Response to Review of HESSD-11-5599-2014 by Hoshin Gupta

Note that original reviewer comments are in blue and author responses are in black throughout.

Overall Review Comments: I congratulate and commend the authors for an excellent job in compiling and reporting this "large-sample" data set of watersheds across the CONUS that can be used for hydrological and modeling investigations emphasizing spatial extensiveness (breadth) and generality of hydrological understanding (and therefore associated model performance). The paper is an excellent example of a relatively comprehensive study and report, and should function as a "model" or "template" for similar studies to be conducted for the other continents on our planet. In addition to reporting a carefully considered and reasoned approach to basin selection, they authors provide a benchmark assessment of simulation performance based on a "standard" lumped catchment model and calibration approach. I have only a very few suggestions for how the paper might be improved.

The authors thank Hoshin for his positive review and thoughtful consideration for areas of improvement. Regarding the specific comments, we have attempted to address them all in a thorough manner.

1) It might be nice to see (in Section 2) some more analysis of "basin characteristics" that would facilitate comparison/contrasting of the CONUS basins with ones from other continents. This analysis should probably include a) physical descriptors such as size, mean elevation distribution, shape (length to width ratio), and river characteristics such as distribution of stream order, dendritic pattern etc., and b) climatological descriptors including both annual values and monthly climatology.
We have added two new figures (Figs. 2 & 3) to address this comment. Figure 2 provides CDFs of basin physical descriptors while Figure 3 includes CDFs of annual climatogical variables. There are two new paragraphs discussing these figures in Section 2.2. The input data will be provided with the basin set download as well.

2) While NSE = 0.55 is "OK" as a benchmark, it does reflect a relatively low level of model performance. Perhaps the authors could also slightly expand the discussion in the text to mention also the fraction of catchments exceeding (say) NSE = 0.8. This would help to set the tone for future studies by setting and "OK" level and a "Good level".
This is a good idea and we have included this discussion in the text. We have included 0.55 and 0.8 NSE levels as "OK" and "Good."

3) I think it would be good to more strongly emphasize the role of large-sample studies to help identify "outlier" catchments (and regions), along with the important function of "characterizing" the nature of the "outlier" (e.g., as being likely due to data errors, model inadequacy, calibration failure, etc).

We have added discussed outlier basins and their underlying causes throughout the text in a more direct manner (e.g. Fig. 4).

4) The issue of performance on basins with strong annual climatology does not come through very strongly in the discussion. I wonder if it would help to include a map showing where the "climatology" is strong and where it is relatively weak (e.g. strength of flow correlation (?))".
We have added a flow autocorrelation plot (Fig. 8d) and discussed how those spatial patterns relate to NSE - MNSE performance differences.

5) The authors report MSE decomposition components for bias and variability. For completeness, perhaps they could also report the obs-sim cross-correlation coefficient.
We've added the correlation coefficient plot (Fig. 5b) and corresponding discussion to the text.

Minor Comments:

1) The sentence "Gupta et al. (2014) emphasize …" beginning on page 5602 line 4 cites the paper twice (at beginning and end).
Fixed sentence

2) The phrase "well know" on page 5605 line 20 should be "well known"
Fixed statement

3) On page 5606 line 3, the phrase "necessitating a snow model is required," either the word "necessitating" or the phrase "is required" should be removed.
Fixed statement

4) On page 5607 line 24 "is shown by" should be "as shown by".
Fixed statement

5) On page 5609 line 23, the term "poorlyfollowing" should be "poorly following".
Fixed statement

6) In Section 4.3 it might be interesting to compare spatial variability results with those reported by Martinez & Gupta (2010, 2011 WRR).
We have added some discussion of these papers. It is interesting to see similar spatial patterns in model performance across CONUS, even with a monthly water balance model.

7) On page 5615 line 8 should this be "which utilized 425 of the basins …"?
Removed statement

Additional Suggestion: As the basis for a complementary study, I think it would be interesting to repeat the calibration-evaluation study while interactively removing the (say) 5% or 10% of the time-steps (used to compute the performance measure) that correspond to the largest simulation errors (and therefore strongly influence the selected “best” parameter set). While this might lead to interesting response surface artifacts during calibration (but none that SCE should not be able to handle), I wonder if this would lead to more stable calibration results, when viewed for the evaluation period (functioning as sort of a fault-detection strategy)?

We agree, this would be an interesting study. We also agree that it may lead to more stable calibration results and may improve the validation period model performance. We do discuss the impact of large errors in the later figures in this text, but it would be very worthwhile to explore this in a rigorous manner. These are the types of studies we are envisioning coming from this dataset and we are pleased to see ideas for future work come quickly.

Response to Review of HESSD-11-5599-2014 by Anonymous Reviewer #2

Note that original reviewer comments are in blue and author responses are in black throughout.

The article presents a new large datasets of catchments in the US built for hydrological modelling applications. The authors shortly present the dataset and then the application of the SAC-SMA model considered as a benchmark. The issues of model spatial variability of model performance and the weight of major model errors in overall performance criteria are discussed.
This is a very valuable contribution, which should encourage the application of models on large datasets for various purposes (validation, regionalization, etc.). The article is generally clear and easy to follow. I have however four main suggestions to improve its content:

We thank the reviewer for their very thorough and thoughtful review. We have added figures and text in many places to address the reviewer's concerns.

A. The introduction should better review and acknowledge the past efforts to gather large datasets for hydrological applications in the US (e.g. the MOPEX dataset among others; see also the review of Gupta et al. (2014) in their supplementary material) and better explain to which extent this new dataset offers new opportunities for model testing compared to existing US datasets.
B. Although the main objective of the paper is to present this new dataset and the benchmark model application, the introduction could raise the scientific questions that the authors wish to specifically investigate in this article, e.g. related to the issues discussion in section 4.

We have reworked the introduction to include more cites from Gupta et al. (2014) and several mentions of MOPEX. We have also included discussion of how this dataset compares to and extends the MOPEX dataset. We have also included in the discussion reference to the scientific questions this paper is aiming to address.

C. The presentation of the data set could be improved, by introducing a more detailed description of the catchment physical characteristics.
We agree with this comment, it was an oversight to not include a more detailed description of the basin set in the paper. We have added two new figures along with panels in other figures to address this comment. We will also be including the basin descriptor data in the downloadable dataset.

D. I think the choice of the authors to use the classical Nash-Sutcliffe efficiency index as objective function for calibrating the benchmark model is questionable, given the clear deficiencies of this criterion, as demonstrated by the work of Gupta et al. (2009). I think this makes the proposed benchmark a bit outdated. It is now five years that Gupta et al. (2009) proposed their KGE criterion, and this paper is a good way to encourage the future users of the dataset to use more up-to-date

approaches for model calibration and give up old habits that are clearly less efficient. Therefore I encourage the authors to present their results using the KGE criterion as objective function instead of RMSE-based criteria.
We thank the reviewer for their thoughtful consideration of the many issues regarding the choice of objective function. While we agree that KGE is likely a “better” objective function than RMSE and that RMSE is outdated, however KGE is essentially a reweighting of RMSE and still subject many of the same issues as RMSE, although to a lesser degree. We do provide the decomposition of RMSE into the three KGE components in Fig. 7.

