At the outset, we wish to thank our reviewers and our handling editor for their deep consideration of our work. Their comments have provided a substantial improvement to the manuscript. The following contains all the original comments, the original author responses and final author responses directing the reader to the relevant changes. All line numbers in author comments refer to the tracked-changes documents included here.

1 First Anonymous Reviewer

Reviewer Comment 1: *This is an interesting paper, which in my opinion could be made easier to follow with some adjustments. I make a few suggestions that I honestly believe could improve its readability and subsequent impact.*
Author Response 1: Thank you for taking the time to review our manuscript. There is certainly a lot of material presented, so we appreciate your recommendations for improving readability. If you continue to have concerns, do not hesitate to comment again.
Final Author Response 1: We wish to thank you again for your consideration of this work. Your suggestions, particularly the figures, have allowed us to better present our material and addressed some of the concerns raised by other reviewers.

Reviewer Comment 2: *Even a specialist of statistical hydrology could use one or two hydrographs (you only show us box plots!). An introductory graph with an example hydrograph and FDC could help the reader understand your methodology.*
Author Response 2: For the revision, based on this and other reviews, we will include a two-panel graphic showing the overlay of an observed and simulated hydrograph (for a representative site) with a second panel showing the overlay of observed and simulated flow duration curves. We will also consider a figure that shows the steps of the methodology.
Final Author Response 2: In the end, we felt that a single graphic of the methodology provided all the context needed without providing too many figures. This additional figure is presented and discussed on page 6, line 17 of the tracked-changes document. With this figure, the reader can easily see how an example hydrograph would be bias corrected.

Reviewer Comment 3: *I do not like the way you deal with all the aspects of the methodology in parallel, it makes things very difficult to understand what you are doing. I would have preferred a paper structure where (a) you show us what a "perfect" simulation of the FDC used for bias correction could give for results, then (b) you would show that due to the inherent uncertainty of FDC prediction at ungauged points you loose a lot of the theoretical advantage, while managing to improve overall bias*
Author Response 3: This is a great suggestion for the flow of the paper. Other reviewers also made suggestions about the flow of the paper. We definitely plan to make revisions for clarity but need to determine which ideas are most appropriate.
Final Author Response 3: The structure proposed is used in the Results section of the paper. With the inclusion of our new methodology figure (page 6, line 17), the methodology is much more accessible.

Reviewer Comment 4: *Last, I believe that in addition to box-plots, you should also show the reader some QQ plots to show that even if on average there is a reduction of bias, there will always be catchments where the bias correction method will increase the bias : e.g. a plot showing the original low flow bias vs the bias corrected low-flow bias (with one point per catchment), and then the same for high flow.*
Author Response 4: This will be a useful figure, so we will consider weaving it into our manuscript.
Final Author Response 4: This new figure is presented and discussed on line 27 of page 13. This excellent addition allows us to highlight the variability of performance across sites and across tails.

Reviewer Comment 5: *P2 L15. "the nature of this approach. . ." : I have difficulties to understand this sentence. . .*
Author Response 5: This sentence is meant to point out that the timing of the raw simulation is not altered, only the magnitude. That is, the methodology assumes that the timing (sequence of relative rankings) is "good". We will find a better way to say this. One possible revision might be, "This approach assumes that, while the streamflow magnitudes of a historical simulation are biased, the timing or rank-order of the streamflows are relatively accurate."
Final Author Response 5: This revision appears on line 19 of page 2.

2 Second Anonymous Reviewer

Reviewer Comment 1: *The work presents a method that aims to produce unbiased time series of daily discharges for ungauged basins. It consists in correcting the distributional bias in the time series of simulated discharges obtained for this ungauged basin from some simulation model (in the present case, the simulation model is a regression model where the daily discharge of each day is estimated via kriging from gauged stations in the neighborhood of the target ungauged station). The correction is a quantile –mapping correction where the reference statistical distribution is obtained from the Flow-Duration-Curve of the target location, estimated via regionalization again from the observed FDC estimated for the neighboring stations.*

The issue is a very relevant one and the approach a promising one. The novelty of the approach with respect to other similar studies (mentioned in the manuscript) has to be clarified.

Author Response 1: Thank you for your deep consideration of our manuscript. Your comments have helped us to identify weaknesses in presentation that substantially improve our presentation.

Final Author Response 1: Thank you, again, for your attention to our work. Addressing your comments surely helped us sharpen the presentaiton of our analyses.

Reviewer Comment 2: *More specifically, the aim of the paper is not well explained and it has taken me more than 2 hours to understand what was done in the paper. One or possibly several graphical schemes are definitively required to figure out the methodology / objective of the paper and data used for the application of the method. A reformulation of the introduction and of the abstract is for me also necessary. Showing figures with hydrographs or FDC is also required (boxplots are not informative enough).*

Author Response 2: Based on your comments and those of another reviewer, we are proposing to reorganize the manuscript and present several additional graphics. One will be a two-panel figure where one panel overlays observed and simulated hydrographs, while the second panel shows observed and simulated flow duration curves. A second figure will show how the methodology proceeds from simulated hydrograph, through estimated FDC to bias-corrected hydrograph. We will also take another look at the clarity of the introduction and the abstract.

Final Author Response 2: On page 2, line 2 of the tracked-changes document we have added an explicit statement of the objective and novelty of this work. Novelty is further discussed on page 6, line 24. In addition, we have added the figure discussed on page 6, line 17, to represent our methodology. This figure allows the reader to better understand both our objective and our methodology.

Reviewer Comment 3: *The formulation in the abstract for instance is somehow clumsy : "Based on an existing approach that separates the simulated streamflow into components of timing and magnitude, the timing component is converted into simulated nonexceedance probabilities and rescaled to new volumes using an independently estimated flow-duration curve (FDC)." I think that the authors do not rescale the timing component. They just scale each daily discharge of the time series with a multiplicative correction factor that depends on the non exceedance probability of that discharge. . . The correction function is estimated from the FDC of the target location, estimated from the observed FDC of gauged stations in the region. I would have understood instantaneously the objective / approach of the work that way (I am perhaps wrong, but I guess this is roughly what you do). This has definitively to be clarified (here and everywhere else in the manuscript).*

Author Response 3: The reviewer is correct, and we propose revising this sentence to "Based on an existing approach that separates the simulated streamflow into components of timing and magnitude, the timing component is converted into simulated nonexceedance probabilities and *the volumes are* rescaled using an independently estimated flow-duration curve (FDC) *derived from regional information*." We will correct this impression throughout the manuscript.

Final Author Response 3: At several places in the manuscript, we have made it clearer that we intend the use of observed FDCs to be an idealized analysis, not an application. For examples, see page 3, line 15, page 8, line 7, and page 9, line 15.

Reviewer Comment 4: *It is also not clear that the authors want to estimate unbiased time series of discharge at ungauged sites. This has to be clearly stated. One thus needs to have a regional simulation model and a regional model for the FDC estimation. The interest of the configuration where an observed FDC is used to bias correct the simulated time series has to be made clear in this context. This approach is presented as an alternative estimation approach. It is actually not and this is really confusing.*

Author Response 4: The primary goal of this work is to estimate distributionally unbiased time series of discharge at ungauged sites. This statement will be reinforced. The use of observed FDCs, which, of course, are not available at ungauged locations are presented as an "idealized" case (P1, L10). It is presented merely as demonstration of the "theoretical potential of this [general] approach" (P6, L10), providing evidence to support the hypothesis that such an approach would work with a suitably-estimated FDC (P6, L30). We will strive to make this point more clearly.

Final Author Response 4: In addressing comments 2 and 3, this comment has been addressed. We included a more explicit statement of the novelty of this work is discussed on page 2, line 3, and on page 6, line 24. Furthermore, we have highlight the idealized utility of the application with observed FDCs (page 3, line 15, page 8, line 7, and page 9, line 15).

Reviewer Comment 5: *The results of the study were rather expected. The authors find that the combined "time series simulation"+"FDC bias correction approach" is unable to correct the bias in the simulated hydrograph. The authors explain this is due to the bad quality of the regionalized FDC which present a rather significant distributional bias for all discharge percentiles. By construction, the "poor" result of the FDC bias correction was thus expected. The interest of the method would have been for a configuration where the distributional bias obtained from the "time series simulation" approach is larger than that of the regionalized FDC. Obviously, this is unfortunately here not the case. At least in average for the 1168 basins considered in the work. More convincing results could have perhaps been presented considering different configurations of basins with different relative performances of the two approaches considered here for the combination. We would have seen that the method works in configurations where reginoalised FDC are good.*

Author Response 5: It is surely expected that this two-step approach will depend on the accuracy of both steps. This project demonstrates the impact that FDC accuracy can have on the results. However, the results were not uniformly poor. As mentioned in the discussion, improvements were seen in certain parts of the distribution of daily streamflow. It is unfortunate that the results were not uniform improvements. In the revision we will identify a few sites that show improvements under the situation the reviewer describes (when bias obtained from the "time series simulation" approach is larger than that of the regionalized FDC). This will supplement our analysis of how performance of the estimated FDC corresponds to improvements in the time series.

Final Author Response 5: In the end, we felt that the addition of figures showing a range of representative performance was not efficient. Instead, we use Figure 10 presented on page 13, line 27, to highlight the variability of performance and augment the discussion in the previous paragraph.

Reviewer Comment 6: *P2. "As attested to by many researchers focused on the reproduction of historical streamflow, this bias commonly appears as a general overestimation of low streamflow and underestimation of high streamflow". This statement is not true. Simulations with any hydrological model will lead to the overestimation of some high stream flow and the underestimation of the others. The same for low flows : some are over-, some are underestimated. In rainfall-runoff models, it clearly depends on the amount of rainfall estimated from the small number of raingauge available, and frequently not numerous enough to have the right estimate of rainfall input. It also depends on the quality of the model which is limited for the simulation of specific hydro processes. The fact that a given model systematically overestimate lowflows and underestimate highflows surely means that the model is not good enough and could likely be improved. As a matter of fact, some models present those limitations and this justify the present work. But a general statement can not be given there on such a case. (Note that this statement is invalidated in your work by the results and comments you mention (p6 – line 20/25). Please reformulate.*

Author Response 6: The citations provided do note the tendency of the bias described. Furthermore, this statement is not made categorically. Instead, what is described is a *common* manifestation of bias, not a *categorical* or *general* statement of all bias. The results presented do not invalidate this statement, as the medians clearly demonstrate, but do show a range of performance. We will try to make this nuance more clear in the revision of this manuscript.

Final Author Response 6: We have added an explicit caveat that our claim is not of a universal truth, but rather of a tendency (page 2, line 16). We have also expanded our discussion of the variability of performance starting on page 13, line 27.

Reviewer Comment 7: *P2. Ln 22. The difference / novelty of the present methodology with that of Fennessey (1994) and Hughes and Smakhtin (1996) is not clear. Please clarify.*

Author Response 7: We will attempt to clarify this in the revision. Fennessey (1994) and Hughes and Smakhtin (1996) presented a method to simulate streamflows using a donor gauge for timing and a regionalized regression for FDCs. The

approach presented in the current work is an extension where the timing is generated from a process-based model (rather than a donor).
Final Author Response 7: The question of novelty is further articulated on page 6, line 24.

Reviewer Comment 8: *P3. The description of the content of the paper is missing at the end of the introduction.*
Author Response 8: We will add a paragraph describing the headers of each section as a signpost to the reader.
Final Author Response 8: This paragraph has been added on page 3, line 11.

Reviewer Comment 9: *P3. Simulated discharges. Â´n initial predictions of daily streamflow values for each streamgauge were obtained by applying the pooled ordinary kriging approach (Farmer, 2016) to each 2-digit Hydrologic Unit". The approach has to be better explained. It is not possible to understand how time series are obtained for any given gauge from what is said in the paragraph ("It builds a single, time-invariant semivariogram model of cross-correlation that is then used to estimate ungauged streamflow as a weighted summation of all contemporary observations"). A summary methodological scheme could be helpful. Is the variogram model estimated independently for each day ? for each target location from stations in its neighborhood (or do you use all stations of the United States to estimate the daily discharge of a given station) ? What is a 2-digit Hydrologic Unit ? Does the leave-one out procedure applies for the target location ?*
Author Response 9: Because the underlying hydrologic model is not the novelty of this method nor is it the only simulation method that can be bias-corrected in this way, we chose not to go into detail on the exact simulation methodology. The approach thoroughly described in Farmer (2016) is followed identically here. The semivariance was calculated for all contemporary pairs of daily streamflow observations, after transformation by dividing by drainage area and taking the logarithm, in a 2-digit HUC. These seminvariances were then summarized with a single seminvariogram cloud, which was used to fit a semivariogram model that applies equally for all days. The 2-digit Hydrologic Units are the large regions used to classify basins in the United States and shown in Figure 1 (Seaber, Paul R., F. Paul Kapanos, and George L. Knapp (1987). "Hydrologic Unit Maps". United States Geological Survey Water-supply Papers. No. 2294: i–iii, 1–63.). These clarifications will be provided in the revision of this manuscript and we will explore the feasibility of a graphic to summarize the simulation routine.
Final Author Response 9: On page 4, line 25, the reader is directed to find further discussion of the simulation method elsewhere. The motivation for not including these details here, namely that they are published elsewhere and do not represent the novelty of this work, is also highlighted on page 3, line 22.

Reviewer Comment 10: *P4. Ln 1-8 : Please clarify. "Daily FDCs were developed independently of the streamflow simulation procedure by following a regionalization procedure similar to that of Farmer et al. (2014). You mention "These same percentiles were then estimated using a leave-one-out cross-validation of regional regression". The objective/process of this estimation was not clear. Please make clear that the FDC used to bias correct your target station is obtained from a regional FDC model, obtained from all (or a part of) the stations close to this target station.*
Author Response 10: We will provide clarification in the revision. Farmer et al. (2014) use unsupervised regional regression that relies on a best-subsets regression to estimate the complete FDC in a way that tries to capture the dependence between quantiles. The leave-one-out procedure is used to quantify performance as if the target site were completely ungauged. Regionalized regression relies on all stations within a pre-defined region (in this case, the 2-digit Hydrologic Unit).
Final Author Response 10: This clarification is made on page 5, lines 1 and 5.

Reviewer Comment 11: *P4. Explanatory variables. Please comment the possible correlations between those. Have you looked / used for uncorrelated sets of explanatory variables ?*
Author Response 11: Farmer et al. (2014) discuss how explanatory correlations were controlled. Most of this detail, being already published elsewhere, is beyond the scope of this current work. However, we also used a limitation of variance inflation factors of each model. This will be added to the revision of the manuscript.
Final Author Response 11: A discussion of explanatory variables has been added on page 5, line 14.

Reviewer Comment 12: *P4 – ln 10 : what are "best-subsets regression". Best sub-sets of what ?*
Author Response 12: Best-subsets regression is a common regression technique that exhaustively searches the predictor space for the best model with a specified number of variables. The specified number of variables is then changed to explore a range

of model sizes. As described by Farmer et al. (2014), of these models that then differ in size, the AIC (or some other metric) is used to select the "best" model in an unsupervised fashion.
Final Author Response 12: A better description of best-subsets regression has been provided on page 5, line 5.

Reviewer Comment 13: *What is the difference with "three contiguous streamflow regimes you mention in line 17.*
Author Response 13: The three contiguous regions allow for different explanatory variables to be used to predict different streamflow quantiles. Each is continguous in that it mast span a contained range of quantiles, e.g. (0.02%, 0.05%, 0.1%, etc.) and not (0.02%, 75%, 99.98%). Because three are allowed, these can be thought of as "low", "medium" and "high" streamflows. We will provide this example in the revision.
Final Author Response 13: The discussion of these regimes has been improved with additions on page 5, line 20.

Reviewer Comment 14: *P4; Please clarify. "the percentiles were grouped into a maximum of three contiguous streamflow regimes based on the behavior of the unit FDCs in the region". What is a unit FDC ? How are defined the regimes ? Are they relative to different parts of the FDCs or to different sets of different basins within the region ? Do you group 95th and 90th percentiles for instance ?*
Author Response 14: The unit FDC is the duration curve of streamflow divided by drainage area; this definition will be provided in the revision. Based on a national analysis of unit FDCs (Over et al., in press [expected to be published before revision]), it is possible to identify low, medium and high streamflow regimes. This explanation will be provided.
Final Author Response 14: The discussion of these regimes has been improved with additions on page 5, line 20.

