisensor QPEs based on long-term gauge-based climatological products has been shown to mitigate the non-stationary bias in MQPE and therefore benefit streamflow simulations. Our analyses, however, suggest the impacts of adjustment on flash flood detection are complex and variable depending on watershed size. The remainder of this section summarizes, and attempts to interpret, the scale-dependent impacts of adjustment.

4.1. Impacts of Adjustments and Their Dependence on Drainage Size

Prior to model calibration, the PRISM-based adjustment itself has a clear tendency to reduce simulated annual discharge peaks for small and large watersheds alike. For the small watersheds, the net impacts are a degradation of accuracy, whereas for the large ones, this reduction actually leads to improved accuracy. This contrast can be explained by the contrasting bias behavior of uncalibrated RDHM in simulating flood peaks for the two groups of watersheds, i.e., underestimation for the former and overestimation for the latter. Reduction of peak, as a consequence of adjustment, worsens the negative bias in the small watersheds but mitigates the positive bias in the larger ones. The question, however, is whether the contrasting outcomes for the two groups of watersheds are in fact reflective of inherent deficiencies in model and parameterization, or those in the precipitation input? Our view is that both factors contribute to the phenomenon, but their relative roles differ.

The contrasting predispositions of the uncalibrated model for small and large watersheds are puzzling. As neither adjustment factor nor streamflow bias exhibit any clear dependence on drainage size, deficiencies in the rainfall-runoff and routing modules of RDHM in either, or both groups of watersheds emerge as the most plausible cause. Despite the advances in development of physically-based *a priori* parameter sets, biases and errors in model simulations may remain large (see e.g., Reed et al., 2004 and Smith et al., 2012). As most of the small watersheds chosen for this study are situated in suburban/urban areas, the flood peaks could be magnified by mechanisms operating at small spatial scale that are not well represented by the model. For example, stormwater runoff could be accelerated through paved surface, and flood peak could be magnified by surcharged sewer (see related discussion in Schmitt et al., 2004). While RDHM does integrate representation of connected impervious areas within each pixel, it is, as shown by our results, hardly adequate in capturing the complexity of these processes. For larger watersheds, there is a possibility of increased role of attenuation due to overbank storage (Woltemade and Potter, 1994).

While model deficiencies may be a key contributor to the observed small-large basin contrasts, roles of precipitation bias can not be completely ruled out. A notable observation for the larger watersheds is that PRISM-based adjustment substantially reduced the bias ratio of median peaks. This could be *prima facie* evidence that MQPEs were indeed biased in a consistent manner (positive bias) for heavier events. The question is, if MQPEs were positively biased, why adjustment led to deterioration of results over mostly small, rather than large, basins? There are two possible explanations. First, as mentioned above, while reduction brought by adjustment helped improve the accuracy of precipitation amounts, it exacerbated the bias in simulated peaks given the backdrop of preexisting, endogenous negative model biases for the small basins. Second, PRISM itself may suffer from negative bias, and the reduction per adjustment was therefore overdone for a significant number of events. Seo et al. (2014) analyzed the gauge-interpolated rainfall fields based on simple Kriging, and found that such fields tend to be slightly positively biased for lighter rainfall but negatively biased for heavier rainfall. Such magnitude-dependent bias, or *conditional bias*, may be a key element underlying the aforementioned negative bias.

To explore possible presence of conditional bias in PRISM-based precipitation accumulation, we plot the monthly adjustment factor against the MAP for each summer month by lumping all watersheds for each group (Fig.13). For each group of watersheds, the adjustment factor exhibit a conspicuous declining tendency with increasing monthly MAP that is statistically significant, with Mann-Kendall's test yielding P values well below 0.05. For drier months, adjustment factor is overall positive, whereas it is becomes slightly negative for the wettest months. These downward trends are consistent with the observations of Seo et al. (2014) on gauge-interpolated rainfall fields, namely that such fields may suffer a slight positive conditional bias for lighter precipitation and a negative one for heavier events. As most of the floods occur during the months with substantial accumulation (Fig.13), the net effect of adjustment is therefore a reduction of simulated flood peaks.

4.2. Interplays between Calibration and Adjustment

Perhaps the most important practical lesson from this work is that calibration does not diminish the impacts of precipitation adjustment. This effect is more conspicuous for small watersheds, where calibration slightly accentuates the outperformance of model with raw MQPEs. For larger watersheds, the limited improvement associated with adjustment remains after model calibration.

In theory, adjustment improves the consistency in the bias of MQPE over time, and therefore should have helped enhance the detection of flooding events, especially when the model is calibrated. Our experiments demonstrate that the opposite is true for small watersheds - adjustment slightly worsened the detection rates and CSI, and calibration in fact slightly amplified this detrimental impact. To explain this dilemma, we zoom in each watershed and compare the discharge peaks for each flood event from the four simulation groups and associated thresholds. It turns out that, for each watershed where POD deteriorated after adjustment, the 2-Y quantile experienced an increase, regardless of whether the model was calibrated. This is hardly surprising, as a substantial portion of the calibration period (1998-2007) lies in the era (1998-2003) when TE was present and induced a negative bias on precipitation. For the same basins, simulated peaks based on both uncalibrated and calibrated model in general declined after adjustment - 34 and 36 of the 46 peaks experienced decline for uncalibrated and calibrated simulations. This combination of declining peaks and increased threshold caused the detection rates to drop. For the larger watersheds, there were roughly equal numbers of events experiencing increase and reduction in peaks. As a result, though adjustment caused thresholds to increase, the effects were rather muted.

The slight amplification of the impacts of adjustment following calibration has to do with the differential change in the threshold after calibration. In general, calibration tends to increase/reduce both the 2-Y quantiles and simulated peaks over the validation period for small/large watersheds. In several watersheds, the magnitude of increases in the 2-Y quantile exceeded that in simulated peaks over the validation period, causing several flooding events to be left out after adjustment. This phenomenon is conceivable: our calibration relied on SLS, a local searching algorithm that can be trapped in a local minimum (Kuzmin et al., 2008). Apparently, adjustment in precipitation was sufficiently large to induce a substantial shift to search path and the resultant optimal parameter set. To fully understand parametric uncertainty and how it influences the perceived role of model adjustment, more sophisticated, global searching mechanisms, such as the Shuffled Complex Evolution Metropolis (SCME, Vrugt et al., 2003) and the Differential Evolution Adaptive Metropolis (DREAM Vrugt et al., 2009), will be needed. Such undertakings will be left for future endeavors.

5. Concluding Remarks

A basic assumption behind the DHM-TF is that simulated discharge peaks will be biased consistently, if not equally, in time. Yet, as our study demonstrates, nonstationarity in precipitation bias is a reality and it complicates the effective discharge threshold from historical simulations. Though adjustment using gauge-based climatological records helped improve the consistency in flow simulations (Zhang et al., 2011a), its impacts on simulated flood peaks and flood detection are mixed. Our analyses pointed to a conspicuous decline in simulated flood peaks after adjustment for a large majority (95%) of watersheds. The median of reduction for median annual peak is about 30%.

This study further shows that adjustment could even lower the detection of flood events, particularly over small, fast-responding watersheds that are prone to flash floods. Prior to calibration, POD declines from 0.56 to 0.46 after adjustment. After calibration, by contrast, 75% of watersheds showed decline, and the POD declines to 0.41. Owing to the limited duration of the experiments (17 years in total) and the number of watersheds involved (8 small watersheds and 11 larger ones; with 86 flood events in total for the validation period), it is premature to write off climatological adjustment as a useful ingredient in future DHM-TF-based flash flood prediction system. Nevertheless, the results are clear enough to warrant cautions against a wholesale adoption of the adjustment approach. The conditional bias in rain gauge based representation of the fields need be better understood and modeled, so do the biases of radar estimates over heavy events. Renalysis efforts, such as one ongoing at National Severe Storm Laboratory and National Climatic Data Center, would be helpful in this respect.