We feel that having discussion on the limits of using RMSE as the objective function and presenting the results in terms of NSE while including the decomposition terms and discussion are worthwhile to the community. It highlights how RMSE performs for the various decomposition flow metrics and again highlights the need to use more innovative and thoughtful objective functions.

This study is intended to provide a benchmark dataset which provides the “old habit” approaches as the benchmark to advance forward and apply, say, KGE as the objective function, then compare to this benchmark for any type of streamflow based performance metric. As Reviewer #1 notes in their additional comment, this dataset allows for advanced calibration methodology experiments, one of which would be to use a more advanced objective function.

Minor Comments:

1. Section 2: The authors should give illustrations of the physical and hydroclimatic characteristics of the selected catchments. Distributions of catchment size, mean elevation, slope, or other descriptors, as well as basic hydroclimatic values (mean Q, P and PE), could give a better idea of the types of selected catchments.
We have added two figures (Figs. 2 & 3) along with two new paragraphs of discussion in the text. We will also include these data in the dataset.

2. Section 2: All readers may not be equally familiar with the geography of the US. Therefore, to better follow the discussions presented in section 4 on the spatial distribution of results, which refers to several specific regions or locations in the US, it might be useful to have a map (e.g. the map in Fig 1.a) that show these regions.
We have changed figure 1 to include references to specific geographic areas and changed the text to help clarify regional descriptions.

3. Section 2.1: A few lines could be added on data quality and availability. Are there indexes to qualify the reliability of streamflow data? What is the range of percentages of gaps in the series?
We have added some discussion of streamflow data quality control flags to the text. We are also including the available flags with the downloadable dataset

4. P. 5603, L. 23: write “contiguous United States (CONUS)”

We have spelled out this abbreviation

5. P. 5603, L. 12: What “MT-CLIM” stands for?

MT-CLIM stands for Mountain Climate simulator (MT-CLIM).

6. Section 3.1: The authors could shortly comment the existing past applications of this model on large datasets, especially in the US. What were the results? What is already known on the possible model limits across the US?

We have tried to find a few applications of this and similar models across CONUS. We have found little to go on. If the reviewer has a specific reference in mind, we would be glad to include it.

7. P. 5606, L. 19-21: By calibrating the model on the first half of the series and validating it on the second half, the authors only applied half of the Klemeš split-sample test (Klemeš, 1986). It would be useful to also do the reverse test, by calibrating the model on the second half and validating it on the first half. This would provide a benchmark simulation in validation mode on all available data (not only half of them) and hence a more comprehensive evaluation of model performance. This would also make the comparison of the difference in model performance between calibration and validation more interesting: by comparing the mean performance in validation on the two periods with the mean performance in calibration on the two periods, one avoid the possible bias resulting from the fact that the two periods may not be similarly difficult/easy to simulate. Last this would give the opportunity to comment the stability of parameter values between the two calibration periods and hence possibly identify regions where model parameter identification appears more robust than others (this discussion could be added in section 4). (the dataset made available could therefore include two benchmark simulated series over the whole period, one using the parameter set calibrated on the first sub-period and one using the parameter set calibrated on the second sub-period.

We agree with this comment and think it is a great addition to the paper and dataset. The full Klemeš (1986) split sample calibration has been included in the dataset along with calibrations for two other forcing inputs (Maurer et al. 2002) and NLDAS-II (Xia et al. 2012) with some basic results in Figure 6 and discussion in the text. The various calibrated parameters and model output for the calibrations shown in Fig. 6 are also included in the dataset.

8. P. 5607, L. 4: As mentioned above, I do not understand the choice of this objective function, given its known limits (also acknowledge by the authors later in the text). Using a KGE-type objective function would also avoid useless discussions later in the article (section 4.2) on the limits of the proposed benchmark given the known problems of the selected objective function! Although I know other objective functions may be even more powerful, the advantage of KGE is that it remains very simple to compute. Note that I better understand however the selection of NSE as a criteria for model performance evaluation in this study to give this commonly-used performance reference.

See above discussion in response to major point D.

9. P. 5607, L. 18-22: It is useless to repeat in the text the information already given in the table.
We respectfully disagree on this point and feel it is worthwhile to discuss the calibrated parameters in the text along with the reasons why they were chosen.

10. P. 5607, L.22-25: This sentence is unclear.
We have tried to improve the clarity of this sentence.

11. P. 5608, L. 13-17: Indicate the units of each term of the equation.
We have included units of each term.

12. P. 5609, L. 1: What are these components?
We have noted the components.

13. P. 5609, L. 8-17: A similar climatological benchmark was advised long ago by Garrick et al. (1978) (see also Martinec and Rango, 1989). This could be mentioned. Why a 30-day smoothing window was deemed necessary compared to the simple reference proposed by Garrick et al. (1978) that simply uses the averaged measured discharge from past years for each day of the period? Is there any difference between these two references in terms of performance?
We thank the reviewer for this comment. It was very helpful looking at these two papers and we have included reference and discussion of them in the text. The 30-day smoothing window was used to provide a smoother daily long term climatology. Just using the 30 values on a given day may be highly influenced by one event, while the smoothed time series provides a historical benchmark more representative of a monthly climatology. We have not examined the difference, but there will likely be a small difference between the two approaches.

14. P. 5610, L. 4-14: This paragraph would probably be better placed at the end of section 2 with a more in-depth analysis of catchments characteristics (see comment above).
We have moved this paragraph and greatly expanded the discussion of the basin characteristics (see response to minor comment #1)

15. Section 4.2: Results on MNSE could be commented in the text.
We have included discussion of MNSE in the text.

16. P. 5611, L. 12-15: I do not agree with this argument. The fact that NSE is widely used does not justify that it should be used here, given it was demonstrated to be a bad choice for model calibration. I think this choice is even counterproductive for the community, since it will encourage a statu quo in the use of RMSE for model calibration if one wants to compare results with the proposed benchmark. I really think the use of KGE-type objective function should be encouraged. (note that I am not one of the developers of the KGE criterion, but I find it useful in practice).
We respectfully disagree with this assertion. Presenting results using RMSE as the objective function and using NSE as the main performance metric give the opportunity to use more advanced calibration methodologies

17. P. 5614, L. 2-6: This issue of data quality in climatic data may also be commented in section 2.2.

We have included mention of this in section 2.2

18. P. 5614, L. 23-25: Indicate if these are calibration or validation results.

We have clarified this statement

19. P. 5615, L. 4: What is a “low-order” hydrologic model?

We have clarified this statement

20. P. 5620, L. 29: Maybe not so useful to cite a paper in preparation if it is ultimately not published and therefore not possible to find it for readers.

We have removed mention of this paper.