Reviewer Comment 15: *P4. "The percentiles in each regime were estimated by the same explanatory variables, allowing only the fitted coefficients to change.." Do you mean that a regression is estimated for each of the Twenty-seven percentiles you considered ? If yes, a very significant risk on a non monotonic behavior of the FDC is likely. Why not working with a analytical model fitted to the FDC (and then propose a regionalization of the parameters of this model ?). This issue should be discussed somewhere.*
Author Response 15: The answer to the reviewers question depends on how the reviewer is defining "regression". Within a given regime, as defined earlier, all of the quantiles have the same explanatory variables in the final equation. However, the coefficients on those variables are fit independently across quantiles. Because the same variables are used, monotonicity is made much more likely within the regime; the only nonmonotonicity therefore occurs at the separation of regimes. We will add a discussion of this point. Furthermore, identifying the optimal method for FDC prediction is beyond the scope of this work. For this reason, we did not explore analytical solutions for the FDC (an aside: https://doi.org/10.5194/hess-21-3093-2017 found it is very difficult to find a suitable analytical solution).
Final Author Response 15: On page 5, line 32, we have added a discussion of the fitting of coefficients and potential issues of non-monotonicity.

Reviewer Comment 16: *P4. "Further details on the approach of percentile grouping this methodology can be explored in the associated data and model archive Farmer et al. (2018)." The given reference is a technical report. The soundness of the ""percentile grouping approach" is thus uncertain. It has thus to be fully explained in the present manuscript or it has to be described in another a peer reviewed journal.*
Author Response 16: The report that the reviewer is looking for is in press and a citation will be added (Over et al., in press [expected to be published before revision). The data release includes all the development code, allowing a user to explore the method in depth, if interested.
Final Author Response 16: The technical report allows the reader to explore the methods further through experimentation. As they are not the focus nor the foundation of this work, they are not explored here. A publication exploring this methodology is currently in press and expected to be released before publication. Its complete citation has been included in the revision.

Reviewer Comment 17: *P4. The section on the bias correction approach has to be made clear. Especially how the bias correction works and with which data in the case of ungauged catchments. Again a synthetic graphical scheme of the BC approach would be usefull. You say "The nonexceedance probabilities were then converted to standard normal quantiles*

and linearly interpolated along two types of independently estimated FDCs: the regionally regressed FDCs and the observed FDCs determined by applying the Weibull plotting position." It is something to do with quantile mapping? I can imagine that a correction function can be established for any given station where you have both observations and simulations. Is the idea here to transfer the correction function obtained for a gauged catchment to neighboring ungauged ones ? If yes, how do you define the neighborhood ? how do you consider the different corrections functions you can obtained for the different gauges stations you may have in the neighborhood of your target location ?

Author Response 17: We will clarify in the revised manuscript that the idealized method that uses observed FDCs is only an example to demonstrate theoretical utility. In practice, only an estimated FDC can be used. There is no development or transferal of correction functions. We will provide a graphical demonstration of the methodology.

Final Author Response 17: On page 6, line 17, we have improved the discussion of the methods by using a figure. On line 9, we clear up some of the confusion with respect to observed FDCs.

Reviewer Comment 18: *P4. Ln – 29/30 : "The nonexceedance probabilities were then converted to standard normal quantiles and linearly interpolated along two types of independently estimated FDCs: the regionally regressed FDCs and the observed FDCs determined by applying the Weibull plotting position." » This is confusing. In the ungauged basin configuration : only one FDC is expected to be used : the regionalized one. The observed FDC is not expected to be available in a ungauged catchment. You use it here only to estimate the added value of the quality of the regionalized FDC on the resulting bias corrected discarhge time series.The "independently estimated FDC" you mention in line 31 (and basically everywhere else in your manuscript) should first refer, I guess, to the regionalized FDC (and not to the FDC from observations).*

Author Response 18: The reviewer is correct, as we have addressed in other comments. The use of an observed case is merely for the purpose of demonstration. The reason we differentiate between "independently estimated FDC" and "regionalized FDC" is that regionalization is not the only way to estimate an ungaged FDC. This will be clarified.

Final Author Response 18: This clarification has been made on page 6, line 8, and elsewhere.

Reviewer Comment 19: *P5. Ln 3: reformulate : "by correcting the simulated volumes to an independently estimated FDC.*

Author Response 19: We will revise to "by *rescaling* the simulated volumes to an independently estimated FDC."

Final Author Response 19: This change appears on page 6, line 15.

Reviewer Comment 20: *P5. Evaluation : Ln 10- the two different evaluations approaches were not clear for me at first. A reformulation would be worth Ln 20 & follow. Evaluation for the whole FDC, or for a given tail : what is the evaluation crietiron : the mean of the bias estimated between each pair of percentiles ? the biais between the mean of the percentiles for the raw and corrected data respectively ?? Evaluation on observation-independent tails and observation dependant tails. A graphical scheme please to explain what is done in the second case, at least in a supplementary material !*

Author Response 20: We will provide a graphical scheme of the two types of tails in the revision. Several criterion are provided for evaluation (e.g., average differences in common logarithms, RMSE, NSE). The median across all sites is taken as the average performance for our data set. There is, of course, some spread around the central tendency, and this is discussed, but using the central tendency is a traditional means of assessing bias.

Final Author Response 20: Upon reconsideration, the addition of a figure showing the two types of tails was not an efficient use of space. We have revised the evaluation section on page 6 and believe the techniques for evaluation have been made more understandable.

Reviewer Comment 21: *Results : all results are given in the form of boxplots. This is likely not enough to understand how the methodology work and how good it is. To give for a selection of stations the different curves (observed / regressed / regressed+bias corrected) would be helpful (with a good performance station and a bad one for instance)*

Author Response 21: We will provide a figure of some example cases in the revision.

Final Author Response 21: Given there are over 1000 sites considered, the addition of figures exploring the range of example cases was not an efficient use of space. Instead, the figure of methodology (2), a map of original bias (figure 7) and the correspondence of tail bias (figure 10) provide the context the reviewer is looking for.

Reviewer Comment 22: *P5 ln33 and p7 . Ln 25. I find the term "timing" and "error in timing" not appropriate. You could perhaps say " an error in timing of the percentiles". This is however more an error in the temporal structure of your simulated time series. This results from an error in volume which is one day an over-estimation of the true volume and the day after an under-estimation.*

Author Response 22: We will consider how to clarify this. "An error in timing of the percentiles" is misleading, as it is not the percentiles that provide timing. Timing is provided by the relative ranks and nonexceedance probabilities. By "timing error", we intend to refer to the "error in the temporal structure". We will use this formulation in the revision. It is true that a transposition of the volumes could appear as a timing error, but this is not explored here (a discussion point will be added).

Final Author Response 22: In several places, we have removed this reference to timing. Examples are on page 1, line 6 and pages 7, 10 and 11.

Reviewer Comment 23: *P7 ln34. You say "These timing errors also almost result in errors in a particular direction: low for high flow and high for low flows". You have perhaps such a mean behavior but as mentioned above, you have a number of low flows that are overestimated but also a number of low flows (less frequent) that are underestimated. . . Please reformulate to put this statement in perspective.*

Author Response 23: We will clarify that we are not making a general statement, but observing a central tendency amid a range of values.

Final Author Response 23: This clarification was made on page 10, lines 24 and 26.

Reviewer Comment 24: *P6. Ln 2 : "Figure 4 and Table 1 summarize the tail bias in all approaches to streamflow simulation considered here.". What are the 3 different approaches ? This has to be clearly explained previously. The BC-Obs is not an "approach" similar to the 2 others as you do not know in principle the observed discharge for the target ungauged basin. It just allows you to identify the influence of the quality of the regionalised FDC. A reformulation is required when relevant in all the manuscript. The presentation of the method has probably also to be restructered to make it clearer.*

Author Response 24: This was addressed in a previous comment. A graphical presentation will be provided to distinguish between observation-dependent and observation-independent tails. The BC-Obs is provided for exactly the reason the reviewer describes: not as a viable alternative for application.

Final Author Response 24: As with comment 3, at several places in the manuscript, we have made it clearer that we intend the use of observed FDCs to be an idealized analysis, not an application. For examples, see page 3, line 15, page 8, line 7, and page 9, line 15. With this improvement, a graphical representation of observation-dependent and observation-independent tails was not deemed efficient.

Reviewer Comment 25: *P7. 1. The analysis of the second paragraph in this section 3.1 is clumsy. The Observed FDC is in principle perfect and thus the bias in simulation for the observation independent evaluation case should fully vanish after correction. You should have nothing or roughly nothing to comment here. The results should be perfect. Why is there some remaining bias with the BC-obs approach ?? Please comment. (could it be a difference in the time period used for the simulation and the time period used to identifiy the observed FDC? Is it something related to the epsilon value you add to discharge data for the logarithmic transformation issue ? to the reduced number of percentile used to describe the FDC ? something else ?) The comment on this in the conclusion section has also to be modified accordingly (p12 – ln 1/10). P7. 24: "To understand the effect of errors in timing further, consider Figure 6, which shows the mean error in the nonexceedance probabilities of the observation-dependent upper and lower tails." I can not understand (just guess) what is refered to here ? Please clarify.*

Author Response 25: The use of an observed FDC will still rely on the timing of the simulated hydrograph. It is for this reason that you are likely to see some residual bias in the observation-dependent tails (Fig. 4, boxplots C and F). With the observation-independent tail (Fig. 4, boxplots I and L), the much smaller residual bias is likely a result of the interpolation along an FDC defined by a finite set of quantiles. There may be some effect from the epsilon value and censored regression used for zero-valued streamflows, but we did not find a major impact. We will add a note on this discussion.

Final Author Response 25: This discussion was added on page 10, line 18. The figure is now Figure 8.

Reviewer Comment 26: *Fig 2 and following : how many data are used for each boxplot (no. stations x no. percentiles ???). Is there one point for each station/percentile ?*

Author Response 26: There are 1168 points in each boxplot (one for each station); we will make this clearer in the caption. The metric is the mean difference in common logarithms of the complete FDC at a given site.
Final Author Response 26: The captions clearly define the number of sites being presented. The revisions of the evaluation section help to better understand the performance metrics used here.

Reviewer Comment 27: *Discussion : what about the likely seasonal dependence of the correction function ? Is there some potential for improvement here ? The estimated FDCs are composed of 27 quantiles, of which the upper and lower tails contain only the eight values with nonexceedance probabilities 95% and larger and 5% and smaller, respectively.). A comment on the number of quantiles used to describe the FDC would be worth (a sensitivity analysis of the results to this number could be also included in a discussion section)*
Author Response 27: No correction function, as such, was used. (It could be conceived as a function, but that is not explored or proposed here.) There is, of course, always room for improvement; we provide a discussion of the sensitivity to the representation of the FDC. We will strengthen that discussion by noting the limitations of an FDC defined by a finite number of quantiles.
Final Author Response 27: This discussion was added on page 13, line 9.

Reviewer Comment 28: *What is the influence of the duration of the observation time period used to estimate the observed FDC > on the quality of the FDC estimation and then on the quality of the bias correction ?*
Author Response 28: As the period of analysis is fixed here, it is not possible to explore this effect. We will make a recommendation for future work, as it will certainly have an impact on something like future projections.
Final Author Response 28: This addition was made on page 13, line 9.

Reviewer Comment 29: *Is there any dependence of the results to the hydroclimatic context of the basin ? How is it structured is space in US ?*
Author Response 29: We could not see any obvious pattern: "Initial exploration did not find a strong regional component to performance of the bias correction method. For some regions, like New England, where FDCs are well estimated by regional regression, there is a general improvement in accuracy under bias correction with regionally regressed FDCs, but the improvement is highly variable. Instead, the strongest link is with the reproduction of the FDC. " It may be that the hydroclimate is driving the ability to reproduce the FDC through regression, but that is left for future research.
Final Author Response 29: We have strengthened this discussion on page 13, line 13 and on.

Reviewer Comment 30: *Minor remark : P2. Ln10. The interest of the "long term forecast term" here is not clear. It seems to be out of the scope of the work. To be better explained. What is long term ?*
Author Response 30: Exploration of long-term forecast is beyond the scope of this work and no effort is made to explore it here. Long-term (decadal and beyond) forecasts are mentioned as an example of a hazard of underlying bias. We will clarify this statement by defining long-term as decades and beyond.
Final Author Response 30: We added this specification on page 2, line 14.

Reviewer Comment 31: *P2. 15/20 : » this paragraph is not clear »> to be clarified / double-formulate. The "interpolation of non-exceedance probabilities along the FDC" is a rather clumsy formulation. What does it mean ?*
Author Response 31: Described further in the methodology, the interpolation must occur because the FDC is being represented as a finite set of quantiles (27). If a no exceedance probability does not fall exactly on one of those percentiles (P4, L10), it must be interpolated. We will reference the methodology in the revision.
Final Author Response 31: The reader has been directed to the discussion of methods on page 2, line 27.

Reviewer Comment 32: *p2. Ln 33 : please clarify what is meant there*
Author Response 32: As before, if the nonexceedance probability falls within the range of the quantiles (P4, L10), it can be interpolated. However, it is falls outside of the range of quantiles (P4, L10), it must be extrapolated. The two nearest points were used for linear extrapolation. We will consider how to clarify.

Final Author Response 32: Our original response unintentionally missed the mark here. By moving the paragraph referenced by the reviewer to the discussion, the relationship with previous methods can be more easily understood.

Reviewer Comment 33: *P2. Ln 35 and following. This does not belong to the introduction but to a discussion section. The discussion should probably give the evaluation of the other method suggested here (Additional research to explore if estimating nonexceedance probabilities directly, as opposed to the conversion of simulated streamflow to nonexceedance probabilities used here, might further improve nonlinear spatial interpolation using FDCs or simulation more generally*
Author Response 33: This will be moved to the discussion. Further evaluation is outside the scope of the present work.
Final Author Response 33: This was moved to the discussion on page 14, line 31.

Reviewer Comment 34: *p3. Ln 3 : give the structure of the paper*
Author Response 34: This will be added, recapitulating the headers of the document.
Final Author Response 34: This was added on page 3, line 11.

Reviewer Comment 35: *p5. "The root-mean-squared error of the common logarithms of streamflow and the differences therein were used to quantify accuracy." Do you mix streamflow and differences between streamflow in the computation of a single RMSE criterion ? If yes, I fear it is not relevant or please clarify / justify.*
Author Response 35: We do not mix streamflow and differences between streamflow in the computation of a single RMSE. The RMSE is calculated for two different approaches; we then observed the differences across approaches. We agree that this setnence is confusing, so will explore other options. One possible revision could be "The differences in the root-mean-squared error of the common logarithms of the predicted streamflow for the two approaches were used to quantify accuracy."
Final Author Response 35: On page 5, line 1, we have added formulas to help the reader understand the computations.

Reviewer Comment 36: *P6. Ln 9 – add a subsection title "simulated hydrographs without correction"*
Author Response 36: This section will be added.
Final Author Response 36: This has been added on page 8, line 21.

Reviewer Comment 37: *P6 – 25. "These results show upward bias in lower tails and downward bias in upper tails." No, this is not the case in general. See your paragraph above.*
Author Response 37: We do not claim that the result is general, but the result do show a tendency in the direction described. We also discuss the variability of this performance, but this will be further highlighted in the revision.
Final Author Response 37: On page 9, line 10, we explicitly state that we are making observations on central tendency and not claiming universality.

Reviewer Comment 38: *P9 "For the observation-independent case, the errors are removed almost completely, and the remaining errors in the observation-dependent case mimic the timing (nonexceedance probability) errors." This is not true (only if the observed FDC is used)*
Author Response 38: The first half of this paragraph only discusses the case where observed FDCs are used. We will clarify this transition.
Final Author Response 38: Clarification was added on page 12, line 13.

Reviewer Comment 39: *P10. 2 : "The change in the absolute bias of the observation-independent lower tail has a 0.72 Pearson correlation with the absolute bias of the lowest eight percentiles of the FDC estimated with regional "regression." I do not understand what is meant here.*
Author Response 39: We will add, "That is, the residual bias in the bias-corrected FDC is strongly correlated with the bias in the independently-estimated FDC."
Final Author Response 39: This has been added on page 13, line 5.

Reviewer Comment 40: *Figures P10- 6 "as regional regression is not the only tool for estimating FDCs, improved methods for DC estimation would further increase the impact of this bias correction procedure." Mention other such methods.*

Author Response 40: Moving and expanding the last paragraph of the introduction will allow us to discuss other methods. Some might include TNDTK, kriging methodologies, index-flood methods, other hydrograph simulations, etc.
Final Author Response 40: This was added on page 6, line 3; page 13, line 9; and page 15, line 1.