To conclude, it is clear from the study that accurate precipitation forcing, proper model structure, and robust parameter combinations are all requisites for DHM-TF to be effective. Calibration, while being able to broadly improve the model performance, is no substitute for improvements in forcing data, and its outcomes can be constrained by initial parameter selections. To improve the robustness of the prediction framework, it is critical to a) further understand the mechanisms underlying the intensity-dependence of adjustment factors, and explore the efficacy of alternative data sources and fusion methods in reconstructing heavy rainfall fields; b) enhance the efficiency of calibration and formulate objective functions that would allow accuracy in flood peak representation to play a more prominent role; and c) explore the sources of model mechanistic deficiencies and devise more robust parameterization scheme to mitigate persistent simulation bias in small domains across geographic settings. In addition, as demonstrated in this study, FFDs constructed using simulations could depart considerably from observed ones, and both calibration and adjustment could widen the departures. Further research will be needed to understand the implications of these departures for detecting and assessing the relative magnitude of extreme floods (i.e., with ARI greater than 50 years). With increasing computational poweress, probabilistic simulations using an ensemble of parameters estimated using strategies such as SCME and DREAM, could become a practical mechanism to account for the compound uncertainty of forcings and parameters.

6. Acknowledgment

This work was in part supported by the NOAA US Weather Research Program. We would like to thank Zhengtao Cui for his generous support and guidance on the implementation of RDHM. We also would like to acknowledge David Welch at LMRFC, Reggina Cabrera at SERFC for valuable insights and suggestions. Two anonymous reviewers provided detailed comments which help greatly improve the quality of the paper, and their contributions are duly noted here.

Acronyms

ARI:	Averaged Recurrence Interval
CSI:	Critical Success Index
DHM-TF:	Distributed Hydrologic Model - Threshold Frequency
FAR:	False Alarm Ratio
FFD:	Flood Frequency Distribution
GPM:	Global Precipitation Measurement
LP3:	Log Pearson type III
MAP:	Mean Areal Precipitation
MPE:	Multisensor Precipitation Estimator
MQPE:	Multisensor Quantitative Precipitation Estimate
NWS:	National Weather Service
POD:	Probability of Detection
PPS:	Precipitation Processing System
PRISM:	Parameter-elevation Regressions on Independent Slopes Model
QPE:	Quantitative Precipitation Estimate
QPF:	Quantitative Precipitation Forecast
RDHM:	Research Distributed Hydrologic Model
TE:	Truncation Error

References

R. M. Anderson, V.I. Koren, and S.M. Reed. Using SSURGO data to improve Sacramento model *a priori* parameter estimates. *J. Hydrology*, 320:103–116, 2006.

A. Berne and W. F. Krajewski. Radar for hydrology: Unfulfilled promise or unrecognized potential? *Adv. in Water Res.*, 51:357–366, 2013.

R. J. C. Burnash. The NWS river forecast system – catchment modeling. In V. P. Singh, editor, *Computer Models of Watershed Hydrology*, pages 311–366. Water Resources Publications, Littleton, Colorado, 1995.

B. A. Cosgrove, D. Lohmann, K. E. Mitchell, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, C. Marshall, J. Sheffiel, Q. Duan, L. Luo, R. W. Higgins, R. T. Pinker, J. D. Tarpley, and J. Meng. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, 108(D22), 2003.

B. A. Cosgrove, E. Clark, S. Reed, V. Koren, Z. Zhang, Z. Cui, and M. Smith. Overview and initial evaluation of the distributed hydrologic model threshold frequency (dhm-tf) flash flood forecasting system. Technical report, U.S. Dept. of Commerce,NOAA/National Weather Service, Silver Spring, MD 20910, 2012.

G. Delrieu, A. Wijbrans, B. Boudevillain, D. Faure, L. Bonnifait, and P.E. Kirstetter. Geostatistical radar-raingauge merging: a novel method for the quantification of rainfall estimation error. *Adv. in Water Res.*, 71:110–124, 2014.

Q. Y. Duan, V. K. Gupta, and S. Sorooshian. Shuffled complex evolution approach for effective and efficient global minimization. *Journal of Optimization Theory and Applications*, 76:501–521, 1993.

R. A. Fulton, J. P. Breidenbach, D. J. Seo, D. A. Miller, and T. O'Bannon. The WSR-88D rainfall algorithm. *Wea. Forecasting*, 13(2):377–395, 1998.

R. A. Fulton, F. Ding, and D. Miller. Truncation errors in historical WSR-88D rainfall products. Seattle, WA, 2003. 31th Conference on Radar Meteorology, Amer. Meteor. Soc.

J. J. Gourley, J. M. Erlingis, and and E. B. Wells Y. Hong. Evaluation of tools used for monitoring and forecasting flash floods in the united states. *Wea. Forecasting*, 27:158–173, 2012.

H. V. Gupta, S. Sorooshian, and P. O. Yapo. Towards improved calibration of hydrologic models: Multiple and non-commensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998.

S. P. Hardegree, S. S. Van Vactor, D. H. Levinson, and A.H. Winstra. Evaluation of NEXRAD radar precipitation products for natural resource applications. *Rangeland Ecology and Management*, 61:346–353, 2008.

Interagency Advisory Committee on Water Data. Guidelines for Determining Flood Flow Frequency. Bulletin 17B of the Hydrology Subcommittee.

Technical report, Office of Water Data Coordination, U.S. Geological Survey, Reston, VA 22092, 1982.

M.G. Kendall. *Rank Correlation Methods*. Charles Griffin, London, UK, 1975.

D. Kitzmiller, S. Van Cooten, F. Ding, K. Howard, C. Langston, J. Zhang, H. Moser, Y. Zhang, J. J. Gourley, D. Kim, and D. Riley. Evolving multisensor precipitation estimation methods: Their impacts on flow prediction using a distributed hydrologic model. *J. Hydromet.*, 12:1414–1431, 2011.

V. Kuzmin, D.-J. Seo, and V. Koren. Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. *J. Hydrology*, 353:109–128, 2008.

J.P. Looper, B. E. Vieux, and M. A. Moreno. Assessing the impacts of precipitation bias on distributed hydrologic model calibration and prediction accuracy. *J. Hydrology*, 418–419:110–122, 2012.

H. B. Mann. Non-parametric tests against trend. *Econometrica*, 13:163–171, 1945.

National Research Council. *Flash Flood Forecasting Over Complex Terrain: With an Assessment of the Sulphur Mountain NEXRAD in Southern California*. The National Academies Press, Washington, DC, 2005. ISBN 978-0-309-09316-3. URL http://www.nap.edu/catalog/11128/flash-flood-forecasting-over-complex-terrain-with-an-assessment-of.

L. Oudin, C. Perrin, T. Mathevet, V. Andreassian, and C. Michel. Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *J. Hydrology*, 320:62–83, 2006.

S. Reed, V. Koren, M. Smith, Z. Zhang, F. Moreda, D-J. Seo, and DMIP Participants. Overall distributed model intercomparison project results. *J. Hydrology*, 298:27–60, 2004.

S. Reed, J. Schaake, and Z. Zhang. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrology*, 337:402–420, 2007.

S. M. Reed and D. R. Maidment. Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. *J. Hydrol. Engrg.*, 4 (2):174–182, 1999.

B. Renard, D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46: n/a–n/a, 2010. doi: 10.1029/2009WR008328.

T. G. Schmitt, M. Thomas, and N. Ettrich. Analysis and modeling of flooding in urban drainage systems. *J. Hydrology*, 299:300–311, 2004.

K. Sene. *Flash Floods: Forecasting and Warning*. Springer Netherlands, Dordrecht, Netherlands, 2012. doi: 10.1007/978-94-007-5164-4.

D.-J. Seo. Real-time estimation of rainfall fields using radar rainfall and rain gage data. *J. Hydrology*, 208:37–52, 1998.

D.-J. Seo and J. Breidenbach. Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *J. Hydromet.*, 3:93–111, 2002.

D.-J. Seo, J. P. Breidenbach, and E. R. Johnson. Real-time estimation of mean field bias in radar rainfall data. *J. Hydrology*, 223:131–147, 1999.

D.-J. Seo, A. Seed, and G. Delrieu. Radar and multisensor rainfall estimation for hydrologic applications. In F. Y. Testik and M. Gebremichael, editors, *Rainfall, State of the Science*, pages 79–104. AGU, 2011.

D. J. Seo, R. Siddique, Y. Zhang, and D. Kim. Improving real-time estimation of heavy-to-extreme precipitation using rain gauge data via conditional bias-penalized optimal estimation. *J. Hydrology*, 519:1824–1835, 2014.

S. K. Singh and A. Bàrdossy. Calibration of hydrological models on hydrologically unusual events. *Adv. in Water Res.*, 38:81–91, 2012.

J. A. Smith, D. J. Seo, M. L. Baeck, and M. D. Hudlow. An intercomparison study of NEXRAD precipitation estimates. *Water Resources Research*, 32 (7):2035–2045, 1996.