21. Fig. 1.b: What RAIM stands for? An interesting graph would also be to plot the ratio of mean flow and mean precipitation ($y=Q/P$) as a function of the ratio of mean precipitation and mean potential evapotranspiration ($x=P/PET$). The graph could show the limit lines $y=1$ and $y=1-1/x$. The advantage of this graph is that it is based on observations only, whereas the graph shown by the authors uses model estimates.

RAIM stands for Rain plus Melt (RAIM). We have added explanation of the abbreviation in the figure caption. We have also added a panel of the proposed figure along with discussion in the text. Just to note, PET is not truly observations only as Daymet estimates shortwave down via regression equations found in MT-CLIM.

Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: Dataset characteristics and assessment of regional variability in hydrologic model performance

A. J. Newman[1], M. P. Clark[1], K. Sampson[1], A. Wood[1], L. E. Hay[2], A. Bock[2], R. J. Viger[2], D. Blodgett[3], L. Brekke[4], J. R. Arnold[5], T. Hopson[1] and Q. Duan[6]

[1] National Center for Atmospheric Research, Boulder CO, USA

[2] United States Geological Survey, Modeling of Watershed Systems, Lakewood CO, USA

[3]United States Geological Survey, Center for Integrated Data Analytics, Middleton WI, USA

[4]U.S. Department of Interior, Bureau of Reclamation, Denver CO, USA

[5] US Army Corps of Engineers, Institute for Water Resources, Seattle WA, USA

[6]Beijing Normal University, Beijing, China

Correspondence to: A. J. Newman (anewman@ucar.edu)

Abstract

We present a community dataset of daily forcing and hydrologic response data for 671 small- to medium-sized basins across the contiguous United States (median basin size of 336 km^2) that spans a very wide range of hydroclimatic conditions. Areally averaged forcing data for the period 1980-2010 was generated for three basin delineations -- basin mean, Hydrologic Response Units (HRUs) and elevation bands -- by mapping ~~the ~~daily, ~~1 km ~~gridded ~~Daymet~~ meteorological ~~dataset~~datasets to the sub-basin (Daymet) and basin polygons~~.~~ (Daymet, Maurer and NLDAS). Daily streamflow data was compiled from the United States Geological Survey National Water Information System. The focus of this paper is to (1) present the dataset for community use; and (2) provide a model performance benchmark using the coupled Snow-17 snow model and the Sacramento Soil Moisture Accounting conceptual hydrologic model, calibrated using the Shuffled Complex Evolution global optimization routine. After optimization minimizing daily root mean squared error, 90% of the basins have Nash-Sutcliffe Efficiency scores ~~>~~≥ 0.55 for the calibration period and 34% ≥ 0.8. This benchmark provides a reference level of hydrologic model performance for a commonly used model and calibration system, and highlights some regional variations in model performance. For example, basins with a more pronounced seasonal cycle generally have a negative low flow bias, while basins with a smaller seasonal cycle have a positive low flow bias. Finally, we find that data points with extreme error (defined as individual days with a high fraction of total error) are more common in arid basins with limited snow, and, for a given aridity, fewer extreme error days are present as basin snow water equivalent increases.

1. Introduction

With the increasing availability of gridded meteorological datasets, streamflow records and computing resources, large sample hydrology studies have become more common in the last decade or more (i.e.~~g.~~ Nathan and McMahon 1990; Perrin et al. 2001; Maurer et al, 2002; Beldring et al. 2003; Merz and Bloschl 2004; Andreassian et al. 2004; Lohmann et al. 2004; ~~Oudin~~Duan et al. 2006; Oudin et al. 2006; Oudin et al. 2010; Samaniego et al. 2010; Martinez and Gupta 2010; Nester et al. 2011; Martinez and Gupta 2011; Nester et al. 2012; Livneh and Lettenmaier 2012, 2013; Kumar et al. 2013; Oubeidillah et al. 2013). Within the United States there have been several studies to produce large sample hydrometeorological datasets (Maurer et al. 2002; Lohmann et al. 2004; Duan et al. 2006; Thornton et al. 2012; Xia et al. 2012; Livneh et al. 2013). Many of these datasets provide gridded data and may need to be further processed by the end user for their specific hydrologic model configuration. The Model Parameter Estimation Project (MOPEX) dataset does provide basin mean hydrometeorological data and observed streamflow records for 438 basins across the contiguous United States (Schaake et al. 2006) over 30+ years; making it one of the few, high quality, freely available hydrometeorological datasets with immediate applicability to catchment type hydrologic models.

Gupta et al. (2014) emphasize that more large-sample hydrologic studies are needed to "balance depth with breadth" – to wit, most hydrologic studies have traditionally focused on one or a small number of basins (depth), which hinders the ability to establish general hydrologic concepts applicable across regions (breadth~~) (Gupta et al. 2014).~~). Gupta et al. (2014) go on to discuss practical considerations for large sample hydrology studies, noting first and foremost that large datasets of quality basin data need to be available and shared in the community.

In support of this philosophy, we present a large-sample hydrometeorological dataset and modeling tools to understand regional variability in hydrologic model performance across the contiguous USA~~.~~ (Fig. 1). The development of the basin dataset presented herein takes advantage of high quality freely-available data from various US government agencies and research laboratories. It includes (1) daily forcing data for 671 basins for multiple delineations over the 1980-2010 time period; (2) daily streamflow data; (3) basic metadata (e.g. location, elevation, size, and basin delineation shapefiles) and (4) benchmark model performance which contains the final calibrated model parameter sets, model output timeseries for all basins as well as summary graphics for each basin. This builds on the MOPEX dataset by providing basin mean forcing data for 233 more basins along with two other spatial configurations and the benchmark model performance parameter sets and model output.

This dataset and benchmark application is intended for the community to use as a test-bed to facilitate the evaluation of hydrologic modeling and prediction questions. To this end, the benchmark consists of the calibrated, coupled Snow-17 snow model and the Sacramento Soil Moisture Accounting conceptual hydrologic model for all 671 basins using the Shuffled Complex Evolution global optimization routine. ~~We provide some basic analysis on how this~~

~~choice of hydrologic modeling method impacts regional variability in model performance.~~ Development of a large sample hydrologic dataset such as this will allow for exploration into many important scientific questions. We provide some basic analysis relating to questions such as: 1) What is the model performance across a large sample of basins and how does model performance vary across basin hydro-climatic conditions? 2) How do error characteristics relate to basin calibration performance and hydro-climatic conditions? This basic analysis is intended to highlight some of the important questions that can be answered through large-sample hydrologic studies and provide example results for further exploration.

The next section describes the development of the basin dataset from basin selection through forcing data generation. It then briefly describes the modeling system and calibration routine. Next, example results using the basin dataset and modeling platform are presented. Finally, concluding thoughts and next steps are discussed.