Reviewer Comment 41: *P10.15 "It may not always be possible to determine the accuracy with which a given FDC estimation technique might perform, making it difficult to determine if these results can be generalized." There is no reason why the accuracy of any FDC regionalization approach could not be assessed (this is done in all FDC regionalization study). At least with a leave-one out procedure. Before applied for the bias correction of any time series simulation model, this quality should be checked and estimated to be better than that of the simulation model.*
Author Response 41: We will revise to, "*At a particular ungaged location,* it may not always be possible to determine the accuracy with which a given FDC estimation technique might perform *(beyond a regional cross-validation assessment)*, making it difficult to determine if these results can be generalized."
Final Author Response 41: The revision can be seen on page 13, line 23.

Reviewer Comment 42: *Fig. 1 – What is meant here : "The outlines of 2-digit Hydrologic Units are provided for further context.*
Author Response 42: The polygons represent the 2-digit Hydrologic Units, which are the regionalization areas. The meaning of these units is described in an earlier comment and will be incorporated into the revised manuscript.
Final Author Response 42: A description of 2-digit hydrologic units has been added on page 4, line 18.

Reviewer Comment 43: *Fig. 2 : BC-RR and BCObs have to be defined in the main text.*
Author Response 43: We will consider adding these into the report. They are not currently used in the main body, as the repetition of opaque acronyms might detract from clarity.
Final Author Response 43: We considered using these in the report, but felt that remove some degree of clarity. The acronyms are defined in teh figure captions alone so that the figures might be understood independently.

Reviewer Comment 44: *Figures and tables : simplify the captions : a number of repetitions could be removed (and a reference to the caption details of one reference figure added in the caption of all other figures)*
Author Response 44: We prefer that the captions allow each figure to standalone as much as possible.
Final Author Response 44: We have retained the formulations in the captions to provide the most clarity for each figure independently.

Reviewer Comment 45: *Are the 3 tables usefull ?*
Author Response 45: We feel they are useful, as they provide a summary of the numbers discussed in the report.
Final Author Response 45: The tables have been retained.

3 Third Anonymous Reviewer

Reviewer Comment 1: *The authors presented a bias-correction procedure useful for improving the accuracy of simulated daily streamflow series by using independently estimated flow-duration curves (FDCs). Although the procedure itself is not completely new (references of previous studies are included in the manuscript), the study is interesting as it considers an extended database in the US and focuses on the reproducibility of upper and lower tails, distinguishing between observation-dependent and observationindependent tails. This aspect is meaningful for highlighting the effect of timing on distributional bias. The study concludes that the significant potential of the bias-correction procedure is limited by the accuracy of the FDCs estimation method.*

The paper is well structured and written, but some changes should be applied for making it more readable and for emphasizing some important aspects. I believe it is suitable for publication in HESS after the authors address some issues reported in the comments below.

Author Response 1: Thank you for your deep consideration of our manuscript. Your kind comments will surely help us to improve the delivery of our results.

Final Author Response 1: We wish to again extend our thanks. Your comments were very useful in thinking about how to better describe our methods and present our findings. The revised version of our manuscript is certainly much stronger because of your input.

Reviewer Comment 2: *I would suggest to explain more in details the "Bias correction" procedure at Section 2.3. For instance, at Page 4 Lines 29-30, the sentence "[. . .] linearly interpolated along two types of independently estimated FDCs" is rather ambiguous: I found difficult to understand whether the authors refer to the resampling of the curves, or perhaps to the prediction of FDC quantiles, which is carried out with a linear regression. I would recommend to rephrase this sentence and add more information to it, in order to clarify better this fundamental aspect of the proposed procedure. Moreover, at least one figure could be useful for clarifying the procedure. I can suggest to show at least two plots, where the authors may report standard normal quantiles vs. logarithmically transformed streamflow percentiles for, in turn, (a) regionally regressed FDC and pooled ordinary kriging curve, and (b) observed FDC and pooled ordinary kriging curve. Finally, in my opinion, the bias-correction section should be extended: I recommend to add more detailed information about how the bias correction is applied to the simulated streamflows, maybe introducing a figure vignette, or, likewise, describing the procedure point by point.*

Author Response 2: In concert with the response of other reviewers, we will be reformatting the section on methodology. It will be substantially enhanced by figures that show example hydrographs and example FDCs, as well as a figure demonstrating the steps of the procedure.

Final Author Response 2: The section on methodology has been revised. In particular, page 6, line 17, of the tracked-changes document presents a figure more completely describing the methodology for bias correction.

Reviewer Comment 3: *I would stress more that in the majority of possible practical applications of the proposed method (i.e. predictions in ungauged sites), using observed FDCs for the bias correction would not be possible. Indeed, the only exception could be represented by those catchments in which we want to simulate streamflows for a given period, even though we have streamflow data for another period. I would add this reasonings to the revised version of the paper.*

Author Response 3: We will add clarification to point out that the use of observed FDCs is only provided as a theoretical demonstration of the upper-limit of performance in ungauged locations. We did not consider use of observed FDCs for record extension, but this approach might be useful in partially gauged locations. Though we did not explore that particular application, we will mention it in the discussion.

Final Author Response 3: In several places we have highlighted the theoretical utility of observed FDCs, noting that they are not intended for application (page 3, line 15; page 8, line 7; and page 9, line 15.). Furthermore, we have acknowledge the potential for record extension on page 8, line 7.

Reviewer Comment 4: *[3] Page 10, Lines 11-13 – The authors highlight that "initial exploration did not find a strong regional component to performance of the bias correction method". In order to better support this sentence and to improve the effectiveness and completeness of the study, the authors could better discuss the spatial distribution of performance, especially given the high climatic variability among the conterminous United States. Therefore, I would suggest to add and discuss a new figure (or figures; e.g. a set of maps, similar to Figure 1), showing the spatial distribution of bias and root mean squared error in the study region for at least a couple of the cases considered in the study (Orig., BC-RR, BC-Obs.; upper tail, lower tail; observation-dependent tail, observation-independent tail; sequential, distributional, etc.).*

Author Response 4: We will develop such a figure.

Final Author Response 4: Figure 7 contains a map of the original bias. Figure 10 shows how this bias in each tail might change under bias correction. Maps were not produced for all cases, as they would have been largely redundant with these additional figures.

Reviewer Comment 5: *In Figure 1, the differences between not-selected and selected streamgauges are not clear: in some areas, crosses overlap with points and the distinction is not simple. Maybe using different colors and symbol sizes might highlight better the differences between these two categories.*

Author Response 5: We will work with the journal to improve the visibility of this figure.
Final Author Response 5: We have changed the symbols to showed only included sites, but will continue to consult with the journal.

Reviewer Comment 6: *Page 3, Line 24 – It is not clear what a 2-digit Hydrologic Unit is. Could you please explain?*
Author Response 6: The Hydrologic Unit system is a common method for delineating watersheds in the US. 2-digit hydrologic units (the polygons in Figure 1) roughly align with the major river basins of the United States. We will add this description with appropriate citations. In the figure, the units are the outlined polygons. (Seaber, Paul R., F. Paul Kapanos, and George L. Knapp (1987). "Hydrologic Unit Maps". United States Geological Survey Water-supply Papers. No. 2294: i–iii, 1–63.)
Final Author Response 6: Hydrologic Units are defined on page 4, line 18.

Reviewer Comment 7: *Page 3, Line 16 (and elsewhere) - "cubic feet per seconds (cfs)" is used. I believe that the International System of Units should be used in scientific papers. At the same time, if cfs is the standard adopted by USGS (and in the GAGES-II database), I think that at least the conversion factor to m3 s-1 should be indicated in parentheses the first time that cfs is mentioned.*
Author Response 7: We will add a conversion factor.
Final Author Response 7: The conversion factor was added on page 4, line 8.

Reviewer Comment 8: *Page 4, Line 3 - It seems that you are referring to period-of-record FDCs; am I correct? Could you please state that explicitly?*
Author Response 8: Yes, you are correct. We will state this more clearly.
Final Author Response 8: This was added on page 2, line 26, and page 4, line 29.

Reviewer Comment 9: *Page 4, Line 10 - It is not clear to me what "best-subsets regression" is. Could you please clarify and/or add at least one reference?*
Author Response 9: Best-subsets regression is a common regression technique that exhaustively searches the predictor space for the best model with a specified number of variables. The specified number of variables is then changed to explore a range of model sizes. As described by Farmer et al. (2014), of these models that then differ in size, the AIC (or some other metric) is used to select the "best" model in an unsupervised fashion.
Final Author Response 9: Best-subsets regression is further defined on page 5, line 5.

Reviewer Comment 10: *Page 4, Lines 13-15 and 20-22; Page 8, Lines 19-21 - I would suggest to remove parentheses.*
Author Response 10: We will remove them.
Final Author Response 10: These were removed on page 5, line 11, 14, 26 and 28; and page 11, lines 15 and 16.

Reviewer Comment 11: *Page 5, Lines 9-10 – I would recommend to add an equation showing the expression "ten to the power of the difference and subtracting one from this quantity". I would also suggest to explain why you are referring to this equation for computing the percentage. Moreover, the authors could show equations for bias and root mean squared error, respectively, when introducing them.*
Author Response 11: We will add these equations and any relevant citations.
Final Author Response 11: These equations were added on page 7, line 1.

Reviewer Comment 12: *Page 5, Line 10 – The authors write "root-mean-squared error", while use "root mean squared error" in the remainder of the text. I would suggest to use the same expression everywhere ("root mean squared error" should be fine).*
Author Response 12: We will make every effort to ensure consistency throughout the manuscript.
Final Author Response 12: This was changed on page 7, line 8.

Reviewer Comment 13: *Page 5, Line 13 – Could you please add a reference for the Wilcoxon signed rank test?*
Author Response 13: We will add this citation: Wilcoxon, Frank (Dec 1945). "Individual comparisons by ranking methods". Biometrics Bulletin. 1 (6): 80–83. doi:10.2307/3001968.

Final Author Response 13: This citation was added on page 7, line 13.

Reviewer Comment 14: *In Tables' captions, the acronyms "OD" and "OI" are used instead of "observation-dependent" and "observation-independent", respectively. You could use these acronyms also in the body of the text, in order to improve the readability; e.g. "OD-tail" instead of "observation-dependent tail" and "OI-tail" instead of "observation-independent tail".*

Author Response 14: A previous reviewer suggested also using BC-Obs, etc. We will attempt to use these abbreviations, and ensure that clarity is not harmed.

Final Author Response 14: After using these abbreviations, we felt they removed a bit of clarity from the prose. We decided not to use abbreviaitons and keep the explicit descriptions.

Reviewer Comment 15: *Page 10, Lines 6-7 – Could you please report some other methods (please provide references) for estimating FDCs?*

Author Response 15: Moving and expanding the last paragraph of the introduction will allow us to discuss other methods. Some might include TNDTK, kriging methodologies, index-flood methods, other hydrograph simulations, etc.

Final Author Response 15: This is appended on page 15, line 1.

Reviewer Comment 16: *Page 11, Line 16 - Please delete "Summary and conclusions".*

Author Response 16: Yes, we will fix this error.

Final Author Response 16: This has been revised on page 15, line 3.

Reviewer Comment 17: *Captions of Figures 2, 3, 4, 5 – Please remove comma in "pooled, ordinary kriging".*

Author Response 17: Commas will be removed.

Final Author Response 17: Commas were removed in the figure cations.

Reviewer Comment 18: *In the body of text, there are some references to a recent study by Pugliese et al. (2017), which is currently under review. I would suggest to update these references after its possible acceptance/publication.*

Author Response 18: We have been keeping an eye on this publication and will update the references when a decision on that manuscript is made.

Final Author Response 18: At the time of writing, this work has not been published. We will continue to monitor its progress in hopes that it is published before this work goes to press.

Bias correction of Simulated Historical Daily Streamflow at Ungauged Locations by Using Independently Estimated Flow-Duration Curves

William H. Farmer[1], Thomas M. Over[2], and Julie E. Kiang[3]

[1]U.S. Geological Survey, Denver, Colorado, United States
[2]U.S. Geological Survey, Urbana, Illinois, United States
[3]U.S. Geological Survey, Reston, Virginia, United States

Correspondence: William H. Farmer (wfarmer@usgs.gov)

Abstract. In many simulations of historical daily streamflow distributional bias arising from the distributional properties of residuals ~~, however small,~~ has been noted. This bias often presents itself as an underestimation of high streamflow and an overestimation of low streamflow. Here, 1168 streamgages across the conterminous United States having at least 14 complete water years of daily data between October 01, 1980, and September 30, 2013, are used to explore a method for rescaling simulated streamflow to correct the distributional bias. Based on an existing approach that separates the simulated streamflow into components of ~~timing~~ temporal structure and magnitude, the ~~timing component is converted into~~ temporal structure is converted to simulated nonexceedance probabilities and ~~rescaled to new volumes~~ the magnitudes are rescaled using an independently estimated flow-duration curve (FDC) derived from regional regression. In this study, this method is applied to a pooled ordinary kriging simulation of daily streamflow coupled with FDCs estimated by regional regression on basin characteristics. The improvement in the representation of high and low streamflows is correlated with the accuracy and unbiasedness of the estimated FDC. The method is verified by using an idealized case~~, though, ~~; however, with the introduction of regionally regressed FDCs developed for this study, the method is only useful overall for the upper tails, which are more accurately and unbiasedly estimated than the lower tails. It remains for future work to determine how accurate the estimated FDCs need to be to be useful for bias correction without unduly reducing accuracy. In addition to its potential efficacy for distributional bias correction, this particular instance of the methodology also represents a generalization of nonlinear spatial interpolation of daily streamflow using FDCs. Rather than relying on single index stations, as is commonly done to reflect streamflow timing, this approach to simulation leverages geostatistical tools to allow a region of neighbors to reflect streamflow timing.

1 Introduction

Simulation of historical daily streamflow at ungauged locations is one of the grand challenges of the hydrological sciences (Sivapalan, 2003; Sivapalan et al., 2003; Hrachowitz et al., 2013; Parajka et al., 2013). Over the past 15 years, research into simulation of historical streamflow has increased. In addition to ongoing international efforts, the U.S. Geological Survey has embarked upon a National Water Census of the United States (Alley et al., 2013)~~seeking~~ , which seeks to quantify hydrology

across the country to provide information to help improve water use and security. However, regardless of the method used for the simulation, uncertainty will always remain and may result in some distributional bias (Farmer and Vogel, 2016). The objective of this work is to present a technique to correct for bias in the magnitudes of a streamflow simulation. While the mechanics of this technique are not novel, the novelty of this work lies in the generalization of this technique for use in bias correction. The method is intended for use at ungauged sites and an idealized experiment is constructed to demonstrate both the potential utility and one example of realized utility.

As defined here, distributional bias in simulated streamflow is a failure to reproduce the tails of streamflow distribution. As attested to by many researchers focused on the reproduction of historical streamflow, this bias commonly appears as a general overestimation of low streamflow and underestimation of high streamflow (Skøien and Blöschl, 2007; Rasmussen et al., 2008; Farmer et al., 2014, 2015; Farmer, 2016; Farmer and Vogel, 2016; Archfield et al., 2010, 2013). The result is an effective squeezing of the streamflow distribution. This distributional compaction is often most notable in the downward bias of extreme high-flow events (as in, e.g., Lichty and Liscum, 1978; Thomas, 1982; Sherwood, 1994). ~~This bias~~ Bias of high streamflows is particularly concerning, as examinations of extreme high-flow events are a common and influential use of historical simulation and long-term (decadal) forecast. Consider, for example, the motivation for work by Archfield et al. (2013)~~.~~ : As simulated streamflows were being routed through a reservoir operations model for flood mitigation, large bias in high streamflows would have severely affected resulting decisions. Of course, this tendency towards distributional compaction is not a categorical truism; the resulting bias will vary widely depending on the structure of the residuals (Farmer and Vogel, 2016).

Because of the importance of accurately representing extreme events, it is necessary to consider how the distributional bias of streamflow simulations can be reduced. The approach presented here ~~uses an independently estimated flow-duration curve (FDC) to rescale estimates from a simulation of historical daily streamflow~~assumes that, while the streamflow magnitudes of a historical simulation are biased, the temporal structure or rank-order of the streamflows is relatively accurate. The nature of this approach is predicated on an assumption that although a historical simulation may produce a distribution of streamflow with biased tails, the sequence of relative rankings or nonexceedance probabilities of the simulated streamflow retains valuable information. With this assumption, it can be hypothesized that distributional bias can be reduced, while not negatively impacting the overall performance, by applying a sufficiently accurate independently estimated representation of the ~~FDC~~ period-of-record flow duration curve (FDC) to rescale the streamflow simulations by interpolating the nonexceedance probabilities of the simulated streamflow along the FDC (see Material and Methods, below).