M. Smith, V. Koren, Z. Zhang, Y. Zhang, S. Reed, Z. Cui, F. Moreda,

B. Cosgrove, N. Mizukami, E. Anderson, and DMIP 2 Participants. Results of the DMIP 2 Oklahoma experiments. *J. Hydrology*, 418-419:17–48, 2012.

M. Strauch, C. Bernhofer, S. Koidec, M. Volkd, C. Lorza, and F. Makeschin. Using precipitation data ensemble for uncertainty analysis in swat streamflow simulation. *J. Hydrology*, 414-415:413–424, 2012.

S. Sun and J. Bertrand-Krajewski. Separately accounting for uncertainties in rainfall and runoff: Calibration of event-based conceptual hydrological models in small urban catchments using bayesian method. *Water Resources Research*, 49:5381–5394, 2013. doi: 10.1002/wrcr.20444.

J. A. Vrugt, H. V. Gupta, W. Bouten, and S. Sorooshian. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39 (8), 2003.

J. A. Vrugt, C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon. Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. international journal of nonlinear sciences and numerical simulation. *Water Resources Research*, 273–290(10), 2009.

I. K. Westerberg, J.-L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C.-Y. Xu. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7):

2205–2227, 2011. doi: 10.5194/hess-15-2205-2011. URL `http://www.hydrol-earth-syst-sci.net/15/2205/2011/`.

H. C. Winsemius, B. Schaefli, A. Montanari, and H. H. G. Savenije. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, 45(12):n/a–n/a, 2009. ISSN 1944-7973. doi: 10.1029/2009WR007706. URL `http://dx.doi.org/10.1029/2009WR007706`. W12422.

C. J. Woltemade and K. W. Potter. A watershed modeling analysis of fluvial geomorphologic influences on flood peak attenuation. *Water Resources Research*, 30(6):1933–1942, 1994. ISSN 1944-7973. doi: 10.1029/94WR00323. URL `http://dx.doi.org/10.1029/94WR00323`.

C. B. Young, B.R. Nelson, A.A. Bradley, J.A. Smith, C.D. Peters-Lidard, A. Kruger, and M.L. Baeck. An evaluation of NEXRAD precipitation estimates in complex terrain. *J. Geophys. Res.*, 104(D16):19691–19703, 1999.

C. B. Young, A. A. Bradley, W. F. Krajewski, A. Kruger, and M. L. Morrissey. Evaluating NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. *J. Hydromet.*, 1:241–254, 2000.

Y. Zhang, J. A. Smith, and M. L. Baeck. The hydrology and hydrometeorology of extreme floods in the great plains of eastern nebraska. *Adv. in Water Res.*, 24(9–10):1037–1050, 2001.

Y. Zhang, J. A. Smith, and M. L. Baeck. Space-time variability of rainfall and extreme flood response in the Menomonee River Basin, Wisconsin. *J. Hydromet.*, 4(3):506–517, 2003.

Y. Zhang, S. Reed, and D. Kitzmiller. Effects of retrospective gauge-based readjustment of multisensor precipitation estimates on hydrologic simulations. *J. Hydromet.*, 12:429–443, 2011a.

Y. Zhang, Z. Zhang, S. Reed, and V. Koren. An Enhanced and Automated Approach for Deriving a Priori SAC-SMA Parameters from the Soil Survey Geographic Database. *Computers and GeoSciences*, 37:219–231, 2011b.

Z. Zhang, V. Koren, S. Reed, M. Smith, Y. Zhang, F. Moreda, and B. Cosgrove. SAC-SMA a priori parameter differences and their impact on distributed hydrologic model simulations. *J. Hydrology*, 420-421:216–227, 2012.

Table 1: Study watersheds

Station	USGS ID	Latitude [°N]	Longitude [°W]	Area [km^2]	T_p	Name
VNOVA	01589300	39°20′45″	76°43′59″	84	4	Gwynns Falls at Villa Nova,MD
NWANAC	01651000	38°57′08″	76°57′57″	128	2	NW. Br Anacostia R,MD
ROCKS	01648000	38°58′21″	77°02′24″	161	4	Rock Ck Sherrill Dr, MD
WASHB	01589352	39°16′17″	76°38′54″	171	9	Gwynns Falls Washington Blvd, DC
CATOC	01637500	39°25′38″	77°33′22″	173	5	Catoctin Ck near Middletown, MD
NEANAC	01649500	38°57′36″	76°55′33″	189	3	NE Branch Anacostia R, MD
WBRANCH	01594526	38°48′51″	76°44′55″	232	5	Western Br. at Upper Marlboro, MD
DAWM2	01645000	39°07′41″	77°20′08″	262	4	Seneca Ck at Dawsonville, MD
LNGP1	01465500	40°10′26″	74°57′26″	544	6	Neshaminy Ck nr Langhorne, PA
CPHP1	01571500	40°13′29″	76°53′54″	552	14	Yellow Breeches Ck nr Camp Hill, PA
SPKP1	01558000	40°36′45″	78°08′27″	570	2	Little Juniata R Spruce Ck, PA
MBGW2	01616500	39°25′25″	77°56′20″	707	9	Opequon Ck nr Martinsburg, WV
ANTIE	01619500	39°26′59″	77°43′48″	728	8	Antietam Ck nr Sharpsburg, MD
WIBP1	01556000	40°27′47″	78°12′00″	754	9	Frankstown Br Juniata R, PA
PNCP1	01555000	40°52′00″	77°02′55″	780	10	Penns Ck Penns CK, PA
LEEV2	01644000	39°01′10″	77°34′40″	860	8	Goose Ck nr Leesburg, VA
PATUXB	01594440	38°57′21″	76°41′37″	901	23	Patuxent R nr Bowie, MD
CANOC	01614500	39°42′59″	77°49′29″	1279	9	Conococheague Ck Fairview, MD
MONOC	01643000	39°24′10″	77°21′57″	2116	9	Monocacy R Jug Bridge, MD

Table 2: Model Parameters for Calibration

Module hline	Parameter Acronym	Parameter Name	Typical Range
SAC-SMA	UZTWM	Upper zone tension water capacity	10-300 mm
	UZFWM	Upper zone free water capacity	5-150 mm
	UZK	Interflow depletion rate,	0.1-0.75 day^{-1}
	ZPERC	Shape parameter of the percolation curve	1-5
	LZTWM	The lower zone tension water capacity	10-500 mm
	LZFSM	The lower zone supplemental free water capacity	5-400 mm
	LZFPM	The lower zone primary free water capacity	10-1000 mm
	LZSK	Depletion rate of lower zone supplemental free water storage	0.01-0.35 day^{-1}
	LZPK	Depletion rate of lower zone primary free water storage	0.001-0.05 day^{-1}
Routing	QMCHN	Rating curve exponent	1-2
	Q0CHN	Channel specific discharge	0.05-0.5 m^3s^{-1}

Table 3: Model Parameters and Calibration Outcome

Module	Parameters	Scalar Multiplier			
		Small/Raw	Small/Adj	Large/Raw	Large/Adj
SAC-SMA	UZTWM	0.16	0.14	0.43	0.45
	UZFWM	1.21	1.16	1.73	1.77
	UZK	1.49	1.52	1.15	1.13
	ZPERC	0.80	0.95	1.45	1.35
	LZTWM	0.44	0.42	0.33	0.34
	LZFSM	1.45	1.49	1.49	1.46
	LZFPM	1.87	1.77	1.46	1.53
	LZSK	0.43	0.52	0.87	0.80
	LZPK	1.17	1.29	1.63	1.55
Routing	QMCHN	1.52	1.53	0.96	0.95
	Q0CHN	1.71	1.71	1.41	1.41

Table 4: % of Basins with Lowered Median Peak and Reduced Bias

Calibration	Period	% Decreased			% Reduced Bias		
		Total	Small	Large	Total	Small	Large
No	1998-2013	58	50	64	63	50	73
	2004-2013	95	100	91	42	0	73
Yes	1998-2013	26	38	18	63	38	82
	2004-2013	74	88	64	32	25	36

Table 5: Net percentage of basins with improvements in LP3 Parameters with Adjustment

	Uncalibrated			Calibrated		
	All	Small	Large	All	Small	Large
Mean	58	100	28	42	-12	82
Std. Dev.	78	76	82	36	50	28
Skew	-6	0	-10	-48	-76	-28