2. Basin Dataset

The development of a freely available large sample basin dataset requires several choices and subsequent data acquisition. Three major decisions were made and are discussed in this section: 1) the selection process for the basins, 2) the various basin delineations to be developed, and 3) selection of underlying forcing dataset used to develop forcing data ~~timeseries~~time series. Additionally, aggregation of the necessary streamflow data is described.

2.1 Basin Selection

The United States Geological Survey (USGS) developed an updated version of their Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II) in 2011 (Falcone et al. 2010; Falcone 2011). This database contains geospatial information for over 9,000 stream gages maintained by the USGS. As a subset of the GAGES-II database, a portion of the basins with minimal human disturbance (i.e. minimal land use changes or disturbances, minimal human water withdrawls) are noted as "reference" gages. A further sub-setting of the reference gages were made as a follow-on to the Hydro-Climatic Data Network (HCDN) 1988 dataset (Slack and Landwehr 1992). These gages, marked HCDN-2009 (Lins 2012), meet the following criteria: 1) have at least 20 years of complete flow data between 1990-2009 and were active as of 2009, 2) are a GAGES-II reference gage, c) have less than 5 percent imperviousness as measured by the National Land Cover Database (NLCD-~~2006~~2011, Jin et al. 2013), and d) passed a manual survey of human impacts in the basin by local Water Science Center evaluators (Falcone et al. 2010). There are 704 gages in the GAGES-II database that are considered HCDN-2009 across the contiguous United States (CONUS~~.~~). This study uses that portion of the HCDN-2009 basin set as the starting point since they should best represent natural flow conditions. After initial processing and data availability requirements, 671 basins are used for analysis in this study (Fig. 1b).~~.~~ Because these basins have minimal human influence they are almost exclusively smaller, headwater-type basins.

2.2 Forcing and Streamflow Data

Hydrologic models are run with a variety of spatial configurations, including entire watersheds (lumped), elevation bands, hydrologic response units (HRUs), or grids. For this dataset, forcing data were calculated (via areal averaging) for watershed, HRU and elevation band delineations. The basin delineations were created from the base national geospatial fabric for hydrologic modeling developed by the USGS Modeling of Watershed Systems (MoWS) group. The geospatial fabric is a watershed-oriented analysis of the National Hydrography Dataset that contains points of interest (e.g. USGS streamflow gauges), hydrologic response unit boundaries and simplified stream segments (not used in this study). This geospatial fabric contains points of interest that include USGS streamflow gauges and allowed for the determination of upstream total basin area and basin HRUs (Viger 2014; Viger and Bock 2014). A digital elevation model (DEM) was applied to the geospatial fabric dataset to create elevation contour polygon shapefiles for each basin. The USGS Geo Data Portal (GDP) developed by the USGS Center for Integrated Data Analytics (CIDA) (Blodgett et al. 2011) was leveraged to produce areally-weighted forcing data for the various basin delineations over our time period. The GDP performs all necessary spatial subsetting and weighting calculations and returns the areally weighted timeseries for the specified inputs.

The Daymet dataset was selected as the primary gridded meteorlogical dataset to derive forcing data for our streamflow simulations (Thornton et al. 2012). Daymet was chosen because of its high spatial resolution, a necessary requirement to more fully estimate spatial heterogeneity for basins in complex topography. Daymet is a daily, gridded (1x1 km) dataset over the CONUS and southern Canada and is available from 1980 to present. It is derived solely from daily observations of temperature and precipitation. The Daymet variables used here are daily maximum and minimum temperature, precipitation, shortwave downward radiation, day length, and humidity; additionally snow water equivalent is included (not used in this work). These daily values are estimated through the use of an iterative method dependent on local station density and the spatial convolution of a truncated Gaussian filter for station interpolation, and the Mountain Climate simulator (MT-CLIM) to estimate shortwave radiation and humidity (Thornton et al. 1997; Thornton and Running 1999; Thornton et al. 2000). Daymet does not include estimates of potential evapotranspiration (PET), a commonly needed input for conceptual hydrologic models or wind speed and direction. Therefore, PET was estimated using the Priestly-Taylor method (Priestly and Taylor 1972) and is discussed further in section 3. Data quality is an ever-present issue in hydrologic modeling, and while the input data to Daymet are subject to rigorous quality control checks (Durre et al. 2008; 2010) potential errors may remain (Menne et al. 2009; 2010; Oubeidillah et al. 2013). Additionally, the Maurer et al. (2002) and National Land Data Assimilation System (NLDAS) (Xia et al. 2012) 12 km gridded datasets were processed to provide daily forcing data for the basin lumped configuration, resulting in three distinct datasets available for future forcing data impact studies.

~~Hydrologic models are run with a variety of spatial configurations, including entire watersheds (lumped), elevation bands, hydrologic response units (HRUs), or grids. For this dataset, forcing data were calculated (via areal averaging) for watershed, HRU and elevation band delineations. The basin delineations were created from the base national geospatial fabric for hydrologic modeling developed by the USGS Modeling of Watershed Systems (MoWS) group. The geospatial fabric is a watershed-oriented analysis of the National Hydrography Dataset that contains points of interest (e.g. USGS streamflow gauges), hydrologic response unit boundaries and simplified stream segments (not used in this study). This geospatial fabric contains points of interest that include USGS streamflow gauges and allowed for the determination of upstream total basin area and basin HRUs~~. ~~A digital elevation model (DEM) was applied to the geospatial fabric dataset to create elevation contour polygon shapefiles for each basin. The USGS Geo Data Portal (GDP) developed by the USGS Center for Integrated Data Analytics (CIDA) (Blodgett et al. 2011) was leveraged to produce areally-weighted forcing data for the various basin delineations over our time period.~~ . ~~The GDP performs all necessary spatial subsetting and weighting calculations and returns the areally weighted timeseries for the specified inputs.~~

Daily streamflow data for the HCDN-2009 gages were obtained from the USGS National Water Information System server (http://waterdata.usgs.gov/usa/nwis/sw) over the same forcing data time period, 1980-2010. While the period 1980-1990 is not covered by the HCDN-2009 review, it was assumed that these basins would have minimal human disturbances in this time period as well. For the portion of the basins that do not have streamflow records back to 1980, analysis is restricted to the available data records. The USGS provides streamflow data flags to identify periods of estimated flow and are included here. However, other data quality information is unavailable without further investigation and not available in this dataset. For reference, 90% (604) of the basins have 20% or fewer flow days estimated and 75% (503 basins) have 10% or less flow values estimated.