~~This approach~~ The approach presented here can be perceived as a generalization of the nonlinear spatial interpolation of daily streamflow using FDCs as conceived by Fennessey (1994) and Hughes and Smakhtin (1996) and widely used thereafter (Smakhtin, 1999; Mohamoud, 2008; Archfield et al., 2010; Shu and Ouarda, 2012). As traditionally applied, nonlinear spatial interpolation proceeds by simulating nonexceedance probabilities at a target location using a single neighboring streamgage (though Hughes and Smakhtin (1996) recommend and Shu and Ouarda (2012) test the use of multiple streamgages) and then interpolating those nonexceedance probabilities along a FDC. The approach tested here seeks to bias-correct a simulated daily hydrograph, and, when viewed in another way, presents a novel form of nonlinear spatial interpolation.

Furthermore, though necessarily explored in this study through the use of a single technique for hydrograph simulation, this approach may be a means to effectively bias-correct any simulation of streamflow, including those from rainfall-runoff models, as presented by Pugliese et al. (2017). Pugliese et al. (2017) used a geostatistical tool to produce site-specific FDCs and then used this information to post-process simulated hydrographs from a deterministic model. Though the underlying methods of producing the FDC and simulated hydrograph are different, the approach proposed by Pugliese et al. (2017) is the same as that explored here. Further discussion of the relationship of the approach presented here to others in the field is provided below.

~~Additional research to explore if estimating nonexceedance probabilities directly, as opposed to the conversion of simulated streamflowto nonexceedance probabilities used here , might further improve nonlinear spatial interpolation using FDCs or simulation more generally. Although the results presented here are promising, they demonstrate that the performance of two-stage modeling, where timing and magnitude are largely decoupled, is limited by the less well performing stage of modeling~~The remainder of this work is organized in the following manner. The section titled "Material and Methods" provides a description of the retrieval of observed streamflow, the estimation of simulated streamflows, the calculations of observed FDCs, the estimation of simulated FDCs, and the application and evaluation of the bias-correction. A section titled "Results" follows and it documents the bias in the original simulated streamflows, analyzes the potential bias correction that could be achieved if it were possible to know the observed FDC at an ungauged location, and the bias correction that would be realized through an application of regional regression. The following section titled "Discussion" considers the implications of these results and hypothesizes how the methodology might be applied and improved. The major findings of this work are then summarized in the section titled "Summary and Conclusions".

2 Material and Methods

This section, which is divided into four subsections, provides a description of the methods applied here. The first subsection describes the collection of observed streamflow as well as the initial simualtion of streamflow. As the approach used here is applicable to any simulated hydrograph, the details of hydrograph simulation are not exhaustively documented. Instead, beyond a brief introduction, the reader is directed to relevant citations, as no modifications to previous methods are introduced here. The second subsection discusses the use of regional regression to define independently estimated FDCs. Again, as any method for the estimation of FDCs could be used and this application is identical to previously reported applications, following a brief introduction, the reader is directed to the relevant citations. The third subsection provides a description of how bias correction was executed, and the fourth subsection describes how the performance of this approach to bias correction was assessed.

2.1 Observed and Simulated Streamflow

The proposed approach was explored using daily mean streamflow data from the reference-quality streamgages included in the GAGES-II database (Falcone, 2011) within the conterminous United States for the period from October 01, 1980, through September 30, 2013. To allow for the interpolation, rather than extrapolation, of all quantiles considered later, streamgages were screened to ensure that at least fourteen (14) complete water years (October 01 through September 30) were available

for each record considered; 1168 such streamgages were available. The selected and not-selected reference streamgages are indicated in Figure ??1. The streamflow data were obtained directly from the website of National Water Information System (NWISWeb, http://waterdata.usgs.gov, accessed 20 Sept. 2017). For each streamgage, associated basin characteristics were obtained from the GAGES-II database (Falcone, 2011).

To control for streamflow distributions that vary over orders of magnitude, the simulation and analysis of streamflow at these streamgages is best explored through the applications of logarithms. To avoid the complication of taking the logarithm of a zero, a small value was added to each streamflow observation. The U.S. Geological Survey rounds all mean daily streamflow to two decimal places in units of cubic feet per second (cfs, which can be converted to cubic meters per second using a factor of 0.0283). As a result, any value below 0.005 cfs is rounded to and reported as 0.00 cfs. Because of this factrounding procedure, the small additive value applied here was 0.0049 cfs. While there may be some confounding effect produced by the use of an additive adjustment, as long as this value is not subtracted on back transformation, the following assessment of bias and bias correction will remain robust. That is, rather than evaluating bias in streamflow, technically this analysis is evaluating the bias in streamflow plus a correction factor. The conclusions remain valid as the assessment still evaluates the ability of a particular method to remove the bias in the simulation of a particular quantity.

Though the potential for distributional bias applies to any hydrologic simulation (Farmer and Vogel, 2016), for this study, initial predictions of daily streamflow values for each streamgage were obtained by applying the pooled ordinary kriging approach (Farmer, 2016) to each 2-digit Hydrologic Unit (figure ??1) through a leave-one-out cross-validation procedure on the streamgages within the 2-digit Hydrologic Unit. The Hydrologic Unit system is a common method for delineating watersheds in the U.S. As described by Seaber et al. (1987), the 2-digit Hydrologic Units, or regions (as seen in Figure 1), roughly align with the major river basins of the U.S. This approach considers all pairs of common-logarithmically transformed unit streamflow (discharge per unit area) at each day and builds a single, time-invariant semivariogram model of cross-correlation that is then used to estimate ungauged streamflow as a weighted summation of all contemporary observations. A spherical semivariogram was used as the underlying model form. Additional information on the time series simulation procedure is provided by Farmer (2016). Note that the choice of pooled ordinary kriging is only made as an example of a streamflow simulation method; it is not implied that the bias observed or methods applied are relevant only to this approach to simulation. Because the novelty of this work is in the application of bias correction, further details on the particular simulation method employed are left for the reader to investigate in the cited works (Farmer, 2016).

2.2 Estimation of Flow Duration Curves

Daily period-of-record FDCs were developed independently of the streamflow simulation procedure by following a regionalization procedure similar to that of Farmer et al. (2014) and Over et al. (2018). Observed FDCs were obtained by determining the percentiles of the streamflow distribution across complete water years between 1981 and 2013 using the Weibull plotting position (Weibull, 1939). Twenty-seven percentiles were considered, having with exceedance probabilities of :0.02%, 0.05%, 0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 90%, 95%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, and 99.98% were considered. The selection of streamgages with at least 14 complete water years

ensures that all percentiles can be calculated from the observed data. These percentiles derived from the observed hydrograph represent the "unknowable observation" in an application for prediction in ungauged basins. Therefore, to simulate the truly ungauged case, these same percentiles were ~~then~~ estimated using a leave-one-out cross-validation of regional regression.

A regional regression across the streamgages in each 2-digit Hydrologic Unit of each of the 27 FDC percentiles was developed using best-subsets regression. Best-subsets regression is common tool for exhaustive exploration of the space of potential explanatory variables. All models with a given number of explanatory variables are computed, exploring all combinations of variables. The top models for a given number of explanatory variables are then identified by a performance metric like the Akaike Information Criterion. This is repeated for several model sizes to fully explore the possibilities for variables and regression size. For each regression, the drainage area was required as an explanatory variable. At a minimum, one additional explanatory variable was used. The maximum number of explanatory variables was limited to the smaller of either six explanatory variables or 5% of the number of streamgages in the region, rounded up to the next larger whole number. ~~(~~The maximum of six arises from what is computationally feasible for the best subsets regression function used, whereas the maximum of 5% of streamgages was determined from a limited exploration of the optimal number of explanatory variables as a function of the number of streamgages in a region. ~~)~~ Explanatory variables were drawn from the GAGES-II database (Falcone, 2011). As documented by Farmer et al. (2014) and Over et al. (2018), a subset of the full GAGES-II dataset was chosen to avoid strong correlations. As the focus of this work is not on the estimation of the FDCs, the reader is referred to the works of (Farmer et al., 2014, 2018) and Over et al. (2018) to explore the exact procedures.

In order to allow different explanatory variables to be used to explain percentiles at different streamflow regimes, the percentiles were grouped into a maximum of three contiguous streamflow regimes based on the behavior of the unit FDCs ~~in the region. The percentiles in~~ (i.e., the FDCs divided by drainage area) in the 2-digit Hydrologic Units. The regimes are contiguous in that only consecutive percentiles from the list above can be included in the same regime; the result is a maximum of three regimes that can be considered "high", "medium" and "low" streamflows, though the number of regimes may vary across 2-digit Hydrologic Units. The percentiles in each regime were estimated by the same explanatory variables, allowing only the fitted coefficients to change. The final regression form for each regime was selected by optimizing the average adjusted coefficient of determination, based on censored Gaussian (Tobit) (Tobin, 1958) regression to allow for values censored below 0.005 cfs, across all percentiles in the regime. ~~(~~The addition of a small value was used to avoid the presence of zeros and enable a logarithmic transformation, but this does not avoid the problem of censoring. Censoring below the small value added must still be accounted for so that smaller numbers do not unduly affect the regression. ~~)~~ This approach to percentile grouping was found to provide reasonable estimates while minimizing the risk of non-monotonic or otherwise concerning behavior. Further details on this methodology can be ~~explored~~ found in the associated data and model archive ~~: Farmer et al. (2018)~~(Farmer et al., 2018) and in Over et al. (2018).

When estimating a complete FDC as realized through a set of discrete points, non-monotonic behavior is likely (Poncelet et al., 2017). If the regression for each percentile were estimated independently, non-monotonicity would be almost unavoidable. By using three regimes and keeping the explanatory variables the same within each, the potential for non-monotonicity is reduced. The greatest risk of non-monotonic behavior occurs at the regime boundaries. If the FDC used to bias-correct is not perfectly

monotonic, the effect will be to alter the relative timing of streamflows. While it would be ideal to avoid any risk of non-monotonic behavior, it is a rather difficult task. An alternative might be to consider the FDC as a parametric function, but Blum et al. (2017) demonstrate how difficult this can be for daily streamflows. Of course, the use of regional regression is not the only tool for estimating an FDC (for reviews, see Castellatin et al., 2013; Pugliese et al., 2014, 2016).

2.3 Bias Correction

To implement bias correction, the initial predictions of the daily streamflow values by the ordinary kriging approach were converted to streamflow nonexceedance probabilities using the Weibull plotting position (Weibull, 1939). The nonexceedance probabilities were then converted to standard normal quantiles and linearly interpolated along ~~two types of independently estimated FDCs: the regionally regressed FDCs and the observed FDCs determined by applying the Weibull plotting position~~an independently estimated FDC. For the linear interpolation, the independently estimated FDC was represented as the standard normal quantiles of the associated nonexceedance probabilities versus the common logarithmic transformation of the streamflow percentiles. In the case where the standard normal quantile being estimated from the simulated hydrograph was beyond the extremes of the FDC, the two nearest percentiles were used for linear extrapolation. In this way, the ordinary kriging simulations were bias-corrected, based on the assumption that the simulated volumes are less accurate than the relative ranks of the simulated values, by ~~correcting~~ rescaling the simulated volumes to an independently estimated FDC. By changing the magnitudes of the simulated streamflow distribution, this approach rescales the distribution of the simulated streamflow.

Figure 2 provides a simplified representation of this bias-correction methodology. Starting in the upper-left panel and proceeding clockwise, after simulating the hydrograph with a given methodology (pooled ordinary kriging was used here), the resulting streamflow value on a given day can be converted to appropriate non-exceedance probabilities by proceeding from point A, through point B and down to point C. Moving then from point C to point D maps the estimated non-exceedance probability onto an independently estimated FDC. Finally, the streamflow value produced at point D is mapped to the original date (point E) to reconstruct a bias-corrected hydrograph. Note that this is a simplified description: as described above a slightly more complex interpolation procedure is used for the FDCs represented by a set of discrete points.

As can be seen in Figure 2, this methodology is quite similar to that conceived by Fennessey (1994) and Hughes and Smakhtin (1996). The novelty of this work lies in its application. That is, both Fennessey (1994) and Hughes and Smakhtin (1996) imagine a case where the original hydrograph from which non-exceedance values will be drawn (upper-left panel of Figure 2) is drawn from an index station of some sort; here the temporal structure could be drawn from any technique for at-site hydrograph simulation. This generalization allows bias-correction of any hydrograph simulation.

2.4 Evaluation

The hypothesis of this work, that distributional bias in the simulated streamflow can be corrected by applying independently estimated FDCs, was evaluated by considering the performance of these bias-corrected simulations at both tails of the distribution. The differences in the common logarithms of both high and low streamflow were used to understand and quantify the

bias (simulation minus observed) and the correction thereof. That is,

$$bias_s = \frac{\sum_{i=1}^{n}(\log_{10}(\hat{Q}_{s,i}) - \log_{10}(Q_{s,i}))}{n} \quad (1)$$

where s indicates the site of interest, $\hat{Q}$ indicates the predicted streamflow, whether the original simulation or the bias-corrected simulation, Q indicates the observed streamflow, and n indicates the number of values being assessed. This difference can be approximated as a percent by computing ten to the power of the difference and subtracting one from this quantity ~~. The root-mean-squared~~ (Eng et al., 2009):

$$bias_{s,\%} = 100 \cdot (10^{bias_s} - 1) \quad (2)$$

The differences in the root mean squared error of the common logarithms of ~~streamflow and the differences therein~~ the predicted streamflow were used to quantify ~~accuracy.~~ improvements in accuracy. The root mean squared error of the common logarithms of streamflow is calculated as

$$rmsel_s = \sqrt{\frac{\sum_{i=1}^{n}(\log_{10}(\hat{Q}_{s,i}) - \log_{10}(Q_{s,i}))^2}{n}} \quad (3)$$

Improvements in accuracy may or may not occur when bias is reduced. The significance of both these quantities, and the effects of bias correction on these quantities, was assessed using a Wilcoxon signed rank test (Wilcoxon, 1945). For assessments of bias, the null hypothesis was that the bias was equivalent to zero. For assessments of the difference in bias or accuracy with respect to the baseline result, the null hypothesis was that this difference was zero.

Distributional bias ~~,~~ and improvement of that bias ~~, was~~ were considered in both the high and low tails of the streamflow distribution. Two methods were used to capture the bias in each tail. One method, referred to herein as an assessment of the observation-dependent tails, considers the observed nonexceedance probabilities to identify the days on which the highest and lowest 5% of streamflow occurred. For each respective tail, the errors were assessed based on the observations and simulations of those fixed days. The other method, referred to herein as an assessment of the observation-independent tails, ~~ignores the nonexceedance probabilities of the observations and~~ compares the ranked top and bottom 5% of observations with the independently ranked top and bottom 5% of simulated streamflow. Errors in the observation-dependent tails are an amalgamation of errors in the sequence of nonexceedance probabilities (~~timing~~the temporal structure) and in the magnitude of streamflow, whereas errors in the observation-independent tails reflect only bias in the ranked magnitudes of streamflow. In the same fashion, evaluation of the complete hydrograph can be assessed sequentially, retaining the contemporary sequencing of observations and simulations, or distributionally, considering the observations and simulations ranked independently. ~~Of course, though~~ Though the overall accuracy will vary between the sequential and distributional case, overall bias will be identical in both cases.

With an analysis of both observation-dependent and observation-independent tails, it is possible to begin to tease out the effect of ~~timing~~ temporal structure on distributional bias. The bias in observation-independent tails is not directly tied to the

~~timing~~temporal structure, or relative ranking, of simulated streamflow. That is, if the independently estimated FDC is accurate, then even if relative sequencing of streamflow is badly flawed, the bias-correction of observation-independent tails will be successful. However, even if the distribution is accurately reproduced after bias correction, the day-to-day performance may still be poor. For observation-dependent tails, the ~~timing~~ temporal structure plays a vital role on the effect of bias correction. If the ~~timing~~ temporal structure is inaccurate in the underlying hydrologic simulation, then the bias correction of observation-dependent tails will be less successful.

The bias correction approach was first tested with the observed FDCs. These observed FDCs would be unknowable in the truly ungauged case, but this test allows for an assessment of the potential utility of this approach. This examination is followed by an application with the regionally regressed FDCs described above, demonstrating one realization of this generalizeable method. This general approach to bias correction could be used with other methods for estimating the FDC and could also be used with an observed FDC for record extension, though neither of these possibilities are explored here.