Table 6: Net Percentage of basins with improvements in POD and FAR

ARI	Metrics	% Uncalibrated			% Calibrated		
		All	Small	Large	All	Small	Large
2-Y	POD	0	-12	9	-21	-75	18
	FAR	21	0	36	-5	-25	9

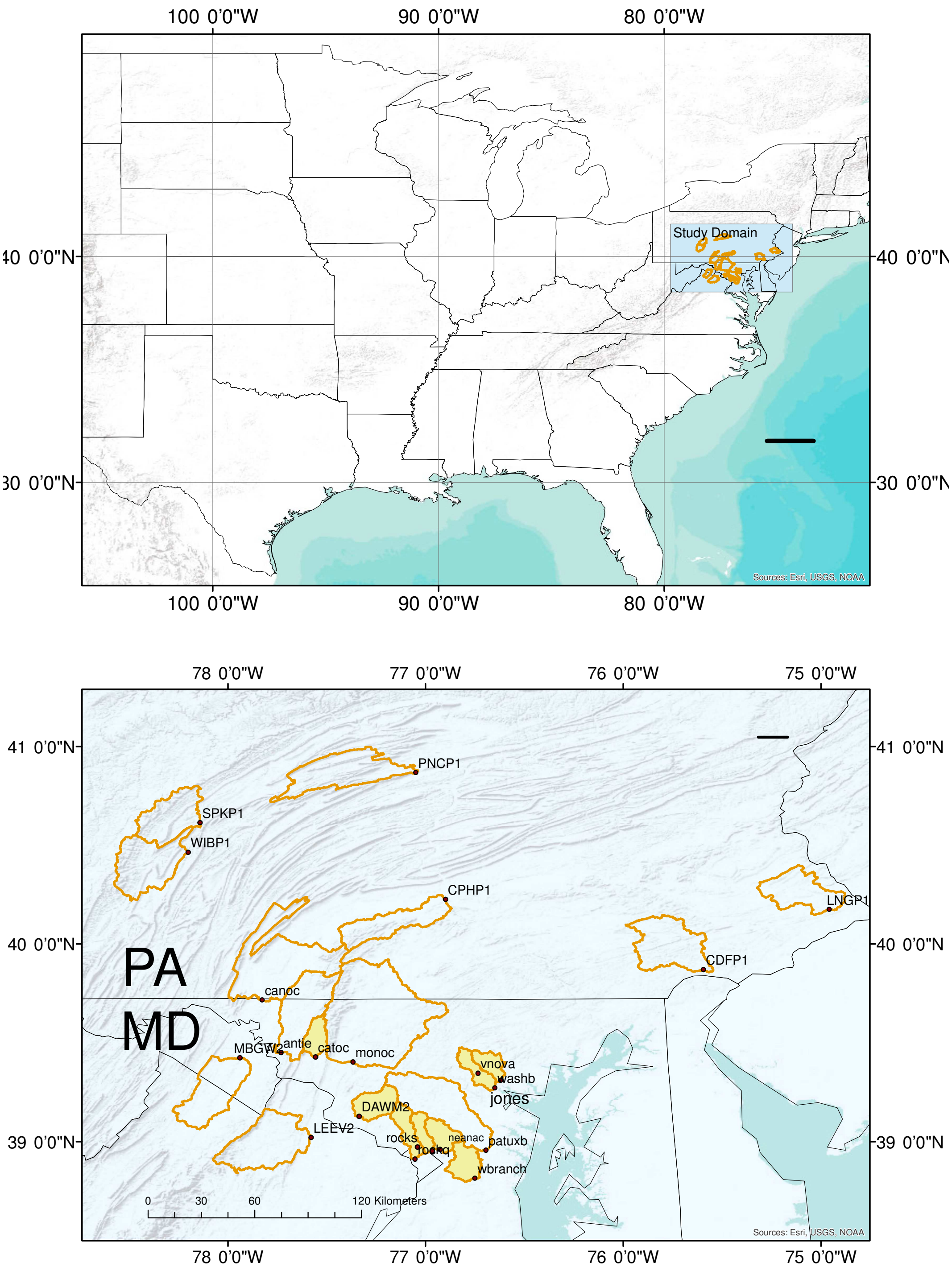

Figure 1: The geographic location of the study domain in the US (top) and the catchments of interest (bottom).

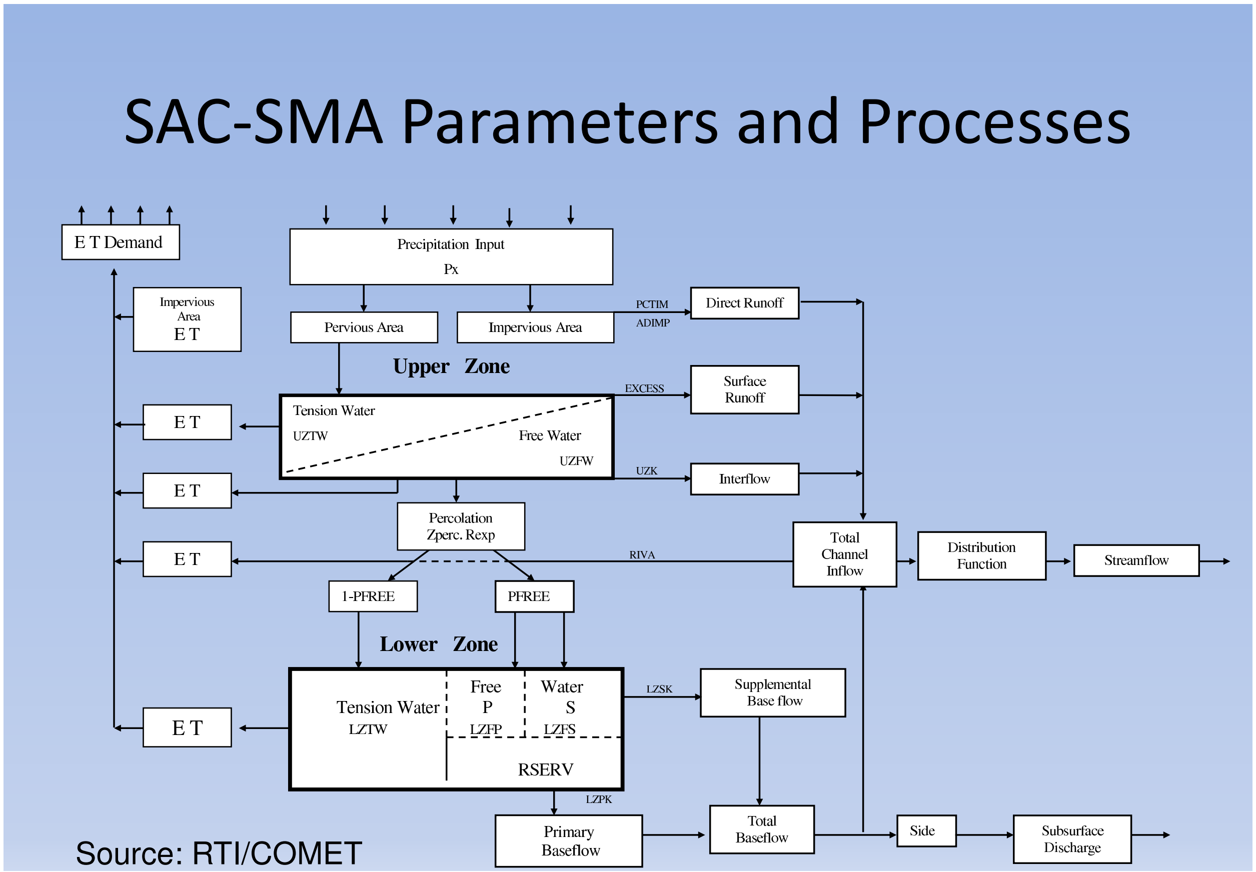

Figure 2: Schematic of SACramento Soil Moisture Accounting (SAC-SMA) model and parameters.