The 671 basins span the entire CONUS and cover a wide range of hydro-climatic conditions. They range from wet, warm basins in the Southeast (SE) US to hot and dry basins in the Southwest (SW) US, to wet cool basins in the Northwest (NW) and dry cold basins in the intermountain (Rocky Mountains in Fig. 1a) western US. Figure 1b displays the basin annual precipitation (colored shading) along with symbols to denote rain and snow dominated basins. In terms of annual mean CDFs, Daymet estimated basin mean temperatures range from -2 ºC to 23 ºC with precipitation amounts of 0.7 to 9.4 mm day^{-1} (Fig. 2). Annual observed mean runoff ranges from 0.01 to 9.3 mm day^{-1} with PET estimates ranging from 1.9 to 4.8 mm day^{-1}. Interestingly, this implies that Daymet precipitation itself is not enough to balance the observed runoff in some basins and is consistent with other recent large sample hydrologic studies (Oubeidillah et al. 2013). Seasonal variations in these four variables are large as well, with some basins reaching mean winter time temperatures lower than -10 ºC and summer time mean temperatures higher than 25 ºC (not shown). The seasonal water balance varies greatly with some basins experiencing much higher precipitation and runoff rates in one season versus another (e.g. spring runoff peaks in mountain snowmelt dominated basins). As expected, PET varies seasonally with a minimum in winter and a maximum in summer.

Figure 3 gives cumulative density functions (CDFs) for various physical descriptors of the basin set. The basins range in size from roughly 1 to 25,800 km^2 with the median basin size being ~335 km^2 and have mean elevations spanning from nearly sea level (10 m) to high alpine elevations (3570 m) with a median elevation of 462 m. Notably, 75 basins have mean elevations > 2000 m. Corresponding to the large range of elevations in the basin set, the mean slopes vary considerably, spanning over 2 orders of magnitude from near zero to over 200 m km^{-1}. The basin set covers a wide range of basin shapes with aspect ratios ranging from 0.08 to ~11. Finally, there is a large range of forest covers across the basin set which may have implications for hydrologic similarity (Oudin et al. 2010) with 20% of the basins having less than (more than) 14% (98%) forest cover and the median basin having ~80% forest cover (NLCD-2011).

This basin set allows us to simulate a variety of energy and water limited basins with different snow storage, elevation, slope, and precipitation characteristics. Figure 4a shows runoff ratio (USGS streamflow/Daymet precipitation) versus the aridity index (Daymet Precipitation/PET). Immediately it can be seen that some basins lie above the water limit line (Y=1) indicating more runoff than precipitation and many basins are near it (Y > 0.9). In these cases the model calibration process would struggle to produce an unbiased calibration, or never in basins above the water limit, because the basic water balance requires nearly zero evapotranspiration (ET) or is not satisfied. This requires a modification to incoming precipitation, which is discussed in the next section. Not coincidentally, the basins near and above the water limit are colder basins (mean annual T < 10 °C) with frozen precipitation during colder months. Additionally, two basins lie to the right of the curved line (Y = 1 – 1/aridity) indicating a surplus of water. These basins may also require modifications to input precipitation, but it is less clear in this case as observations of precipitation are generally underestimates, especially in snowfall (e.g. Yang et al. 1998). Examining the basin set using model output terms in the Budyko framework, there are many energy limited basins with dryness ratios as small as 0.2 and many water limited basins with model estimated dryness ratios as large as 4.5 (Fig. 4b). Note that now no basins lie above the water limit, indicating bulk precipitation corrections were applied as needed during the calibration process. Examination of hydrometeorlogical forcing datasets across a large spatial extent through the lens of water & energy balance draws attention to gross errors in the forcing or streamflow datasets and permits any identified errors to be placed into spatial and temporal context, a benefit of large sample studies.

As noted above, no additional quality control was performed on the candidate basins before calibration. For completeness and to more fully highlight some of the benefits and tradeoffs made when performing large sample hydrologic studies, all basins are kept for analysis in this work.

3. Hydrologic modeling benchmark

As stated in the introduction, the intended purpose of this dataset is a test-bed to facilitate assessment of hydrologic modeling and prediction questions across broad hydroclimatic variations, and we focus here on providing a benchmark performance assessment for a widely used calibrated, conceptual hydrologic modeling system. This type of dataset can be used for many applications including evaluation of new modeling systems against a well ~~know~~known benchmark system over wide ranging conditions, or as a base for comprehensive predictability experiments exploring importance of meteorology or basin initial conditions~~ (e.g. Wood et al. 2014).~~. To this end, we have implemented and tested an initial model and calibration system described below, using the primary models and objective calibration approach that have been used by the US National Weather Service River Forecast Centers (NWSRFCs) in service of operational short-term and seasonal streamflow forecasting.

3.1 Models

The HCDN-2009 basins include those with substantial seasonal snow cover (Fig. ~~1~~1b), necessitating a snow model~~ is required~~ in addition to a hydrologic model. Within the NWSRFCs, the coupled Snow-17, Sacramento Soil Moisture Accounting Model (Snow-17 and SAC-SMA) system is used. Snow-17 is a conceptual air temperature index based snow accumulation and ablation model (Anderson 1973). It uses near surface air temperature to determine the energy exchange at the snow-air interface and the only time-varying inputs are typically air temperature and precipitation (Anderson 1973; Anderson 2002). The SAC-SMA model is a conceptual hydrologic model that includes representation of physical processes such as evapotranspiration, percolation, surface flow, sub-surface lateral flow. Required inputs to SAC-SMA are potential evapotranspiration and water input to the soil surface (Burnash 1973; Burnash 1995). Snow-17 runs first and determines the partition of precipitation into rain and snow and the evolution of the snowpack. Any rain, snowmelt or rain passing unfrozen through the snowpack for a given timestep becomes direct input to the SAC-SMA model. Finally, streamflow routing is accomplished through the use of a simple two-parameter, Nash-type instantaneous unit-hydrograph model (Nash 1957).

3.2 Calibration

We employed a split-sample calibration approach~~,~~, following Klemes (1986), assigning the first 15 years of available streamflow data for calibration and the remainder for validation then repeating the calibration using the last 15 years and the initial remaining period for validation; thus, approximately 5500 daily streamflow observations were used for each calibration. To initialize the model calibration moisture states on October 1st, we specified an initial wet SAC-SMA soil moisture state that was allowed to spin down to equilibrium for a given basin by running the first year of the calibration period repeatedly and assume no snow pack. This was done until all SAC-SMA state variables had minimal year over year variations, which is a spin-up approach used by the Project for Intercomparison of Land-Surface Process Schemes (e.g. Schlosser et al. 2000). Determination of optimal calibration sampling and spin-up procedures is an area of active research. Spin-up was performed for every parameter set specified by the optimization algorithm, then the model was integrated for the calibration period and the RMSE for that parameter set was calculated.

Objective calibration was done by minimizing the root mean squared error (RMSE) of daily modeled runoff versus observed streamflow using the Shuffled Complex Evolution (SCE) global search algorithm of Duan et al. (1992, 1993). The SCE algorithm uses a combination of probabilistic and deterministic optimization approaches that systematically spans the allowed parameter search space and also includes competitive evolution of the parameter sets (Duan et al. 1993). Prior applications to the SAC-SMA model have shown good results (Sorooshian et al. 1993; Duan et al. 1994). In the coupled Snow-17 & SAC-SMA modeling system, 35 potential parameters are available for calibration, of which we calibrated 20 parameters having either *a priori* estimates (Koren et al. 2000) or those found to be most sensitive following Anderson (2002) (Table 1). The SCE algorithm was run using 10 different random seed starts for the initial parameter sets for each basin, in part to evaluate the robustness of the optimum in each case, and the optimized parameter set with the minimum RMSE from the ten different optimization runs was chosen for evaluation.