3 Results

Figures ~~?? and ??~~ 3 and 4 show the overall bias and accuracy of the reproduced hydrographs; these figures are quantified in Tables ~~?? and ??. Figure ?? and Table ??~~ 1 and 2. Figure 5 and Table 1 summarize the tail bias in all approaches to streamflow simulation considered here. Similarly, Figure ~~?? and Table ??~~ 6 and Table 2 summarize the tail accuracy of all approaches. These results are discussed in detail below, beginning with a discussion of the bias and accuracy in the original kriged simulations. This is followed by a consideration of the effectiveness of bias-correction with observed FDCs as emblematic of the theoretical potential of this approach. The realization of this theoretical potential through the regionally regressed FDCs is subsequently presented. Complete results can be explored and reproduced using the associated model and data archive ~~:~~ ~~Farmer et al. (2018)~~(Farmer et al., 2018).

3.1 Simulated Hydrographs without Correction

There is statistically significant overall bias at the median (-7.1%; $10^{-0.0318} - 1$) in the streamflow distribution simulated by the kriging approach applied here (Figure ~~??~~3, boxplot A), but more significant bias is apparent in the upper and lower tails of the distribution (Figure ~~??~~5, boxplots A, D, G and J). Both the observation-dependent and observation-independent upper tails of the streamflow distribution demonstrate significant downward bias (Figure ~~??~~5, boxplots D and J). At the median, the observation-dependent upper tail is underestimated by approximately 38% (Table ~~??~~1, row 1; Figure ~~??~~5, boxplot D), while the observation-independent upper tail is underestimated by approximately 23% (Table ~~??~~1, row 2; Figure ~~??~~5, boxplot J). For the lower tail, the observation-dependent tail shows a median overestimation of 36% (Table ~~??~~1, row 1; Figure ~~??~~5, boxplot A), while the observation-independent tail is underestimated by less than one percent (table ~~??~~1, row 2; Figure ~~??~~5, boxplot G). The bias is much more variable, producing greater magnitudes of bias more often, in the lower tails than in the upper tails. Generally, biases in the observation-independent tails are less severe, both in the median and in range, than those in the

observation-dependent tails. To provide some information on regional performance and incidence of bias, Figure 7 shows the spatial distribution of bias in each tail.

In both observation-dependent and –independent cases, downward bias in the upper tail is more probable than upward biases in the lower tail. For the observation-dependent tails, approximately 89% of streamgages show downward bias for the upper tail (Figure ??5, boxplot D), and approximately 61% of the streamgages upward bias in the lower tail (Figure ??5, boxplot A). For the observation-independent tails, approximately 80% of streamgages show downward bias in the upper tail (Figure ??5, boxplot J) and approximately 50% of the streamgages exhibit upward bias in the lower tail (Figure ??5, boxplot G), indicating, as does the small median bias value, that the lower tail biases are relatively well balanced around zero for the observation-independent case for these simulations.

~~These~~ With respect to their central tendencies, these results show upward bias in lower tails and downward bias in upper tails ~~.~~ of the distribution of streamflows from the original simulations for both observation-dependent and observation-independent cases. There is, of course, a great degree of variability around this central tendency. With these baseline results, the bias-correction method presented here seeks to mitigate these biases.

3.2 Bias Correction with Observed FDCs

The results ~~provide evidence to support the hypothesis~~ for this idealized case that could not be applied in practice provide clear evidence that distributional bias in simulated streamflow can be reduced by rescaling using independently estimated FDCs. This evidence is apparent in the reduction of the magnitude and variability of overall bias (Figure ??3, boxplot C; Table ??1, rows 5 and 6) and of the bias in the observation-independent tails of the streamflow distribution (Figure ??5, boxplots I and L) when observed FDCs are used for rescaling. Similarly, the overall distributional accuracy is much improved (Figure ??4, boxplot F; Table ??2, rows 5 and 6), as is the accuracy of observation-independent tails (Figure ??6, boxplot I and L). The effect on observation-dependent tails (Figure ??5, boxplots C and F) and overall sequential accuracy (Figure ??4, boxplot C) is less compelling but still substantial.

~~While~~ Whereas the measures of bias and accuracy are summarized in Tables ~~?? and ??, Tables ?? and ??~~ 1 and 2, Tables 3 and 4 summarize the change in absolute bias and in accuracy, respectively. With the use of observed FDCs, the overall bias is reduced to a tenth of a percent at the median (Table ??1, rows 5 and 6). This represents a significant median reduction of 0.14 common-logarithm units in the overall absolute bias (Table ??3, rows 3 and 4). Overall, the distributional accuracy is improved by a median of 0.21 common-logarithm units (Table ??4, row 4). Of all streamgages considered, 99% saw a reduction in the overall absolute bias, and all saw improvements in overall distributional accuracy. These improvements extend to both observation-independent tails of the distributions. The lower observation-independent tails have a median 0.35 common-logarithm unit reduction in absolute bias (Table ??3, row 4). For the upper tail, the median reduction in absolute bias is 0.14 common-logarithm units (Table ??3, row 4). Nearly all streamgages (99%) saw reduction in absolute bias of the observation-independent tails. Table ?? 4 (row 4) shows similar improvements in tail accuracy~~;~~: -0.37 and -0.15 units in the lower and upper tails, respectively, with nearly all streamgages (excepting the lower tail of a single streamgage~~;~~, likely the result of the interpolation procedure) showing improved tail accuracy.

With the use of a perfect, observed FDC for bias correction, one would expect that nearly all bias would disappear, but the results do not show this. The temporal of the simulated hydrograph continues to play a role in the bias of observation-dependent tails. The observation-independent tails continues to exhibit a small degree of residual bias, though it is still slightly non-intuitive. This residual bias arises from the effect of representing the FDC as a set of discrete points and interpolating between them. There may be some additional effect from the small value added to avoid zero-valued streamflows or the censoring procedure, but initial exploration found little impact.

The overall sequential performance (Figure ??4, boxplot C) and the performance of observation-dependent tails (Figures ~~?? and ??~~5 and 6, boxplots C and F) demonstrate the degree to which errors in ~~timing~~ the temporal structure result in bias in the observation-dependent case even when observed FDCs are used for bias correction. Both the observation-dependent lower and upper tails exhibit bias~~;~~: 30% and -20%, respectively, at the median (Table ??1, row 5). Absolute bias in both tails show median reductions; sequential accuracy and observation-dependent tail accuracy ~~is~~ are also improved at the median (Tables ~~?? and ??~~3 and 4, row 3). Proportionally, 82% of the observation-dependent lower tails and 86% of the observation-dependent upper tails showed reduction in absolute bias; 85% of observation-dependent lower tails and 79% of observation-dependent upper tails showed improvements in accuracy. Despite improvements in overall bias and accuracy from rescaling with observed FDCs, the residual bias in the observation-dependent lower tail (Figure ??5, boxplot C) is almost always positive (upward bias) and upper tails (Figure ??5, boxplot F) are almost negative (downward bias), a result which arises primarily from errors in ~~timing~~the temporal structure.

To understand the effect of errors in ~~timing~~ the temporal structure further, consider Figure ??8, which shows the mean error in the nonexceedance probabilities~~of the~~, i.e., the difference in the ranks of the observed and simulated streamflows, of the observation-dependent upper and lower tails. The nonexceedance percentages in the lower tail are overestimated by a median of 3.8 points with 5th and 95th percentiles of 0.9 and 20.5, while the percentages in the upper tail are underestimated by 2.4 points, with 5th and 95th percentiles of -0.5 and -12.6 points. The distributions of errors in the nonexceedance probabilities closely reflect the distribution of bias in the observation-dependent tails (Figure ??5, boxplots C and F). These results show that the inaccuracy in the nonexceedance probabilities (i.e., ~~timing errors~~ errors in temporal structure) will obscure, at least partially, the improvement offered by bias correction when considering the observation-dependent errors, even when an observed FDC is used for bias correction. These ~~timing errors also almost~~ errors in temporal structure also almost always result in errors in a particular direction ~~:-~~ low for high flow and high for low flows.

3.3 Bias Correction with Regionally Regressed FDCs

When the uncertainty of regionally regressed FDCs is introduced into the bias correction procedure, the potential value of the bias correction procedure is not as convincing. There is a slight, but significant, increase in the overall bias (Table ??3, rows 1 and 2). Whereas the original estimated streamflow displays a median bias of approximately 7.1%, the median overall bias is approximately 7.6% after bias correction with estimated FDCs, (Table ??1, rows 3 and 4). Though statistically significant, the distribution of bias does not appear to have changed in a meaningful way (Figure ??3, boxplots A and B). The overall accuracy,

sequential and distributional, is also degraded (Figure ??4, boxplots B and E; Table ??4, rows 3 and 4), with more than 60% of streamgages showing degradation in sequential and distributional accuracy.

The observation-independent tails, which are not affected by errors in ~~relative timing~~temporal structure, show a divergence in performance between the results obtained using observed FDCs and those obtained using regionally regressed FDCs. With observed FDCs, both tails demonstrated substantial reductions in absolute bias and improvements in accuracy. With regionally regressed FDCs, the upper observation-independent tails continue to show reductions in absolute bias (Table ??3, row 2; Figure ??5, boxplots J and K) and improvements in accuracy (Table ??4, row 2; Figure ??6, boxplots J and K), while the lower observation-independent tails show a significant increase in absolute bias (Table ??3, row 2; Figure ??5, boxplots G and H) and a degradation of accuracy (Table ??4, row 2; Figure ??6, boxplots G and H). ~~Only 44% of observation-dependent lower tails produced after~~ After bias correction with regionally regressed FDCs, only 44% of observation-dependent lower tails showed reductions in absolute bias; 58% of upper tails showed reductions in absolute bias.

The effects of the rescaling with FDCs estimated with regional regression on overall and observation-independent tail bias and accuracy can be better understood if the properties of the estimated FDCs are considered. Figure ?? 9 shows the bias ~~and accuracy of the upper and lower~~ (left panel) and accuracy (right panel) of the lower and upper tails of the regionally regressed FDCs. ~~(~~Recall that the estimated FDCs are composed of 27 quantiles, of which the upper and lower tails contain only the eight values with nonexceedance probabilities 95% and larger and 5% and smaller, respectively. ~~)~~ The upper tails are reproduced through regional regression with an insignificant 2.5% median downward bias, but the lower tails exhibit a significant negative median bias of 38.35% (Table ??1, row 7). Because of this bias in the lower tail of the regionally regressed FDCs, the regionally regressed FDCs are unable to correct the bias in the simulated hydrograph, instead turning a small median bias into large one. As there is no ~~timing~~ temporal uncertainty in the observation-independent tails, the resulting bias arises from the bias of the regionally regressed FDC. Illustrating this fact~~:~~, the -38% bias in the lower tail of the regionally regressed FDC approximates the -33% in the observation-independent lower tail, while the -2.5% bias in the upper tail of the regionally regressed FDC approximates the -3.7% bias in the observation-independent upper tail. The introduction of this additional bias, beyond failing to correct any underlying bias in the simulated hydrograph, also markedly increased the variability of both bias and accuracy.

The results are similar for the observation-dependent tails produced after bias correction with regionally regressed FDCs, even when complicated by the addition of ~~timing~~ temporal uncertainty as discussed in section 3.2 with reference to Figure ??8. In some cases, the errors in ~~timing~~ the temporal structure (nonexceedance probability) counteract the additional bias from regionally regressed FDCs. For example, the observation-dependent lower tails have a median bias of 13%, which possesses a smaller magnitude and different sign than the median -33% bias seen in the observation-independent lower tail (Table ??1, rows 3 and 4). The addition of ~~timing~~ temporal uncertainty actually reduced the increase in absolute bias (Table ??3, rows 1 and 2) and reduced the degradation of accuracy in the lower tail (Table ??4, rows 1 and 2). These slight improvements result from an offsetting of the underestimated regionally regressed FDCs by the overestimated nonexceedance probabilities. While interesting, it seems unlikely that this result can be generalized in a simple way: that is, the errors in estimated FDCs cannot be expected to balance out the errors in nonexceedance probabilities without deleterious effects on other properties. To this point~~:~~

, as noted, rescaling by these regionally regressed FDCs with underestimated lower tails result in similarly underestimated observation-independent lower tails.

The introduction of uncertainty from regionally regressed FDCs diminishes the advantages gained by biased correction with observed FDCs. Considering the observation-independent lower tails, 55% of streamgages show reductions in absolute bias with observed FDCs that were reversed into increases of absolute bias by the introduction of regionally regressed FDCs. Another 43% of streamgages show smaller reductions in absolute bias when observed FDCs were replaced with regionally regressed FDCs. For the observation-dependent lower tails, 37% of streamgages have reversals and 31% show smaller reductions in absolute bias. For the observation-independent upper tails, 41% show reversals and 56% yield smaller reductions in absolute bias. For the observation-dependent upper tails, 24% produce reversals and 40% provide smaller reductions in absolute bias. Results are similar with respect to accuracy: while many streamgages saw reversals, a large proportion of streamgages continue to demonstrate improvements.

4 Discussion

~~The approach~~Though the first analysis presented, which utilized observed FDCs for bias correction, represents only an assessment of hypothetical potential of this general approach, the approach to bias correction presented here produced near universal and substantial reduction in bias and improvements in accuracy, overall and in each tail, for both observation-dependent and –independent cases when the uncertainty in independently estimated FDCs was minimized. For the observation-independent case, the errors are removed almost completely, and the remaining errors in the observation-dependent case mimic the ~~timing~~ temporal structure (nonexceedance probability) errors. These results~~based on observed FDCs demonstrate~~ , which are not applicable under the conditions of the true ungauged problem, demonstrate that the bias-correction approach introduced here is theoretically valid. However, this improvement becomes inconsistent with respect to bias and generally reduces the accuracy when the ~~bias and uncertainty of regional regression of the FDCs is introduced~~biased and uncertain regionally regressed FDCs are used. Furthermore, in both the observation-dependent and observation-independent tails in the case of rescaling by ~~regional regression~~regionally regressed FDCs, the improvements in the lower tails are much more variable than the improvements in the upper tail (Figures ~~?? and ??; Tables ?? and ??~~5 and 6; Tables 3 and 4). This result is not surprising, given the more-variable nature of lower-tail bias and accuracy (Figures ~~?? and ??~~5 and 6).

The regional regressions developed here were much better at estimating the upper tail of the streamflow distribution than estimating the lower tail. This provides a convenient comparison: the bias correction of lower tails with regionally regressed FDCs only improved the bias in the observation-dependent case when the low bias of the regionally regressed FDC offset the high bias of the observation-dependent tails, and did not improve accuracy in either case. However, the bias correction of upper tails with regionally regressed FDCs, which produced the upper tails with much less bias, continued to show, like in the case of observed FDCs, improvements in bias and accuracy, though to a much smaller degree than the improvements produced by observed FDCs.

Particularly in the lower tail of the distribution, the effectiveness of this bias-correction method is strongly influenced by the accuracy of the independently estimated FDC. The change in the absolute bias of the observation-independent lower tail has a 0.72 Pearson correlation with the absolute bias of the lowest eight percentiles of the FDC estimated with regional regression, showing that the residual bias in the FDC of the bias-corrected streamflow simulations is strongly correlated with the bias in the independently estimated FDC. The analogous correlation for the upper tail is 0.31. For the observation-dependent these correlations are only 0.33 for each tail, the reduced correlation for the lower tail being a result of the combination of the uncertainty in ~~timing and~~ the temporal structure and in the regionally regressed FDC. Therefore, as regional regression is not the only tool for estimating FDCs (for other examples, see Castellatin et al., 2013; Pugliese et al., 2014, 2016), improved methods for FDC estimation would further increase the impact of this bias-correction procedure. There are also hints that the representation of the FDC as a set of discrete points degraded performance. Further work might address the question of improving FDC simulation. Still further, seasonal FDCs or some other methods of increasing the temporal variability of FDCs could improve performance of this general bias correction approach.

While this method of bias correction, as implemented here using regionally regressed FDCs, improves the bias in the upper tails, it had a negative impact on lower tails. This makes the question of application or recommendation more poignant. Under what conditions might this approach be worthwhile? Initial exploration did not find a strong regional component to performance of the bias correction method. Figure 7 shows the original tail bias from pooled ordinary kriging; at each point the accuracy of the bias correction method is dependent on the original bias present as well as the error in the independently estimated FDC. For some regions, like New England, where FDCs are well estimated by regional regression, there is a general improvement in accuracy under bias correction with regionally regressed FDCs, but the improvement is highly variable. Instead, the strongest link is with the reproduction of the FDC. When magnitude of tail biases of the regionally regressed FDC was under 20%, more than 50% of streamgages showed improvements in bias, both overall and in the tails of the distribution. ~~It~~ At a particular ungauged site, it may not always be possible to determine the accuracy with which a given FDC estimation technique might perform beyond a regional cross-validated assessment of general uncertainty, making it difficult to determine if these results can be generalized. If accuracy of the estimated FDCs can be estimated, it may also be useful to consider rescaling one tail and not the other, depending on the estimated accuracy. Further work might explore the effects of hydroclimates on the ability to reproduce reliable FDCs with which to implement this bias-correction procedure.