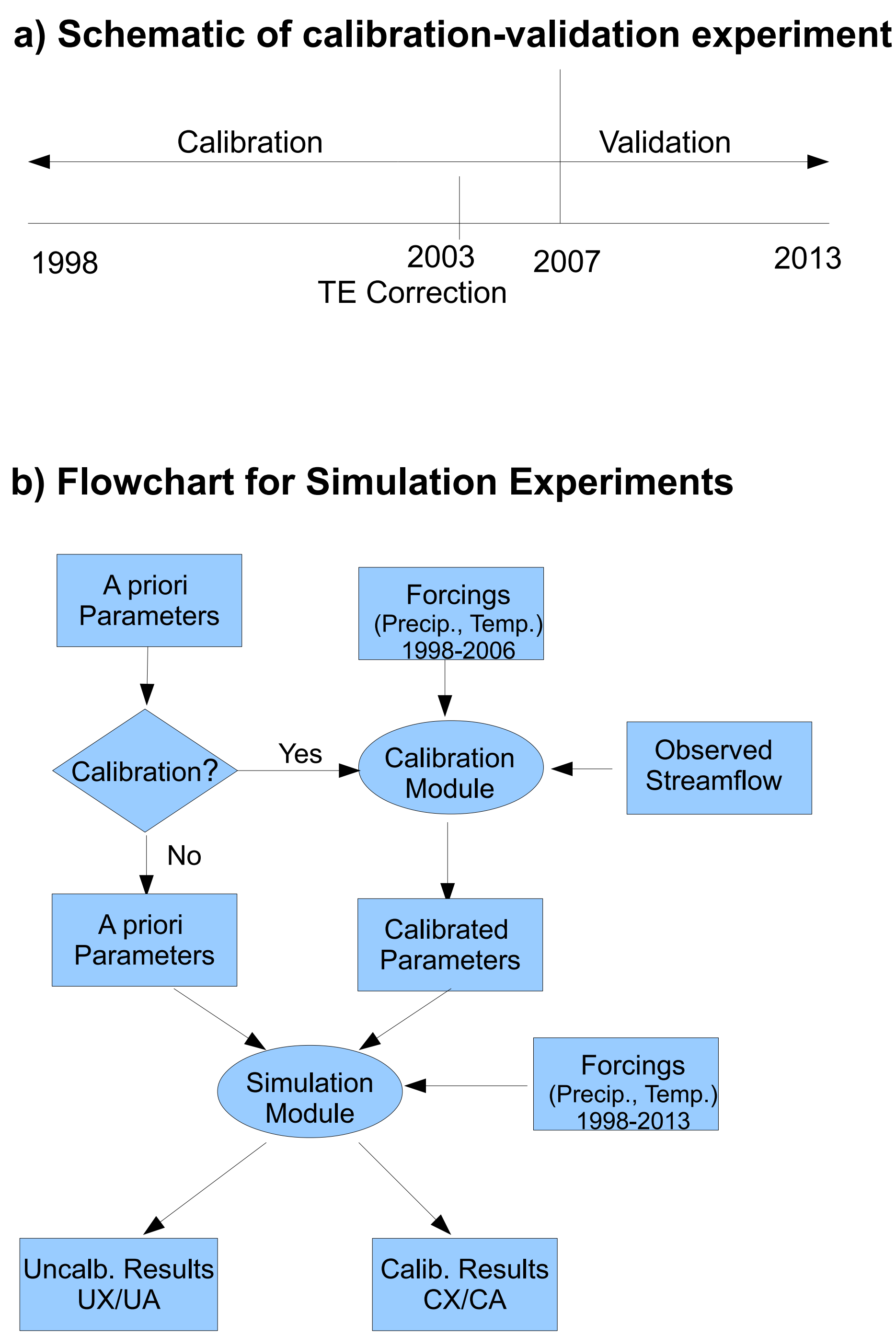

Figure 3: a) Schematic of the calibration-validation process and b) flowchart of the simulation experiment.

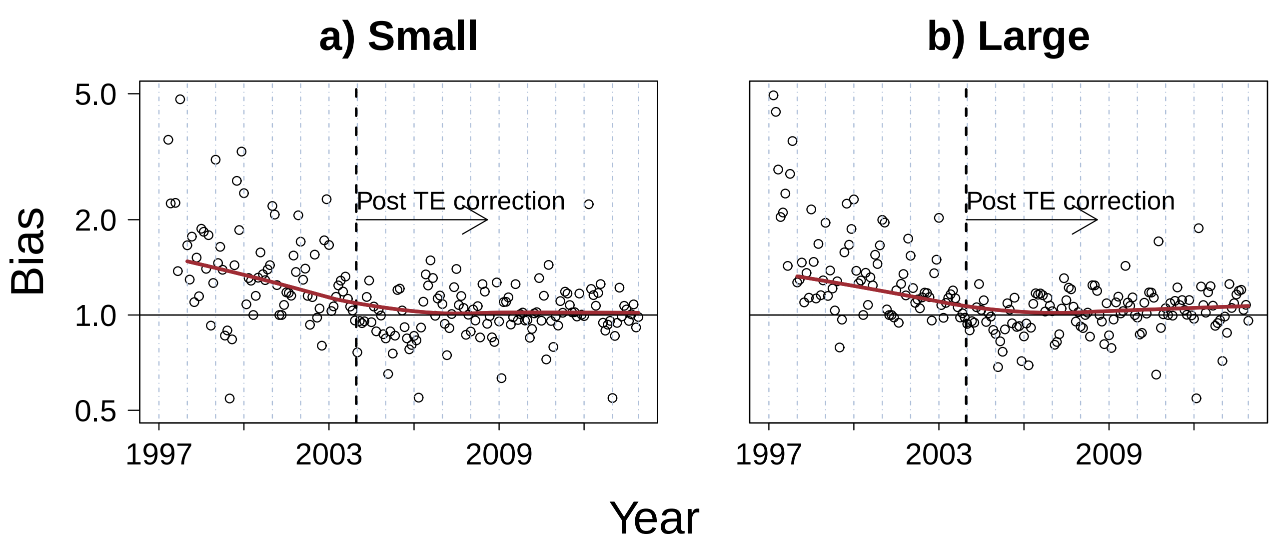

Figure 4: Time series of multi-basin mean of monthly adjustment factors for a) small and b) large basins. Vertical lines mark the approximate date when the truncation error was corrected. Superimposed is the locally weighted regression smoother curve. Note the conspicuous downward trend of adjustment factors prior to the TE correction.

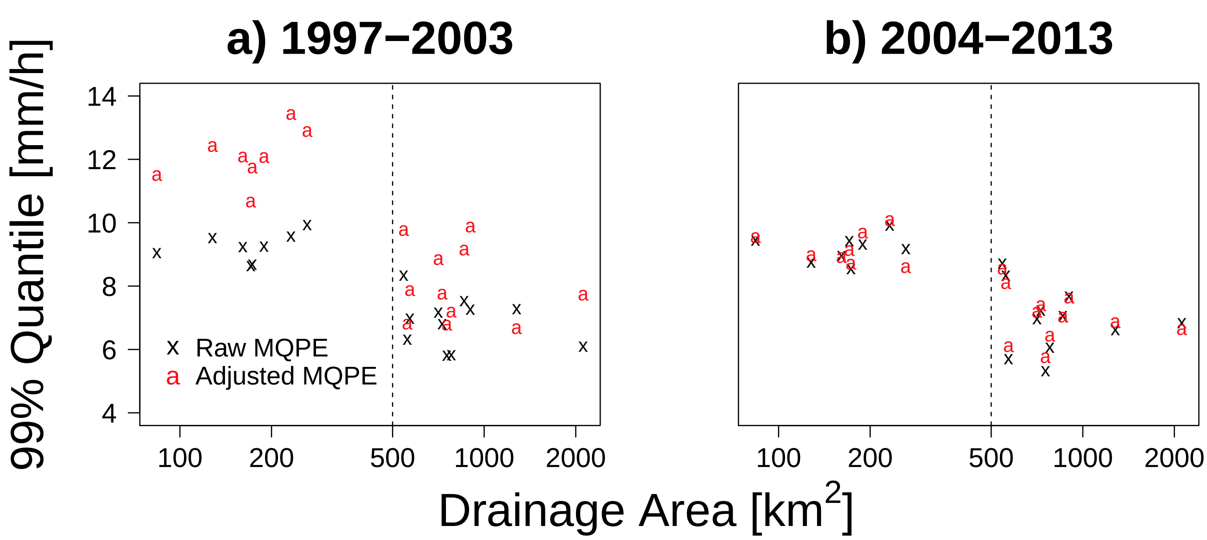

Figure 5: 99% quantiles of hourly mean areal precipitation before and after adjustment versus drainage area, for a) the entire record and b) the post-TE era (2004-2013).