For Snow-17, six parameters were chosen for optimization (Table 1): The minimum and maximum melt factors (*MFMIN*, *MFMAX)*, the wind adjustment for enhanced energy fluxes to the snow pack during rain on snow *(UADJ*~~) ,~~), the rain/snow partition temperature, which may not be 0°C *(PXTEMP)*, the snow water equivalent for 100% snow covered area (*SI)*, and the gauge catch correction term for snowfall only (*SCF)*. These six parameters were chosen because *MFMIN*, *MFMAX*, *UADJ*, *SCF*, and *SI* are defined as major model parameters by Anderson ~~2002, with the addition of*PXTEMP* is shown by Mizukami et al. 2013.~~ (2002). *PXTEMP* was also shown to be important in the Snow-17 model by Mizukami et al. (2013). The *SCF* is critical in many snow dominated basins as precipitation is generally underestimated in these types of basins (e.g. Yang et al. 1998) and is certainly underestimated in some basins in Daymet as shown in Figures 3 and 4.

The areal depletion curve (ADC) is considered a major parameter in Snow-17. However, to avoid expanding the parameter space by the number of ordinates on the curve (typically 10), we manually specified the ADC according to regional variations in latitude, topographic characteristics (e.g. plains, hills or mountains) and typical air mass characteristics (e.g. maritime polar, continental polar) (as suggested in Anderson, 2002). The remaining Snow-17 parameters were set in the same manner. Following the availability of *a priori* parameter estimates for SAC-SMA from a variety of datasets and various calibration studies with SAC-SMA (Koren et al. 2000; Anderson et al. 2006; Pokhrel and Gupta 2010; Zhang et al. 2012) 11 parameters from

SAC-SMA are included for calibration (Table 1). We use an instantaneous unit hydrograph, represented as a two-parameter Gamma distribution for streamflow routing (Sherman 1932; Clark 1945; Nash 1957; Dooge 1959), the parameters of which were inferred as part of calibration. .

Finally, the scaling parameter in the Priestly-Taylor PET estimate is also calibrated. The Priestly-Taylor (P-T) equation (Priestly and Taylor 1972) can be written as:

$$PET = \frac{a}{\lambda} \cdot \frac{s \cdot (R_n - G)}{s + \gamma} \tag{1}$$

Where λ (MJ kg^{-1}) is the latent heat of vaporization, R_n (MJ m^{-2} day^{-1}) is the net radiation estimated using day of year, all Daymet variables and equations to estimate the various radiation terms (Allen et al. 1998; Zotarelli et al. 2009), G (MJ m^{-2} day^{-1}) is the soil heat flux (assumed to be zero in this case), s (kPa °C^{-1}) is the slope of the saturation vapor pressure-temperature relationship, γ (kPa °C^{-1}) is the psychrometric constant and a (unitless) is the P-T coefficient. The P-T coefficient replaces the aerodynamic term in the Penman-Monteith equation and varies by the typical conditions of the area where the P-T equation is being applied with humid forested basins typically having smaller values and exposed arid basins having larger values (Shuttleworth and Calder 1979; Morton 1983; ASCE 1990). Thus the P-T coefficient was included in the calibration since it should vary from basin to basin.

4. Benchmark results

4.1 Assessment Objectives and Metrics

Assessment of the models will focus on overall performance across the basin set, regional variations, and error characteristics. Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970) and two of the decomposition components of NSE, variance bias (α) and total volume bias (β) (Gupta et al. 2009) are the first metrics examined in two variations. Because NSE scores model performance relative to the observed climatological mean, regions in which the model can track a strong seasonal cycle (large flow autocorrelation) perform relatively better when measured by NSE, and this seasonal enhancement may be imparted when using NSE as the objective function for both the calibration and validation phases (e.g. Schaefli et al. 2007). Additionally, basins with higher streamflow variance and frequent precipitation events have better model performance. Therefore, to give a more standardized picture of model performance across varying hydroclimatologies, the NSE was recomputed using the long-term monthly mean flow instead of mean flow (denoted MNSE hereafter), thus preventing climatological seasonality from inflating the NSE and more accurately ranking basins by the degree to which the model added value over climatology in response to weather events (Garrick et al. 1978; Martinec and Rango 1989; Schaefli et al. 2005). MNSE in this context is defined for each day of year (DOY) via a 31-day window centered on a given DOY. The long-term flow for that 31-day "month" is computed giving rise to a "monthly" mean flow. Using this type of climatology as the base for an NSE type analysis provides improved standardization in basins with large flow autocorrelations. This definition is similar to the one proposed by Garrick et al. (1978) but with the addition of the 31-day smoother, which is done to provide a smoother reference climatology.

Also, several other advanced, more physically based, metrics of model performance are provided. First, three diagnostic signatures based on the flow duration curve (FDC) from Yilmaz et al. (2008) are computed: 1) the top 2% flow bias, 2) the bottom 30% flow bias and 3) the bias of the slope of the middle portion (20-70 percentile) of the FDC. Second, examination of the time series of squared error contribution to the RMSE statistic was performed to highlight events in which the model performs ~~poorlyfollowing~~poorly following Clark et al. (2008). This analysis was performed to gauge the representativeness of performance metrics over the model record by using the sorted (highest to lowest) time series of squared error to identify the *N* number of the largest error days and determine their fractional error contribution to the total. Finally, we extend this analysis to introduce, a simple, normalized general error index for application and comparison across varying modeling and calibration studies. We coin the index, E50, the fraction of calibration points contributing 50% of the error~~ (Fig 7c).~~. This captures the number of points determining the majority of the error and thus the optimal parameter set.

4.2 ~~Overall Performance~~ Spatial variability

It is informative to examine spatial patterns of the aforementioned metrics to elucidate factors leading to weak (and strong) model performance. This also allows for identification of outlier basins and characterization of contributing factors (i.e. forcing or streamflow data issues or poor calibration). Poor performing basins are most common along the high plains and desert southwest (Fig 5a, section 3c). When examining MNSE (Fig 5b), basins with high non-seasonal streamflow variance and frequent precipitation events (SE and NW US) have the highest model MNSE, while most of the snowmelt dominated basins see MNSE scores reduced relative to NSE, particularly in the validation phase (Fig. 5c). This indicates that RMSE as an objective function may not be well suited for model calibration in basins with high flow autocorrelation (Kavetski and Fenicia 2011; Evin et al. 2014). This is confirmed by comparing Fig. 5d to Fig. 5c, basins with large flow autocorrelations (one week mean flow for example) generally have lower MNSE scores.