The results of this work were also discussed in reference to earlier work that suggested a prevalence, though not a universality, of underestimation of high streamflows and over-estimation of low streamflows. Similarly, the bias correction approach produced a wide variability of results; where the high tails might have been improved, the lower tails might have been degraded. Figure 10 shows the correspondence of tails across all sites. While there is a general move towards unbiasedness in many cases (along the vertical axis), there is a great degree of variability. All methods will produce variability, it remains to future research to determine if a more consistent representation of the FDC might reduce the variability of this performance.

When looked at from the point-of-view of the estimated FDCs that need ~~timing~~ temporal information in order to simulate streamflow, this approach to bias correction is as akin to an extension of the non-linear spatial interpolation using FDCs developed by Fennessey (1994) and Hughes and Smakhtin (1996) as it is bias correction. Here it is approached as a method for

bias correction, but it can also be thought of as a novel approach to simulate the nonexceedance probabilities at an ungauged location to be used with estimated distributional information (FDCs) to simulate streamflow. In the early uses of nonlinear spatial interpolation using FDCs, the simulated nonexceedance probabilities were obtained from a hydrologically appropriate neighboring or group of neighboring streamgages (Shu and Ouarda, 2012), though the approach to identifying a hydrologically appropriate neighbor has varied. Here, the entire network is used to approximate the ungauged nonexceedance probabilities, much like the indexing problem was overcome with ordinary kriging of streamflow directly (Farmer, 2016). Two major sources of uncertainty are inherent in nonlinear spatial interpolation using FDCs: uncertainty in the nonexceedance probabilities and uncertainty in the FDC. This work addresses the general approach by attacking the former and observing that performance may be further limited by the latter. The potential success of this approach to bias correction is likely not specific to simulation with ordinary kriging.

That this approach to bias correction does improve the observation-dependent tails and the overall performance when observed FDCs are used shows that the ~~timing~~ temporal structure of the underlying simulation retains useful information even if the tails of the original simulation are biased. However, some error remains in the simulated nonexceedance probabilities. A natural extension would be to ~~wonder~~ investigate if it might be more reasonable to estimate nonexceedance probabilities directly rather than extracting their implicit values from the estimated streamflow time series as was done here. Farmer and Koltun (2017) executed a kriging approach to estimate daily nonexceedance probabilities in a smaller data set in Ohio. They found that modeling probabilities directly resulted in similar tail biases of nonexceedance probability to that observed when, as in Farmer (2016), simulating streamflow directly. In earlier work, Farmer (2015) showed that kriging nonexceedance probabilities directly and then redistributing them via an estimated FDC, as compared with kriging streamflow directly, had only a marginal effect on bias in the tails. Further exploration of this question, whether to estimate nonexceedance probabilities directly or derive them from streamflow simulations, is left for future research. This current study focuses on the more general question of whether the distributional bias in a set of simulated streamflow, the provenance thereof being more or less inconsequential, could be reduced using a regionally regressed FDC.

As mentioned earlier, recent work by Pugliese et al. (2017) explore how this generalization of non-linear spatial interpolation using FDCs can be used to improve simulated hydrographs produced by a continental scale deterministic model. They consider it as an approach to inform a large-scale model with local information, thereby improving local application without further calibration. In 46 basins in Tyrol, Pugliese et al. (2017) saw universal improvement in the simulated hydrographs, though they did not explore tails biases. The results presented here provide an analysis across a wider range of basin characteristics and climates, demonstrating a link between how well the FDC can be reproduced and ultimate improvements in performance or reductions in bias.

Additional research to explore if estimating nonexceedance probabilities directly, as opposed to the conversion of simulated streamflow to nonexceedance probabilities used here, might further improve nonlinear spatial interpolation using FDCs or simulation more generally. Although the results presented here are promising, they demonstrate that the performance of two-stage modeling, where temporal structure and magnitude are largely decoupled, is limited by the less well performing stage

of modeling. In this case, alternative methods for estimating the FDC might prove worthwhile (e.g., see Castellatin et al., 2013; Pugliese et a

.

5 Summary and Conclusions

Regardless of the underlying methodology, simulations of historical streamflow often exhibit distributional bias in the tails of the distribution of streamflow, usually an overestimate of the lower tail values and an underestimate of the upper tail values. Such bias can be extremely problematic, as it is often these very tails that affect human populations and other water management objectives the most and, thus, these tails that receive the most attention from water resources planners and managers. Therefore, a bias-correction procedure was conceived to rescale simulated time series of daily streamflow to improve simulations of the highest and lowest streamflow values. Being akin to a novel implementation of nonlinear spatial interpolation using flow-duration curves, this approach could be extended to other methods of streamflow simulation.

In a leave-one-out fashion, daily streamflow were simulated in each 2-digit hydrologic unit code using the pooled ordinary kriging. Regional regressions of 27 percentiles of the flow-duration curve in each 2-digit hydrologic unit code were independently developed. Using the Weibull plotting position, the simulated streamflow were converted into nonexceedance probabilities. The nonexceedance probabilities of the simulated streamflow were used to interpolate newly simulated streamflow volumes from the regionally regressed flow-duration curves. Assuming that the sequence of relative magnitudes of streamflow retains useful information despite possible biases in the magnitudes themselves, it was hypothesized that simulated magnitudes can be corrected using an independently estimated flow-duration curve. This hypothesis was evaluated by considering the performance of simulated streamflow observations and the performance of the relative timing of simulated streamflow. This evaluation was primarily focused on examination of errors in both the high and low tails of the streamflow distribution, defined as the lowest and highest 5% of streamflow, and considering changes in both bias and accuracy.

When observed flow-duration curves were used for bias correction, representing a case with minimal uncertainty in the independently estimated flow-duration curve, bias and accuracy of both tails were substantially improved and overall accuracy was noticeably improved. The use of regionally regressed flow-duration curves, which were observed to be approximately unbiased in the upper tails but were biased low in the lower tails, corrected the upper tail bias but failed to consistently correct the lower tail bias. Furthermore, the use of the regionally regressed flow-duration curves degraded the accuracy of the lower tails but had relatively little effect on the accuracy of the upper tails. Combining the bias-correction and accuracy results, the test with regionally regressed flow-duration curves can be said to have been successful with the upper tails (for which the regionally regressed flow-duration curves were unbiased) but unsuccessful with the lower tails. The effect on accuracy of the bias correction approach using estimated flow-duration curves was correlated with the accuracy with which each tail of the flow-duration curve was estimated by regional regression.

In conclusion, this approach to bias-correction has significant potential to improve the accuracy of streamflow simulations, though the potential is limited by how well the flow-duration curve can be reproduced. While conceived as a method of bias correction, this approach is an analog to a previously applied nonlinear spatial interpolation method using flow-duration curves

to reproduce streamflow at ungauged basins. While using the nonexceedance probabilities of kriged streamflow simulations improves upon the use of single index streamgages to obtain nonexceedance probabilities, further improvements are limited by the ability to estimate the flow-duration curve more accurately.

Code and data availability. The data and scripts used to produce the results discussed herein can be found in Farmer et al. (2018).

Competing interests. No competing interests are present.

Acknowledgements. This research was supported by the U.S. Geological Survey's National Water Census. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The R Programming language was used to implement all analyses in this work. The data, including model simulations and scripts, used to produce the figures and numbers in this report are available in a publicly-available data release that will be published upon acceptance of this manuscript.

References

Alley, W. M., Evenson, E. J., Barber, N. L., Bruce, B. W., Dennehy, K. F., Freeman, M. C., Freeman, W. O., Fischer, J. M., Hughes, W. B., Kennen, J. G., Kiang, J. E., Maloney, K. O., Musgrove, M., Ralston, B., Tessler, S., and Verdin, J. P.: Progress toward establishing a national assessment of water availability and use, Circular 1384, U.S. Geological Survey, https://pubs.usgs.gov/circ/1384, 2013.

Archfield, S. A., Vogel, R. M., Steeves, P. A., Brandt, S. L., Weiskel, P. K., and Garabedian, S. P.: The Massachusetts Sustainable-Yield Estimator: A decision-support tool to assess water availability at ungaged stream locations in Massachusetts, Scientific Investigations Report 2009-5227, U.S. Geological Survey, https://pubs.usgs.gov/sir/2009/5227/, 2010.

Archfield, S. A., Steeves, P. A., Guthrie, J. D., and Ries III, K. G.: Towards a publicly available, map-based regional software tool to estimate unregulated daily streamflow at ungauged rivers, Geoscientific Model Development, 6, 101–115, https://doi.org/10.5194/gmd-6-101-2013, 2013.

Blum, A. G., Archfield, S. A., and Vogel, R. M.: On the probability distribution of daily streamflow in the United States, Hydrology and Earth System Sciences, 21, 3093–3103, https://doi.org/10.5194/hess-21-3093-2017, 2017.

Castellatin, A., Botter, G., Hughes, D., Liu, S., Ouarda, T., Parajka, J., Post, D., Sivapalan, M., Spence, C., Viglione, A., and Vogel, R.: Prediction of flow-duration curves in ungauged basins, in: Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, edited by Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H., Cambridge Univeristy Press, 2013.

Eng, K., Chen, Y.-Y., and Kiang, J. E.: User's guide to the weighted-multiple-linear-regression program (WREG version 1.0), Techniques and Methods 4-A8, U.S. Geological Survey, https://pubs.usgs.gov/tm/tm4a8/, 2009.

Falcone, J.: Geospatial Attributes of Gages for Evaluating Streamflow, digital spatial dataset, http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml, 2011.

Farmer, W. H.: Estimating records of daily streamflow at ungaged locations in the southeast United States, Ph.d. disertation, Tufts University, 2015.

Farmer, W. H.: Ordinary kriging as a tool to estimate historical daily streamflow records, Hydrology and Earth System Sciences, 20, 2721–2735, https://doi.org/10.5194/hess-20-2721-2016, 2016.

Farmer, W. H. and Koltun, G.: Geospatial tools effectively estimate nonexceedance probabilities of daily streamflow at ungauged and intermittently gauged locations in Ohio, Journal of Hydrology: Regional Studies, 13, 208 – 221, https://doi.org/https://doi.org/10.1016/j.ejrh.2017.08.006, 2017.

Farmer, W. H. and Vogel, R. M.: On the deterministic and stochastic use of hydrologic models, Water Resources Research, 52, 5619–5633, https://doi.org/10.1002/2016WR019129, 2016.

Farmer, W. H., Archfield, S. A., Over, T. M., Hay, L. E., LaFontaine, J. H., and Kiang, J. E.: A comparison of methods to predict historical daily streamflow time series in the southeastern United States, Scientific Investigations Report 2014-5231, U.S. Geological Survey, https://doi.org/10.3133/sir20145231, 2014.

Farmer, W. H., Knight, R. R., Eash, D. A., Hutchinson, K. J., Linhart, S. M., Christiansen, D. E., Archfield, S. A., Over, T. M., and Kiang, J. E.: Evaluation of statistical and rainfall-runoff models for predicting historical daily streamflow time series in the Des Moines and Iowa River watersheds, Scientific Investigations Report 2015-5089, U.S. Geological Survey, https://doi.org/10.3133/sir20155089, 2015.

Farmer, W. H., Over, T. M., and Kiang, J. E.: Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow-duration curves: Data release, Tech. rep., U.S. Geological Survey, https://doi.org/10.5066/F7VD6XNG, 2018.

Fennessey, N. M.: A hydro-climatological model of daily stream flow for the northeast United States, Ph.d. dissertation, Tufts University, 1994.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Hughes, D. A. and Smakhtin, V.: Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves, Hydrological Sciences Journal, 41, 851–871, https://doi.org/10.1080/02626669609491555, 1996.

Lichty, R. W. and Liscum, F.: A rainfall-runoff modeling procedure for improving estimates of T-year (annual) floods for small drainage basins, Water Resources Invesgations Report 78-7, U.S. Geological Survey, https://pubs.er.usgs.gov/publication/wri787, 1978.

Mohamoud, Y. M.: Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves, Hydrological Sciences Journal, 53, 706–724, https://doi.org/10.1623/hysj.53.4.706, 2008.

Over, T., Farmer, W., and Russell, A.: Refinement of a regression-based method for prediction of flow-duration curves of daily streamflow in the conterminous United States, Scientific Investigations Report 2018-5072, U.S. Geological Survey, https://doi.org/10, 2018.

Parajka, J., Andréassian, V., Archfield, S., Bàrdossy, A., Blöschl, G., Chiew, F., Duan, Q., Gelfan, A., Hlavcova, K., Merz, R., McIntyre, N., Oudin, L., Perrin, C., Rogger, M., Salinas, J., Savenije, H., Skøien, J., Wagener, T., Zehe, E., and Zhang, Y.: Prediction of runoff hydrographs in ungauged basins, in: Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, edited by Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H., Cambridge Univeristy Press, 2013.

Poncelet, C., Andréassian, V., Oudin, L., and Perrin, C.: The Quantile Solidarity approach for the parsimonious regionalization of flow duration curves, Hydrological Sciences Journal, 62, https://doi.org/10.1080/02626667.2017.1335399, 2017.

Pugliese, A., Castellarin, A., and Brath, A.: Geostatistical prediction of flow-duration curves in an index-flow framework, Hydrology and Earth System Sciences, 18, 3801–3816, https://doi.org/10.5194/hess-18-3801-2014, 2014.

Pugliese, A., Farmer, W. H., Castellarin, A., Archfield, S. A., and Vogel, R. M.: Regional flow duration curves: Geostatistical techniques versus multivariate regression, Advances in Water Resources, 96, 11 – 22, https://doi.org/https://doi.org/10.1016/j.advwatres.2016.06.008, 2016.

Pugliese, A., Persiano, S., Bagli, S., Mazzoli, P., Parajka, J., Arheimer, B., Capell, R., Montanari, A., Blöschl, G., and Castellarin, A.: A geostatistical data-assimilation technique for enhancing macro-scale rainfall-runoff simulations, Hydrology and Earth System Sciences Discussions, 2017, 1–29, https://doi.org/10.5194/hess-2017-589, 2017.

Rasmussen, T. J., Lee, C. J., and Ziegler, A. C.: Estimation of constituent concentrations, loads, and yields in streams of Johnson County, northeast Kansas, using continuous water-quality monitoring and regression models, October 2002 through December 2006, Scientific Investigations Report 2008-5014, U.S. Geological Survey, https://pubs.usgs.gov/sir/2008/5014/, 2008.

Seaber, P. R., Kapanos, F. P., and Knapp, G. L.: Hydrologic Unit Maps, Water Supply Paper 2294, U.S. Geological Survey, https://pubs.usgs.gov/wsp/wsp2294/, 1987.

Sherwood, J. M.: Estimation of peak-frequency relations, flood hydrographs, and volume-duration-frequency relations of ungaged small urban streams in Ohio, Water-Supply Paper 2432, U.S. Geological Survey, https://pubs.er.usgs.gov/publication/wsp2432, 1994.

Shu, C. and Ouarda, T. B. M. J.: Improved methods for daily streamflow estimates at ungauged sites, Water Resources Research, 48, 1–15, https://doi.org/10.1029/2011WR011501, 2012.

Sivapalan, M.: Prediction in ungauged basins: a grand challenge for theoretical hydrology, Hydrological Processes, 17, 3163–3170, https://doi.org/10.1002/hyp.5155, 2003.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., and Zehe, E.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydrological Sciences Journal, 48, 857–880, https://doi.org/10.1623/hysj.48.6.857.51421, 2003.

Skøien, J. O. and Blöschl, G.: Spatiotemporal topological kriging of runoff time series, Water Resources Research, 43, 1–21, https://doi.org/10.1029/2006WR005760, 2007.

Smakhtin, V.: Generation of natural daily flow time-series in regulated rivers using a non-linear spatial interpolation technique, Regulated Rivers Research and Management, 15, 311–323, https://doi.org/10.1002/(SICI)1099-1646(199907/08)15:4<311::AID-RRR544>3.0.CO;2-W, 1999.

Thomas, W. O.: An evaluation of flood frequency estimates based on runoff modeling, Journal of the American Water Resources Association, 18, 221–229, https://doi.org/10.1111/j.1752-1688.1982.tb03964.x, 1982.