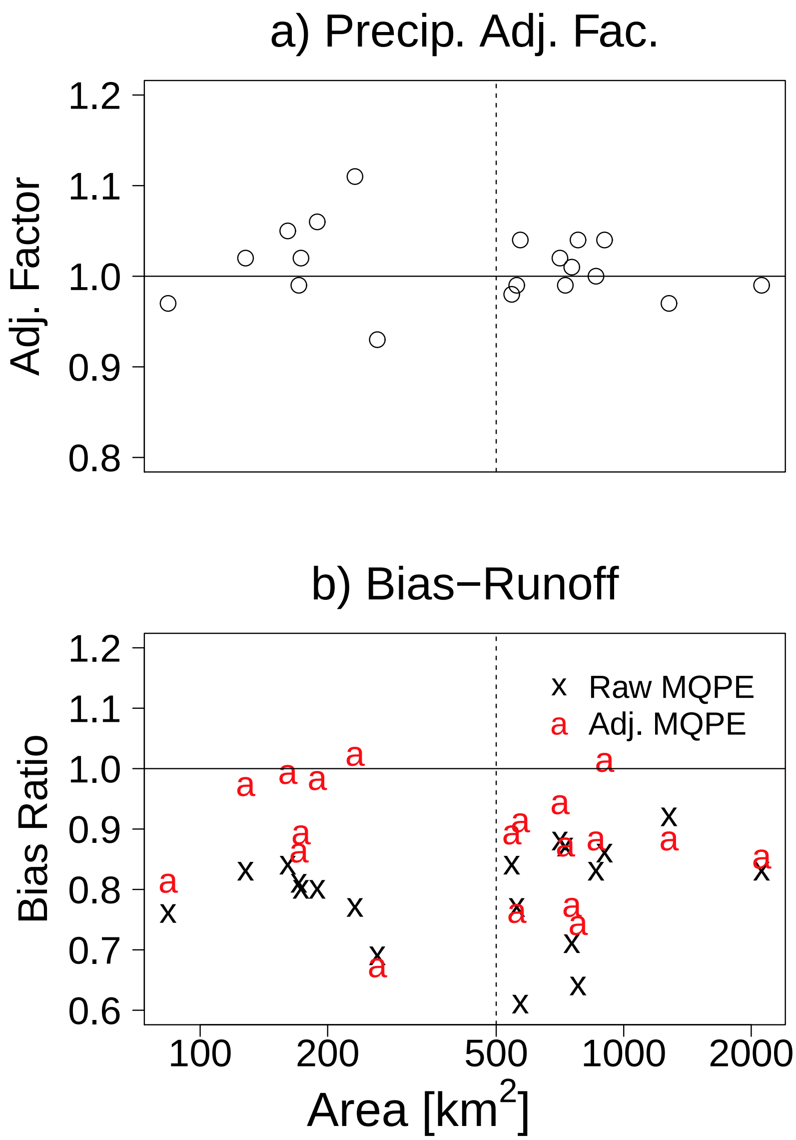

Figure 6: Dependence of a) precipitation adjustment factor (ratio of PRISM to MQPE-based totals), and b) bias ratio (simulation/observation) of cumulative runoff over 1998-2013 for each basin as a function of drainage area. Simulations based on both raw (x) and adjusted (a) MQPE are shown in b).

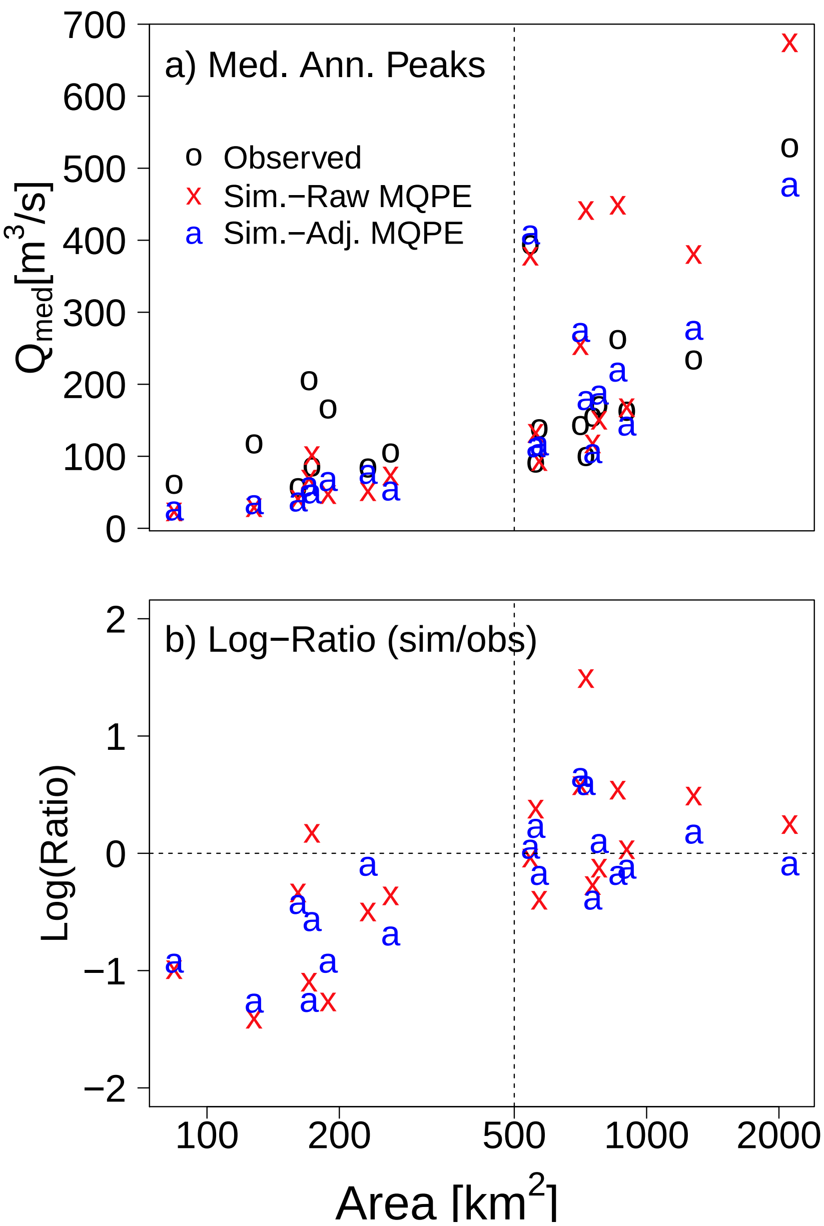

Figure 7: a) Median annual peaks from observed ('o'), simulated discharge with raw and adjusted MQPE ('x' and 'a') using *a priori* model parameters as a function of drainage area computed for the entire period (1998-2013) and b) the associated ratios of simulated to observed median peaks.