Areas with low validation NSE and MNSE scores have generally large biases when looking at FDC metrics as well (Fig. 6). Focusing on the high plains, high flow biases of ± 50% are common. Extreme negative low flow biases are also present along the high plains and desert SW along with a general model trend to have large negative FDC slope biases, consistent with a poorly calibrated model. For the 72% of basins with validation NSE > 0.55 (basins with yellow-green to dark red colors in Fig. 6a), there is no noticeable spatial pattern across CONUS in regard to high flow periods. However, basins with a more pronounced seasonal cycle (e.g. ~~The 671 basins span Northwest (NW) and dry cold basins in the intermountain western US (Fig. 1). This allows us to simulate a variety of energy and water limited basins with different snow storage, elevation,~~

~~slope, and precipitation characteristics. There are many energy limited basins with dryness ratios as small as 0.2 and many water limited basins with dryness ratios as large as 4.5 (Fig.~~ snowpack dominated watersheds, central West coast) generally have a negative low flow bias, while basins with a smaller seasonal cycle have a positive low flow bias (Fig. 6b). Correspondingly, basins with a pronounced seasonal cycle generally have a near zero or positive slope of the FDC bias, while basins with a smaller seasonal cycle have a negative slope bias (Fig. 6c).

Past applications similar conceptual snow and hydrologic modeling systems across the CONUS have shown comparable spatial performance patterns. Clark et al. (2008) applied many conceptual models to a subset of the MOPEX basin set and found poor performance in arid regions. Martinez and Gupta (2010), using a monthly water balance model found the best performance generally along the east coast, most of SE CONUS, and along the west coast with scattered good performance in the Rocky Mountains. They found that many basins along the High Plains and north side of the Appalachian Mountains perform poorly. They also note that arid regions have high variability error (variability bias term in KGE).

4.3 Cumulative Performance

Two basic cumulative thresholds for model performance are highlighted here, NSE values of 0.55 and 0.8. An NSE of 0.55 indicates some model skill, and an NSE of 0.8 suggests reasonably good model performance. ~~1b). As noted in section 2b, no additional quality control was performed on the candidate basins before calibration. For completeness and to highlight some of the tradeoffs made when performing large sample hydrologic studies, all basins are kept for analysis in this work.~~

For the calibration period, 90% (604) of the basins ~~produce~~have a NSE greater than 0.55, while 72% (484) of the basins had a validation period NSE > 0.55 (Fig ~~2a).~~7a). At the NSE > 0.8 level, 34% (225) basin models perform better during calibration and 12% (78) basin models meet that criteria during the validation phase.   When using MNSE, 85% and 57% (568 and 385) of the basins lie above 0.55 and 17% and 4% (114 and 29) of the basins lie above 0.8 during the calibration and validation phases. The decomposition of the NSE (Gupta et al. 2009) shows that and 90% of basins have a calibration (validation) model-observation flow correlation > 0.75 (0.68) and 30% (12%) of basins have a model-observation flow correlation > 0.9 (Fig 7b). However, nearly all basins have too little modeled variance (values less than one) for both the calibration and validation phases (Fig. ~~2b~~7c). The total volume biases are generally small with 94% (79%) of the basins having a calibration (validation) period total flow bias within 10% of observed (Fig. ~~2c~~7d). These are expected results when using RMSE for the objective function (Gupta et al. 2009) and reaffirm that our implementation of SCE is calibrating the model properly.

~~The~~Figure 8 highlights the full split sample approach for calibration following Klemes (1986). It is seen that the calibration and validation statistics give quite similar results regardless of which time period is used for calibration and validation using the Daymet data. This could indicate that both halves of the data are equally challenging to model with this modeling system. We have also included basin calibrations using the first 15 years only for the Maurer et al. (2002) and NLDAS-II (Xia et al. 2012) datasets. It can be seen that the Daymet forcing provides better model performance overall than both Maurer et al. and NLDAS forcing data. This likely relates to the coarser resolution of the Maurer et al. and NLDAS data (12 km) and the somewhat small basin sizes in this basin set. More importantly the inclusion of the Klemes (1986) split-sample approach provides users of this dataset two parameter estimates for each basin using different calibration periods, while the inclusion of three total forcing datasets begins to allow for ensemble type forcing data impact studies across a large basin sample size. In the remaining discussion, only model performance results using the first half of the split sample for calibration are presented.

With respect to advanced diagnostics, the model under predicts high flow events in nearly all basins during calibration and slightly less so for the validation period (Fig. ~~3a~~9a). This is an expected result when using RMSE as the objective function because the optimal calibration underestimates flow variability (Gupta et al. 2009). Low flow periods are more evenly over and under predicted (Fig. ~~3b~~9b) for both the calibration and validation time frames with 58% and 61% of basins having more modeled low flow. Finally, the bias in the slope of the FDC is generally under predicted with ~75% of basins having a negative model bias (FDC slope is negative, thus a negative bias indicates the model slope is more positive and that the modeled flow variability is too compressed). The slope of the FDC indicates the variance of daily flows, which primarily relate to the seasonal cycle or the “flashiness” of a basin. Again this indicates model variability is less than observed, at both short and longer time scales. In aggregate, these results agree with ~~Fig. 2~~Figure 5 and are expected based on the analysis of Gupta et al. (2009). Optimization using RMSE or NSE as the objective function generally results in under prediction of flow variance and near zero total flow bias (Fig ~~2~~7). This manifests itself in the simulated hydrograph as under predicted high flows, generally over predicted low flows and a more positive slope to the middle portion of the FDC (Fig~~ 3~~. 9). It is worth repeating that the goal of this initial application is to provide to community with a benchmark of model performance using well known models, calibration systems and widely used, simple objective functions, thus the use of RMSE.

~~*4.3 Spatial variability*~~

~~It is informative to examine spatial patterns of the aforementioned metrics to elucidate factors leading to weak (and strong) model performance. Poor performing basins are most common along the high plains and desert southwest (Fig 4a, section 3c). When examining MNSE (Fig 4b), basins with high non-seasonal streamflow variance and frequent precipitation events (Gulf Coast and Pacific NW) have the highest model MNSE, while most of the snowmelt dominated basins see MNSE scores reduced relative to NSE, particularly in the validation phase (Fig 2a, Fig. 4c). This indicates that RMSE as an objective function may not be well suited for~~

~~model calibration in basins with high flow autocorrelation (Kavetski and Fenicia 2011; Evin et al. 2014).~~

~~Areas with low validation NSE and MNSE scores have generally large biases when looking at FDC metrics as well (Fig 5). Focusing on the high plains, high flow biases of ± 50% are common. Extreme negative low flow biases are also present along the high plains and desert SW along with a general model trend to have large negative FDC slope biases, consistent with a poorly calibrated model. For the 72% of basins with validation NSE > 0.55 (basins with yellow-green to dark red colors in Fig. 5a), there is no noticeable spatial pattern across CONUS in regard to high flow periods. However, basins with a more pronounced seasonal cycle (e.g. snowpack dominated watersheds, central California) generally have a negative low flow bias, while basins with a smaller seasonal cycle have a positive low flow bias. Correspondingly, basins with a pronounced seasonal cycle generally have a near zero or positive slope of the FDC bias, while basins with a smaller seasonal cycle have a negative slope bias.~~

4.4 Error Characteristics

When examining fractional error statistics for the basin set, 15 basins have single days that contribute at least half the total squared error~~,~~ (potential outlier basins), whereas at the median, the largest error day contributes 8.3% of the total squared error for the median basin (Fig ~~6~~10). The fractional error contribution for the 10, 100 and 1000 largest error days for the median basin are 33%, 70% and 96% of the total squared error respectively. This indicates that for nearly all basins, there are 100 or fewer points that drive the RMSE and therefore optimal model parameters. This type of analysis can be undertaken for any objective function to identify the most influential points and allow for more in-depth examination of forcing data, streamflow records, calibration strategies (i.e. Kavetski et al. 2006; Vrugt et al. 2008; Beven and Westerberg 2011; Beven et al. 2011; Kauffeldt et al. 2013), or if different model physics are warranted.