Tobin, J.: Estimation of Relationships for Limited Dependent Variables, Econometrica, 26, 24–36, https://doi.org/10.2307/1907382, 1958.

Weibull, W.: A statistical theory of strength of materials, Ing. Vetensk. Akad. Handl., 151, 1–45, 1939.

Wilcoxon, F.: Individual Comparisons by Ranking Methods, Biometrics Bulletin, 1, 80–83, https://doi.org/10.2307/3001968, 1945.

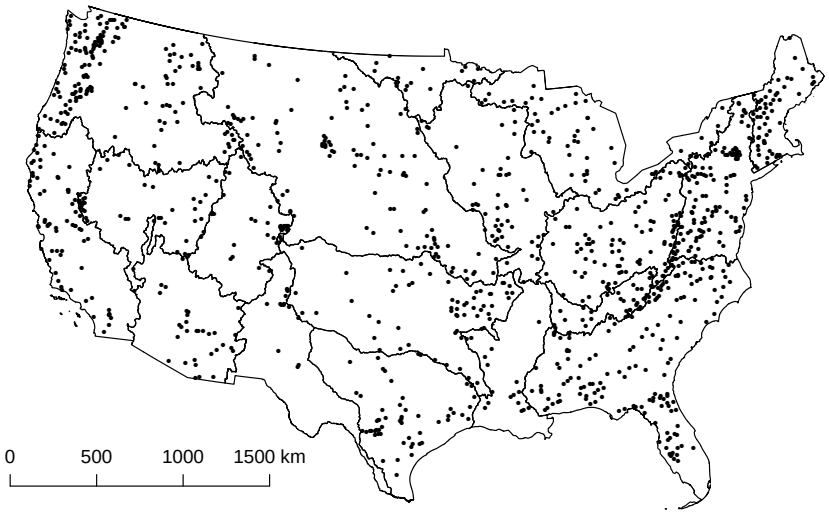

Figure 1. Map of the locations of ~~streamgages used for analysis. All~~ 1168 reference quality streamgages from the GAGES-II database (Falcone, 2011) ~~are included here~~used for analysis. ~~Only those marked with an X were retained, having~~ All streamgages used have more than 14 complete water years between 01 October 1980 and 30 September 2013. ~~With this criterion, 1168 streamgages were retained.~~ The outlines of 2-digit Hydrologic Units, which define the regions used here, are provided for further context.

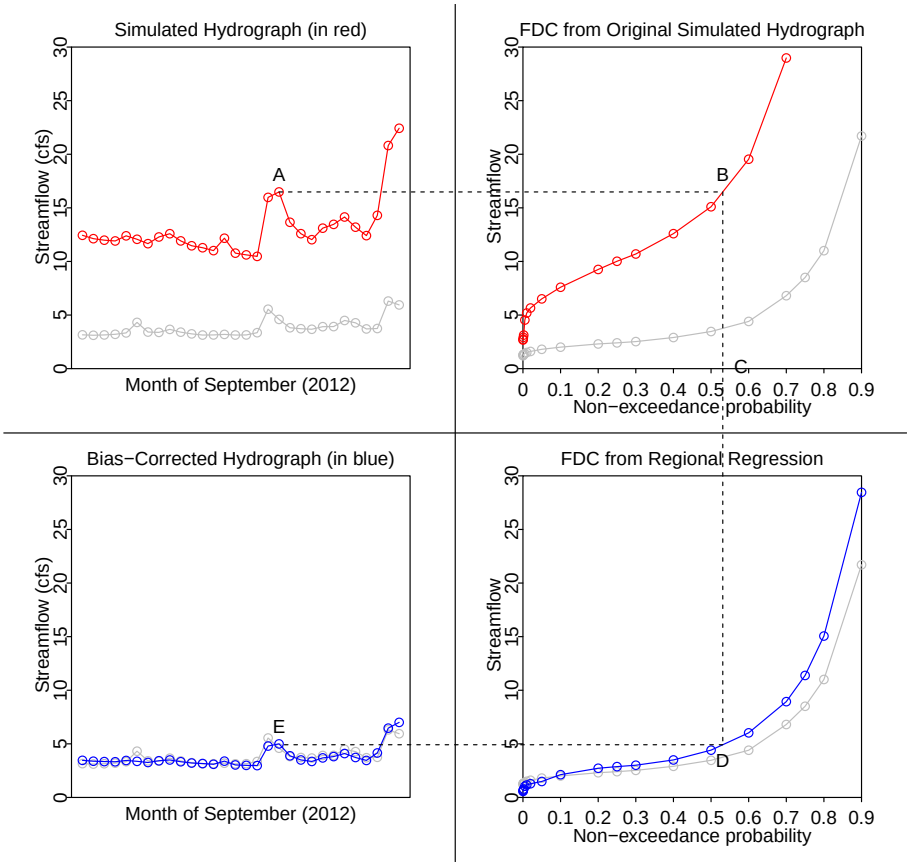

Figure 2. Diagram showing the bias-correction methodology applied here. The simulated daily hydrograph at the ungauged site is presented in the upper-left panel. For any particular point on the hydrograph (point A) the daily volume of streamflow can be mapped to a non-exceedance probability using the rank order of simulated streamflows (points B and C). With an independently estimated flow duration curve (FDC) from some procedure such as regional regression, the non-exceedance probability can be rescaled to a new volume (point D) and placed back in same sequence as the simulated streamflows (point E) to produce a bias-corrected hydrograph. This example is shown for one month, though the FDC applies across the entire period of record. As this data is based on an example site, the observed streamflows and FDC are shown in grey on each figure.

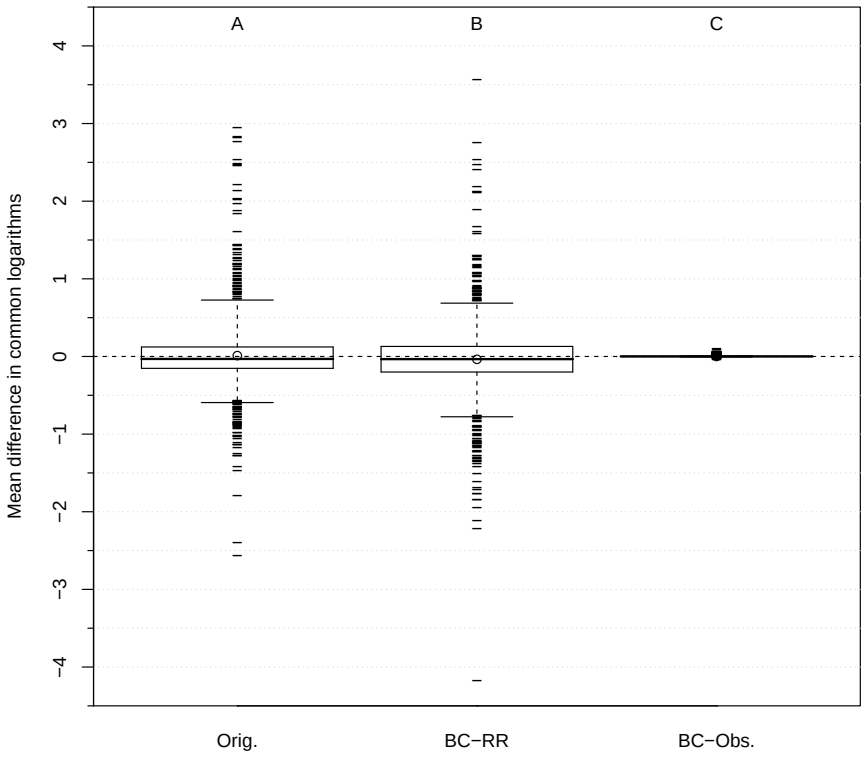

Figure 3. Distribution of logarithmic bias, measured as the mean difference between the common logarithms of simulated and observed streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow- duration curves. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

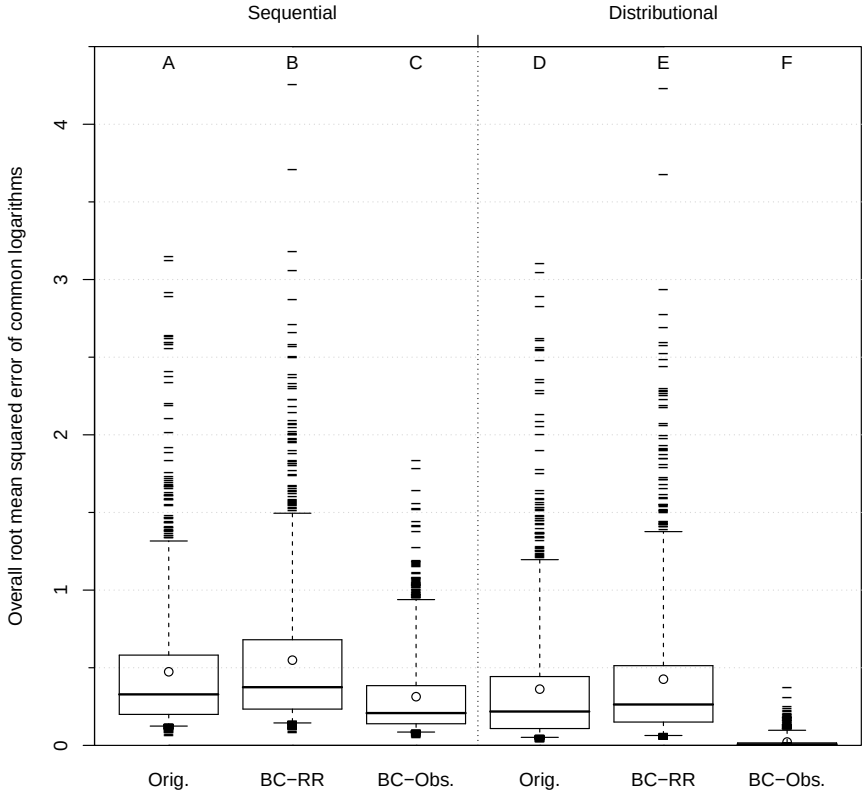

Figure 4. Distribution of logarithmic accuracy, measured as the root mean squared error between the common logarithms of observed and simulated streamflow at 1168 streamgages across the conterminous United States. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. Sequential indicates that contemporary days were compared, while distributional indicates that days of equal rank were compared. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

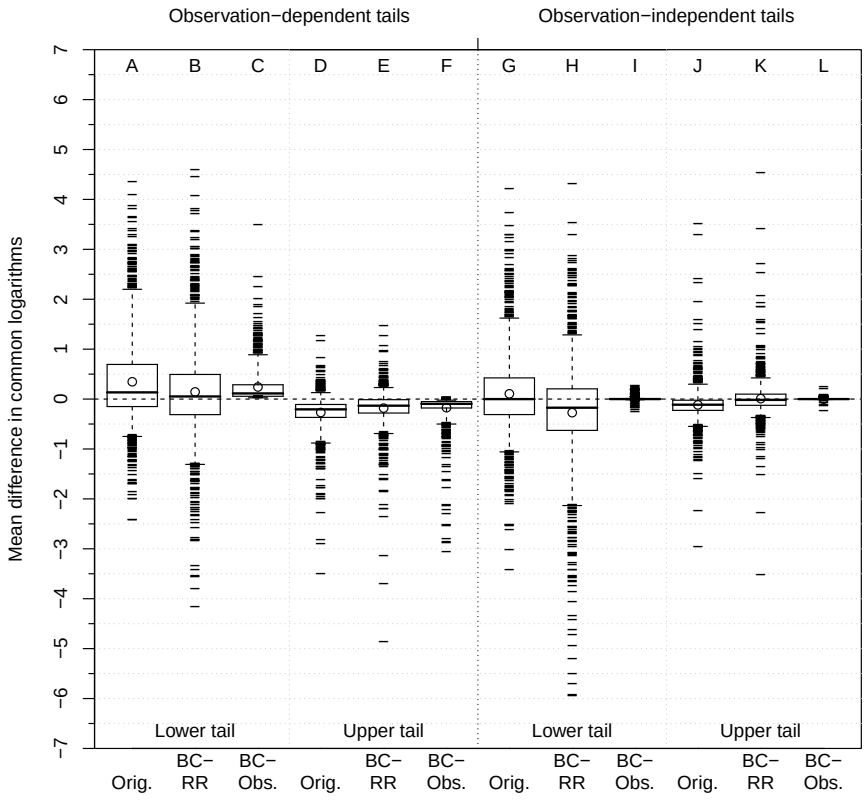

Figure 5. Distribution of logarithmic bias, measured as the mean difference between the common logarithms of simulated and observed streamflow ~~(simulated minus observed)~~ at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Observation-dependent tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent tails rank observations and simulation independently. The upper tail considers the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

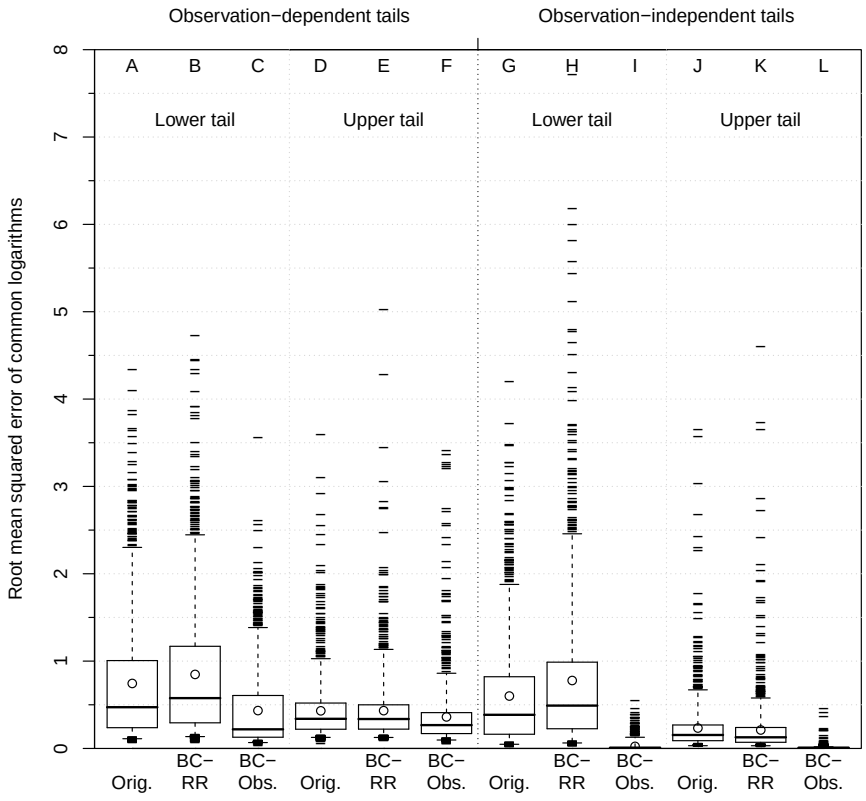

Figure 6. Distribution of logarithmic accuracy, measured as the root mean squared error between the common logarithms of simulated and observed streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Observation-dependent tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent tails rank observations and simulation independently. The upper tail considers the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

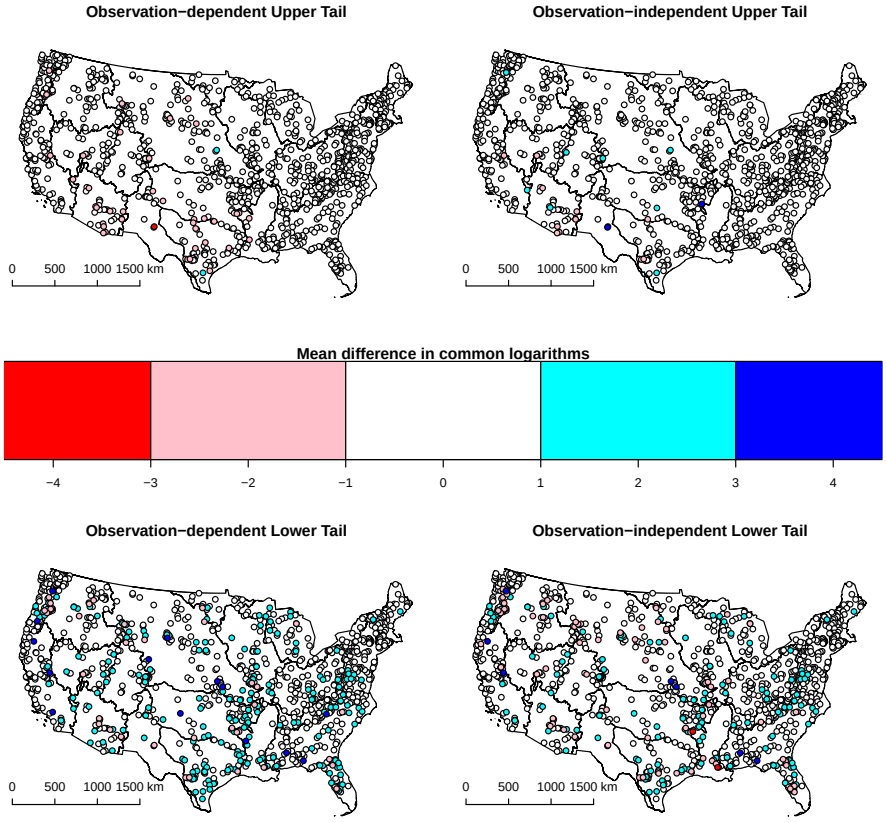

Figure 7. Maps showing the distribution of logarithmic bias, measured as the mean difference between the common logarithms of simulated and observed streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Observation-dependent tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent tails rank observations and simulation independently. The upper tail considers the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. The bias is derived from the original simulation of daily streamflow using pooled ordinary kriging at 1168 sites regionalized by the 2-digit Hydrologic Units (polygons).