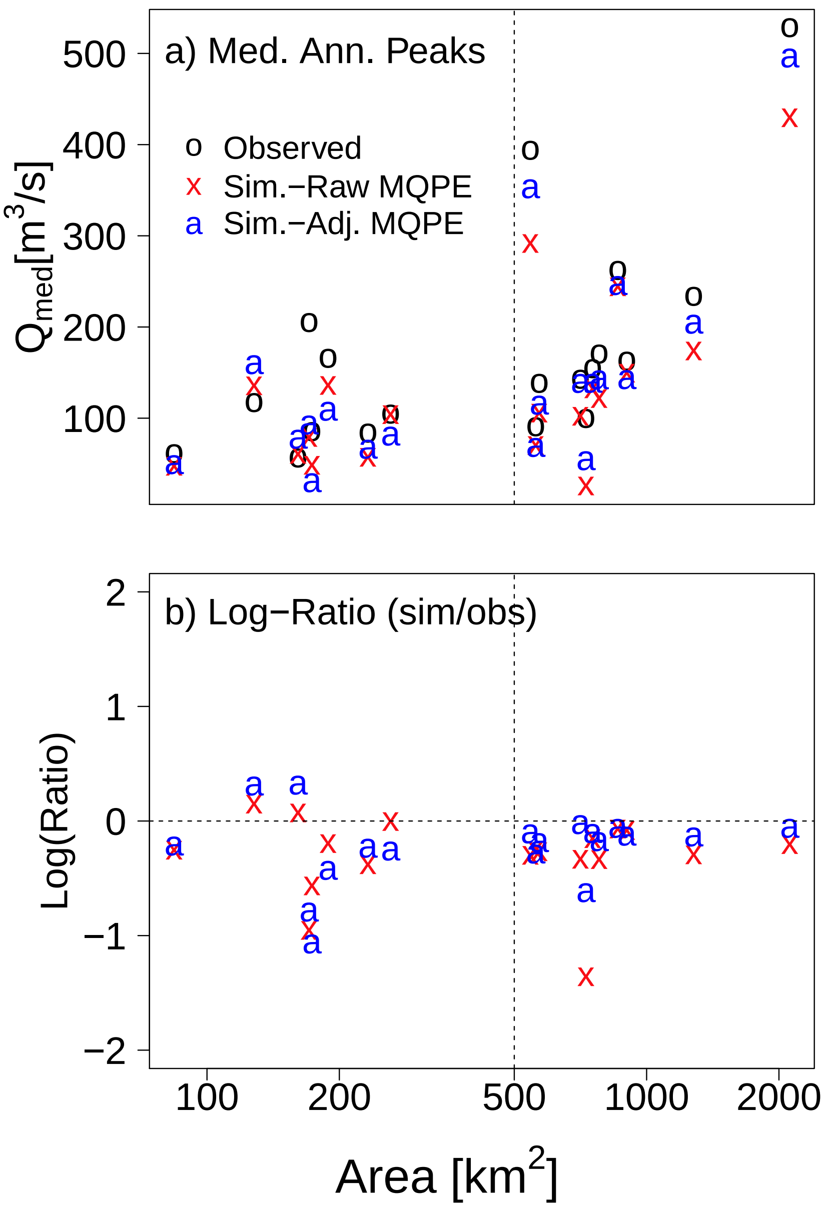

Figure 8: As in Fig.7, except based on calibrated model simulations.

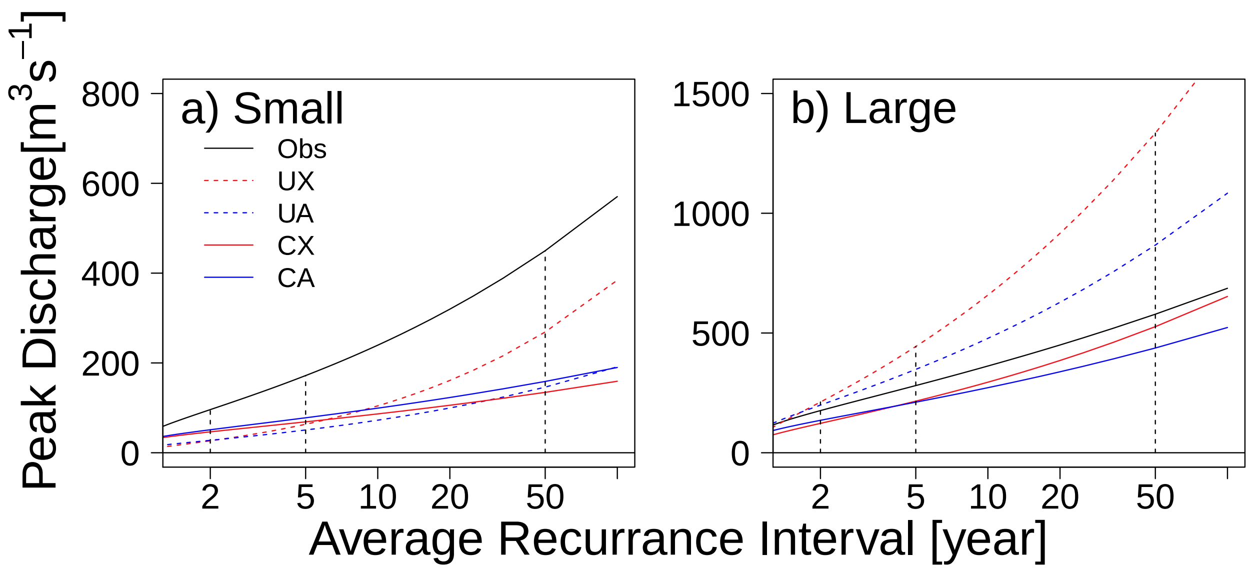

Figure 9: Sensitivity of the flood frequency (FF) curve based on the Log Pearson Type III (LP3) distribution to variations in LP3 parameters among UX, UA, CX and CA for a) small and b) large watersheds. These FFD curves are constructed using the multi-basin mean of parameters derived from each set of simulation results.

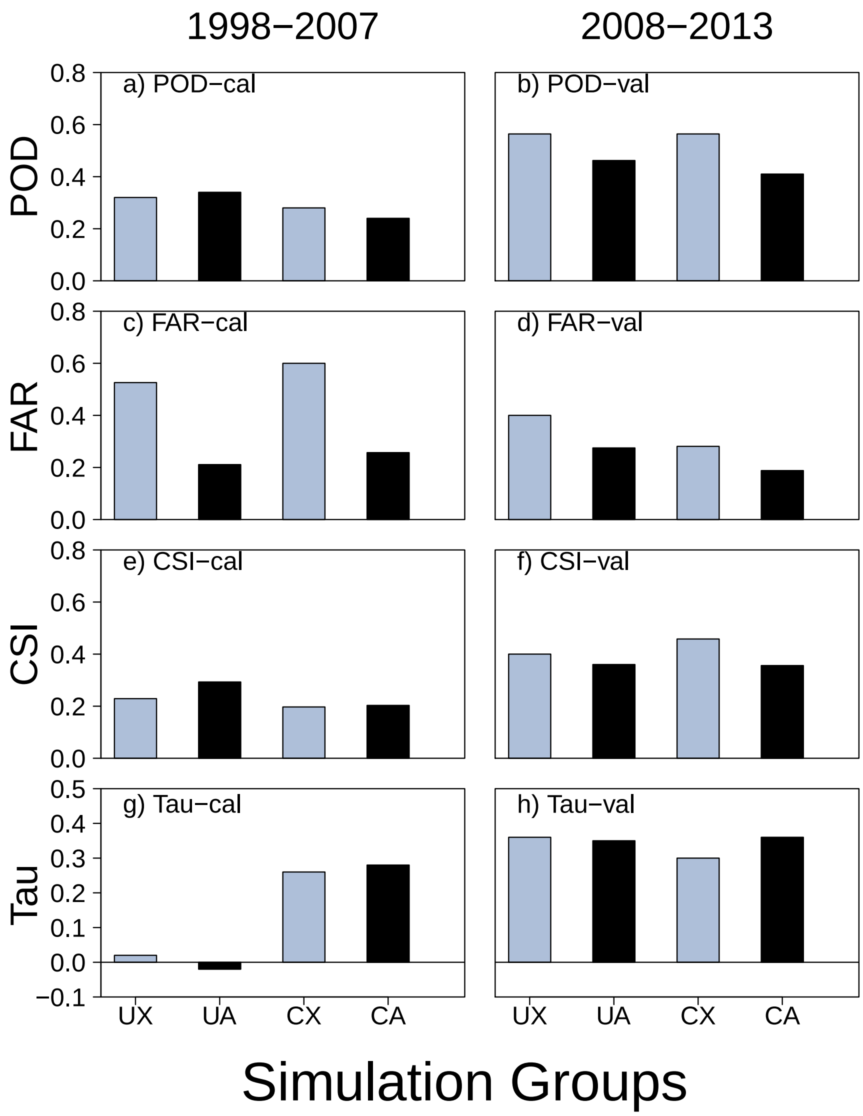

Figure 10: Accuracy of model simulations in capturing the flood events as gauged by multi-basin aggregate probability of detection (POD), false Alarm Ratio, critical success index (CSI), and ranked correlation (Tau) for small basins. Shown on the left and right panels are the outcomes for the calibration (1998-2007; denoted by "cal") and validation (2008-2013; denoted by "val"). As in Fig. 9., "UX" and "UA" denote the results of uncalibrated model runs with raw and adjusted MQPE, respectively; whereas "CX" and "CA" denote those for calibrated model runs.