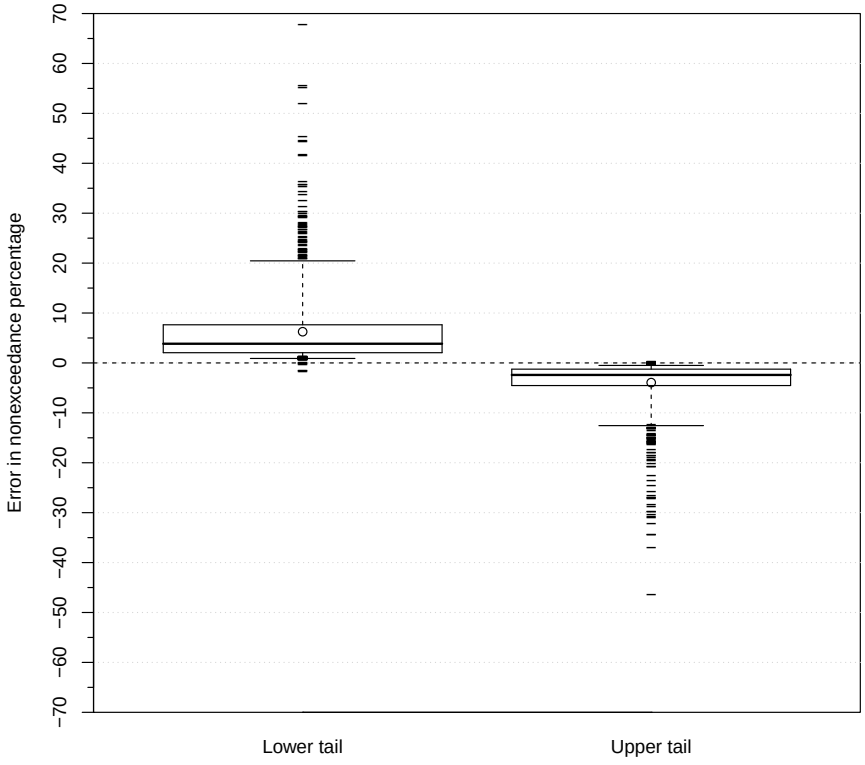

Figure 8. Distribution of mean error in the simulated nonexceedance probabilities of the lowest and highest 5% of observed daily streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States. The upper tail considers the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

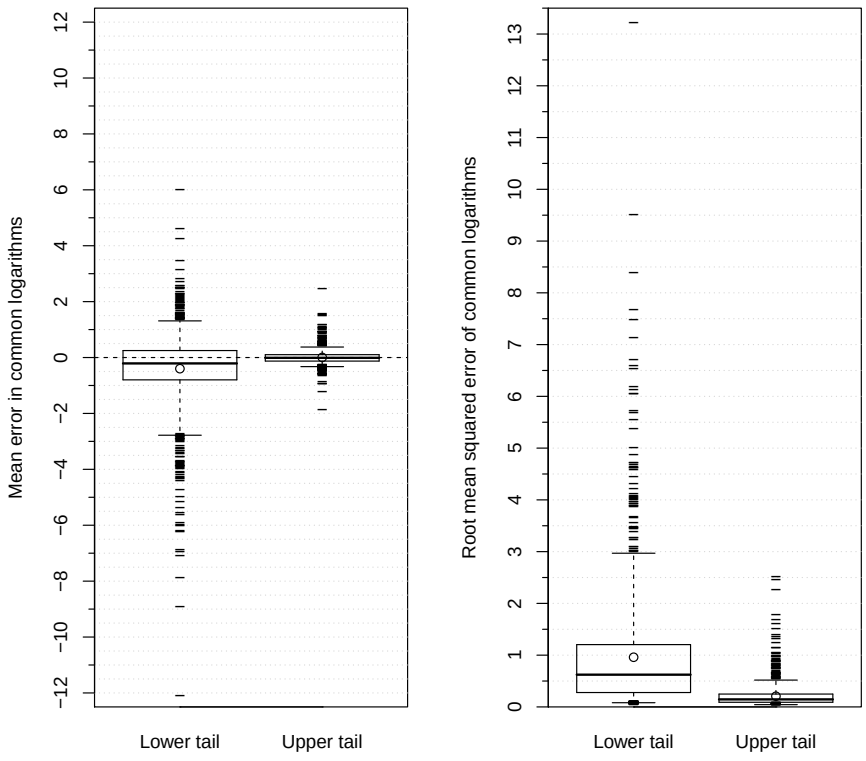

Figure 9. Distribution of logarithmic bias (left panel), measured as the mean difference between the common logarithms of quantiles of observed and simulated streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States, and logarithmic accuracy (right panel), measured as the root mean squared error between the common logarithms of quantiles of observed and simulated streamflow at the same streamgage, in the upper and lower quantiles of regionally regressed flow-duration curves. The upper tail considers the 8 quantiles in the highest 5% of streamflow, while the lower tail considers the 8 quantiles in the lowest 5% of streamflow. The tails of the boxplots extend to the 5th and 95th percentiles of the distribution; the ends of the boxes represent the 25th and 75th percentiles of the distribution; the heavier line in the box represents the median of the distribution; the open circle represents the mean of the distribution; outliers beyond the 5th and 95th percentile are shown as horizontal dashes.

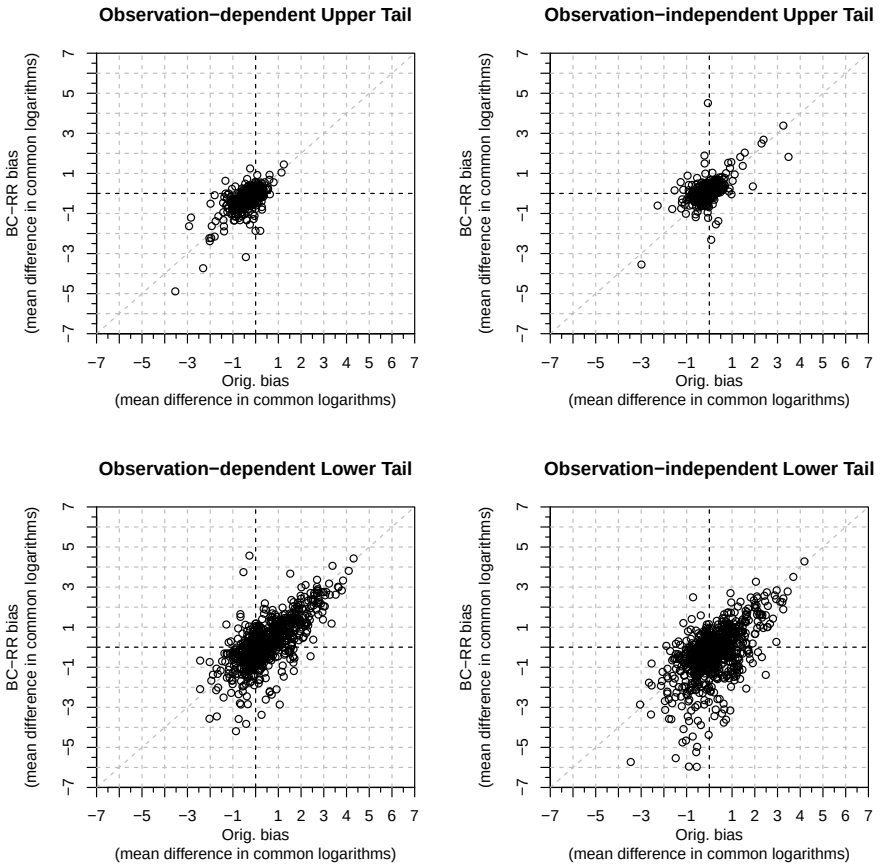

Figure 10. Scatter plots showing the correspondence of logarithmic bias, measured as the mean difference between the common logarithms of simulated and observed streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Observation-dependent tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent tails rank observations and simulation independently. The upper tail considers the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Orig. refers to the original simulation with pooled ordinary kriging, and BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves.

Table 1. ~~Distribution~~ Measures of the distribution of logarithmic bias, ~~measured~~ computed as the mean difference between the common logarithms of simulated and observed streamflow (simulated minus observed) at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed ~~flow- duration~~ flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. Observation-dependent (OD) tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent (OI) tails rank observations and simulation independently. The upper tail observes the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Significance is the p-value resulting from a Wilcoxon signed rank test with continuity correction, with the null hypothesis that the median of distribution is equal to zero ~~, while~~ and the alternative hypothesis that median is not equal to zero.

		Overall					Lower Tail					Upper Tail				
		Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance
Orig	OD	-0.0318	0.0108	0.2752	0.4574	0.0067	0.1340	0.3469	0.8437	0.8918	<0.0001	-0.2060	-0.2685	0.2590	0.3532	<0.0001
	OI	-0.0318	0.0108	0.2752	0.4574	0.0067	-0.0007	0.1058	0.7347	0.8323	0.0245	-0.1129	-0.1165	0.2036	0.3420	<0.0001
BC-RR	OD	-0.0344	-0.0364	0.3298	0.4992	<0.0001	0.0539	0.1446	0.8040	0.9664	<0.0001	-0.1326	-0.1808	0.2678	0.3827	<0.0001
	OI	-0.0344	-0.0364	0.3298	0.4992	<0.0001	-0.1732	-0.2723	0.8323	1.0893	<0.0001	-0.0162	0.0085	0.2240	0.3670	0.0547
BC-Obs.	OD	0.0004	0.0004	0.0017	0.0078	<0.0001	0.1151	0.2426	0.2281	0.3225	<0.0001	-0.0957	-0.1735	0.1284	0.2943	<0.0001
	OI	0.0004	0.0004	0.0017	0.0078	<0.0001	0.0000	0.0014	0.0035	0.0316	0.0018	0.0004	0.0009	0.0051	0.0144	<0.0001
Estimated FDC		-0.0796	-0.1270	0.4091	0.5525	<0.0001	-0.2101	-0.3988	1.0485	1.3589	<0.0001	-0.0108	0.0047	0.2302	0.2611	0.1336

Table 2. ~~Distribution~~ Measures of the distribution of logarithmic accuracy, ~~measured~~ computed as the root mean squared error between the common logarithms of observed and simulated streamflow at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails. Orig. refers to the original simulation with pooled ~~,~~ ordinary kriging, BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. Observation-dependent (OD) tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent (OI) tails rank observations and simulation independently. The upper tail observes the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow.

		Overall				Lower Tail				Upper Tail			
		Median	Mean	Interquartile Range	Standard Deviation	Median	Mean	Interquartile Range	Standard Deviation	Median	Mean	Interquartile Range	Standard Deviation
Orig	OD	0.3286	0.4741	0.3818	0.4293	0.4722	0.7448	0.7649	0.7197	0.3394	0.4310	0.2998	0.3501
	OI	0.2182	0.3623	0.3347	0.4164	0.3852	0.6003	0.6583	0.6171	0.1542	0.2338	0.1800	0.2969
BC-RR	OD	0.3747	0.5489	0.4466	0.4827	0.5763	0.8476	0.8802	0.7633	0.3371	0.4331	0.2785	0.3913
	OI	0.2634	0.4264	0.3631	0.4609	0.4905	0.7780	0.7622	0.8626	0.1277	0.2116	0.1696	0.3209
BC-Obs.	OD	0.2080	0.3137	0.2454	0.2660	0.2186	0.4350	0.4789	0.4558	0.2674	0.3612	0.2400	0.3716
	OI	0.0066	0.0218	0.0115	0.0369	0.0066	0.0240	0.0096	0.0556	0.0084	0.0114	0.0050	0.0234
Estimated FDC		0.4073	0.6220	0.5560	0.6699	0.6227	0.9600	0.9259	1.1495	0.1455	0.2068	0.1585	0.2207

Table 3. ~~Distribution~~ Measures of ~~change~~ the distribution of changes in absolute logarithmic bias with bias correction, ~~measured~~ where absolute logarithimic bias is computed as the absolute value of the mean difference between the common logarithms of bias-corrected and simulated streamflow at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails, where the simulated streamflow was obtained with pooled ~~,~~ ordinary kriging. BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow- duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow-duration curves. Observation-dependent (OD) tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent (OI) tails rank observations and simulation independently. The upper tail observes the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Significance is the p-value resulting from a paired Wilcoxon signed rank test with continuity correction, with the null hypothesis that the median difference with respect to the original simulation is equal to zero, while the alternative hypothesis that median difference is not equal to zero.

		Overall					Lower Tail					Upper Tail				
		Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance
BC-RR	OD	0.0215	0.0385	0.2117	0.3274	<0.0001	0.0163	0.0269	0.4866	0.5953	0.3710	-0.0545	-0.0526	0.1857	0.2690	<0.0001
	OI	0.0215	0.0385	0.2117	0.3274	<0.0001	0.0508	0.1588	0.5922	0.7885	<0.0001	-0.0273	-0.0261	0.1946	0.2813	<0.0001
BC-Obs.	OD	-0.1382	-0.2605	0.2280	0.3718	<0.0001	-0.1996	-0.3859	0.5354	0.5572	<0.0001	-0.1111	-0.1334	0.1551	0.3017	<0.0001
	OI	-0.1382	-0.2605	0.2280	0.3718	<0.0001	-0.3492	-0.5615	0.6255	0.6081	<0.0001	-0.1424	-0.2160	0.1891	0.2821	<0.0001

Table 4. ~~Distribution~~ Measures of ~~change~~ the distribution of changes in logarithmic accuracy between original and bias-corrected simulations, ~~measured~~ where the logarithmic accuracy is computed as the root mean squared error between the common logarithms of bias-corrected and simulated streamflow at 1168 streamgages across the conterminous United States for observation-dependent and observation-independent upper and lower tails, where the simulated streamflow was obtained using pooled ~~,~~ ordinary kriging. BC-RR refers to the Orig. hydrograph bias-corrected with regionally regressed flow-duration curves, and BC-Obs. refers to the Orig. hydrograph bias-corrected with observed flow- duration curves. Observation-dependent (OD) tails retain the ranks of observed streamflow, while matching simulations by day. Observation-independent (OI) tails rank observations and simulation independently. The upper tail observes the highest 5% of streamflow, while the lower tail considers the lowest 5% of streamflow. Significance is the p-value resulting from a paired Wilcoxon signed rank test with continuity correction, with the null hypothesis that the median difference with respect to the original simulation is equal to zero, while the alternative hypothesis that median difference is not equal to zero.

		Overall					Lower Tail					Upper Tail				
		Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance	Median	Mean	Interquartile Range	Standard Deviation	Significance
BC-RR	OD	0.0331	0.0749	0.1636	0.2966	<0.0001	0.0422	0.1028	0.3897	0.5377	<0.0001	-0.0019	0.0020	0.1111	0.2348	0.1532
	OI	0.0377	0.0641	0.2159	0.3294	<0.0001	0.0601	0.1777	0.5646	0.7791	<0.0001	-0.0222	-0.0222	0.1800	0.2751	<0.0001
BC-Obs.	OD	-0.0658	-0.1604	0.1615	0.2794	<0.0001	-0.1554	-0.3098	0.4079	0.4597	<0.0001	-0.0436	-0.0698	0.1051	0.2621	<0.0001
	OI	-0.2056	-0.3405	0.3138	0.3957	<0.0001	-0.3702	-0.5763	0.6399	0.6027	<0.0001	-0.1450	-0.2224	0.1805	0.2852	<0.0001