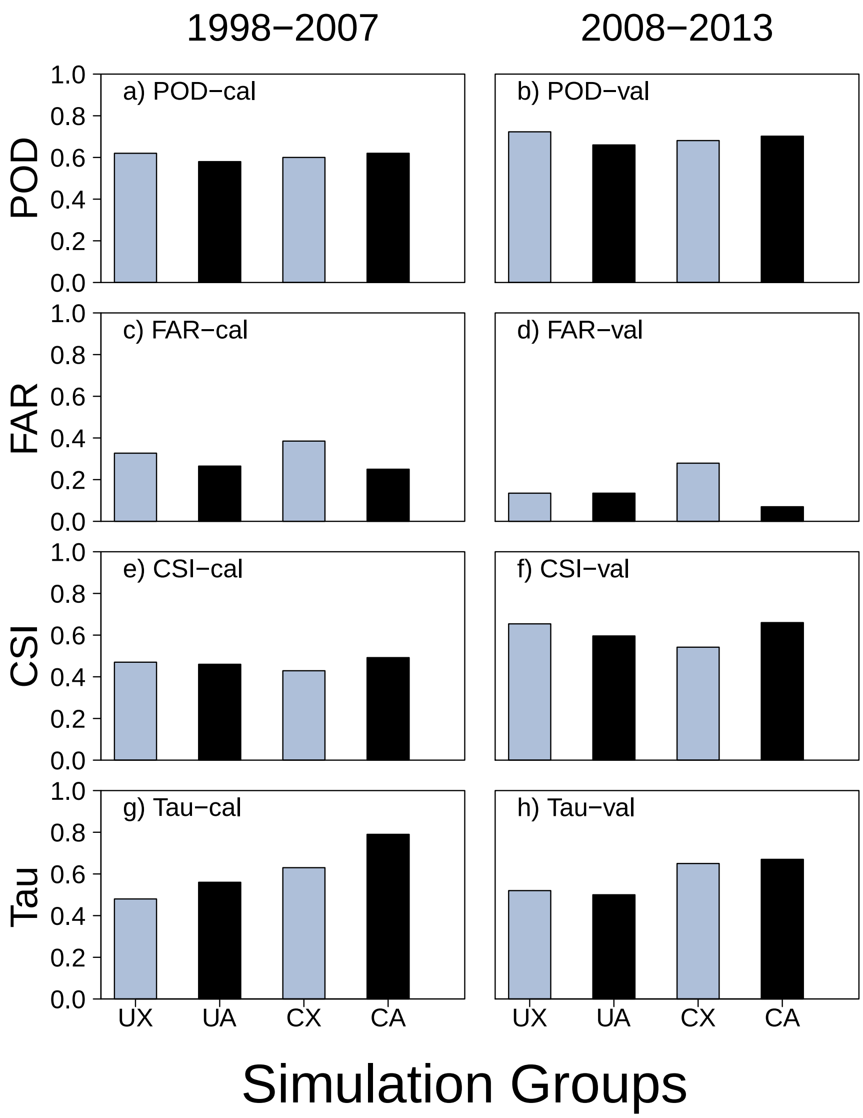

Figure 11: As in Fig. 10, except for larger watersheds.

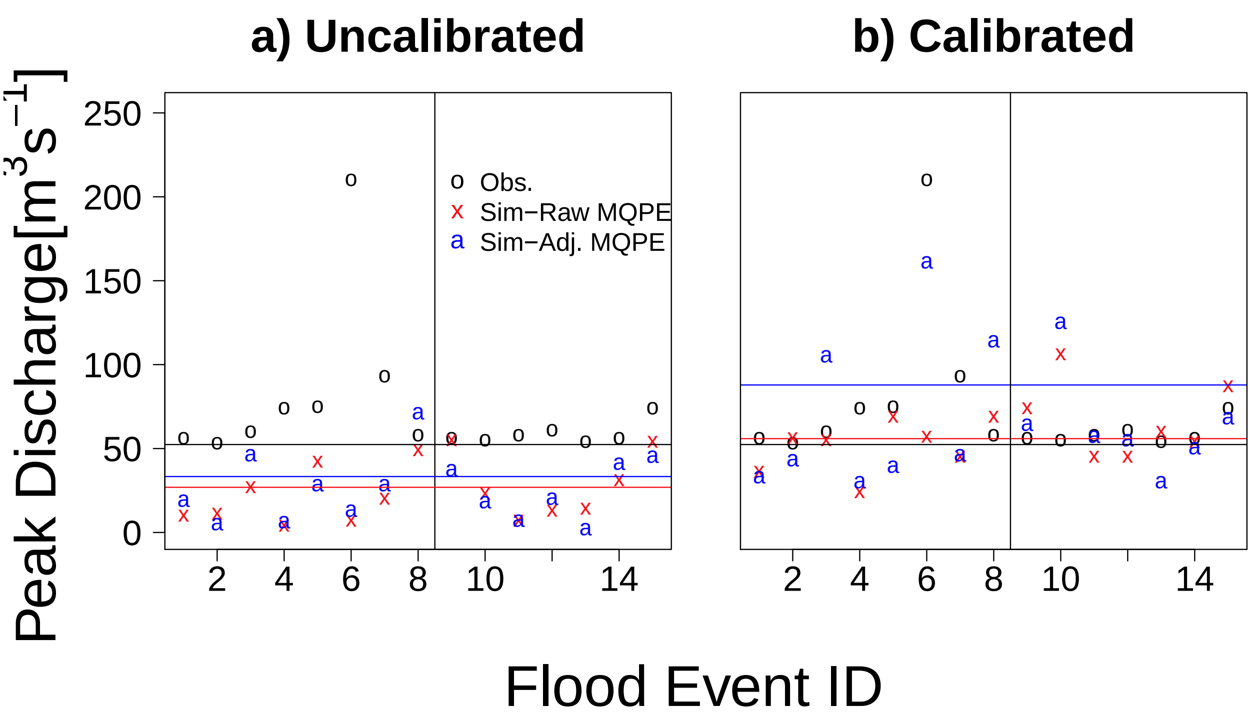

Figure 12: Simulated peak discharge based on a) uncalibrated and b) calibrated model runs for the basin **ROCKS**. Horizontal lines represent the thresholds (2-Y quantile) based on observed and simulated annual peak discharge computed using raw and adjusted MQPE as forcing. The vertical line in each panel separates the calibration (left) and validation (right) periods.

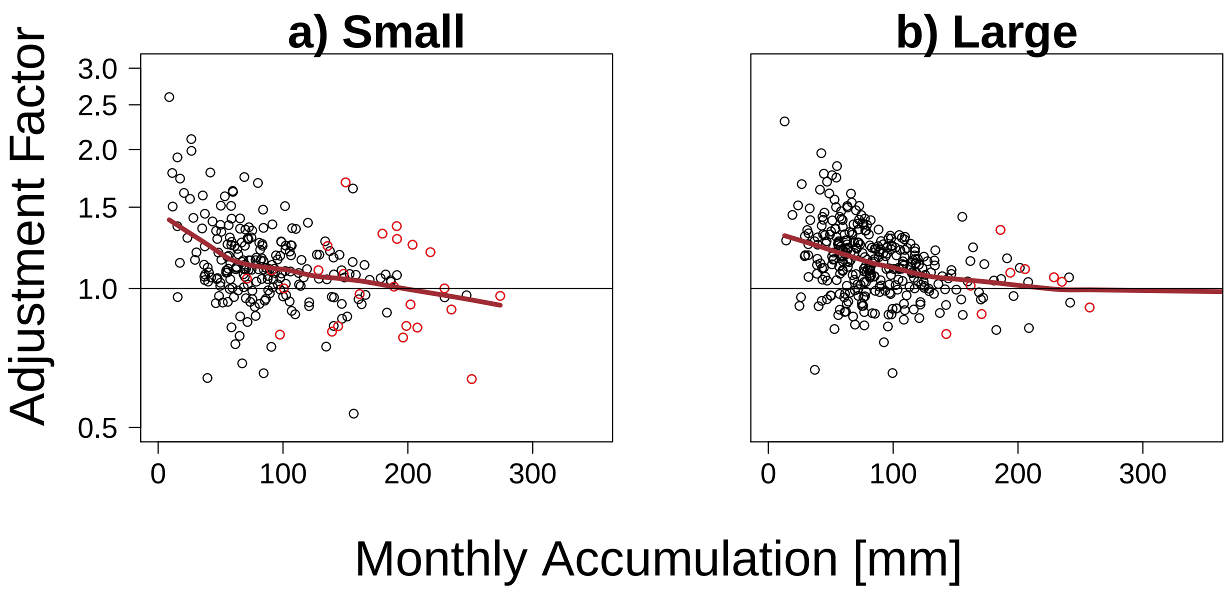

Figure 13: Monthly adjustment factor (ratio of accumulation based on raw to that based on adjusted MQPE) versus precipitation accumulation for the summer (June-August), for a) small and b) large watershed groups. Months with at least one flood reported are highlighted in red.