The spatial distribution of fractional error contributions show that the issue of model performance being explained by a relatively small set of days is more prevalent in arid regions of CONUS (desert SW US and high plains) as well as basins slightly inland from the east coast of CONUS. (Fig ~~7a~~11a-b). The arid basins are generally dry with sporadic high precipitation (and flow) events, while the Appalachian basins are wetter (Fig. ~~1~~1b) with extreme precipitation events interspersed throughout the record. Basins with significant snowpack tend to have lower error contributions from the largest error days (Fig. ~~7a~~11a-b). The E50 metric highlights mean peak snow water equivalent (SWE) and frequent precipitation basins as well. These regions contain and order of magnitude more days than the high plains and desert SW, giving insight into how representative of the entire streamflow timeseries the optimal model parameter set really is.

Additionally, ranking the basins using their fractional error characteristics provides a similar insight. As the aridity index increases, the fractional error contribution increases for basins with little to no mean peak SWE. For basins with significant SWE, the fractional error contribution decreases with increasing aridity (Fig. ~~8~~12). Alternatively, for a given aridity index the fractional error contribution for *N* days will decrease with increasing SWE. This dynamic arises because more arid basins with SWE produce a relatively greater proportion of their runoff from snowmelt, without intervening rainfall. This implies that the optimized model produces a more uniform error distribution with less heteroscedasity in basins with more SWE. Moreover, as the fractional error contribution for the 10 largest error days increases, model NSE generally decreases in the validation phase (Fig. ~~9~~13). This indicates fractional error metrics are related to overall model performance and that calibration methods to reduce extreme error days should improve model performance. This is not unexpected due to the fact that the residuals from an RMSE type calibration are heteroscedastic. Arid basins typically have few high flow events, which are generally subject to larger errors when minimizing RMSE. Using advanced calibration methodologies that account for heteroscedasticy (Kavetski and Fenicia 2011; Evin et al. 2014) may produce improved calibrations for arid basins in this basin set and provide different insights into model behavior using this type of analysis.

4.5 Limitations and ~~uncertainties~~Uncertainties

One interesting example of the usefulness (and a potential limitation) of large sample hydrology stemming from this work lies in the identification of issues with forcing datasets. Figures 3 and 4 show Daymet has too little precipitation in certain regions which is also seen in Oubeidillah et al. (2013). When examining calibrated model performance in the Pacific Northwest, it is seen that several basins along the ~~Olympic Peninsula~~west coast have low outlier NSE scores. Tracing this unexpected result, we find the Daymet forcing data available for those basins has a negative temperature bias, preventing mid-winter rain and melt episodes in the modeling system, identifying scope to improve the Daymet forcing. Moreover, winter periods of observed precipitation and streamflow rises coincide with subzero T_{max} in the Daymet dataset, also suggesting areas to improve the Daymet forcing. The large sample of basins in this region (91) allowed for identification of the outlier basins and the underlying causes.

This ~~limits~~may also limit interpretation of these results and other large sample hydrologic studies. As noted by Gupta et al. (2014), large sample hydrology requires a tradeoff between breadth and depth. The lack of depth ~~inhibits~~may inhibit discovery and identification of all data quality issues and ~~introduces~~ the underlying causes of outliers in any analysis (e.g. Fig ~~9~~13). Explanation of these outliers is sometimes difficult and not complete in the initial development and analysis due to the lack of ~~familiarty~~familiarity with ~~those~~ specific basins and any forcing or validation data peculiarities. However, providing forcing data, model parameters and model output permits additional focused studies and helps reduce these limitations. Additional prescreening using the methods of Martinez and Gupta (2011) can also help identify outliers due to data quality issues and help identify basins and regions where model physics errors are present.

5. Summary and Discussion

Most hydrologic studies focus in detail on a small number of watersheds, providing comprehensive but highly local insights, and may be limited in their ability to inform general hydrologic concepts applicable across regions (Gupta et al. 2014). To facilitate large-sample hydrologic studies, large-sample basin datasets and corresponding benchmarks of model performance using standard methodology across all basins need to be freely available to the community. To that end, we have compiled a community dataset of daily forcing and streamflow data for 671 basins and provide a benchmark of performance using a widely used conceptual a hydrologic modeling and calibration scheme over a wide range of conditions.

Overall, application of the basin set to assessing an objectively calibrated conceptual hydrologic model representation of the 671 watersheds yielded calibration Nash-Sutcliffe Efficiency (NSE) scores of > 0.55 (0.8) for 90 (34) percent of the basins. Performance of the models varied regionally, and the main factors influencing this variation were found to be aridity and precipitation intermittency, contribution of snowmelt, and runoff seasonality. Analysis of the cumulative fractional error contributions from the largest error days showed that the presence of significant snow water equivalent (SWE) offset the negative impact of increasing aridity on simulation performance. This study has identified potential outlier basins for this modeling system and has provided insights into potential forcing data limitations. Although this modeling application utilized ~~low-order~~a conceptual hydrologic ~~models~~model with a single-objective calibration strategy, the findings provide a baseline for assessing more complex strategies in each area, including multi-objective calibration of more highly distributed hydrologic models (e.g., in Shi et al 2008).~~ The dataset and model demonstration also provides a starting point for hydrologic prediction experiments (e.g. Wood et al. (2014), which utilized 425 of the models to investigate the sources of seasonal streamflow prediction skill).~~ The unusually broad variation of hydroclimatologies represented by the dataset, which contains ~~forcings~~forcing and streamflow data obtained by consistent methodology and retains outlier basins, makes it a notable resource for these and other future large-sample watershed-scale hydrologic analysis efforts.

This dataset and applications presented are made available to the community. ~~(see http://ral.ucar.edu/projects/hap/flowpredict/subpages/modelvar.php~~(see http://ral.ucar.edu/projects/hap/flowpredict/subpages/modelvar.php or http://dx.doi.org/10.5065/D6MW2F4D)

6. Acknowledgements

This work is funded by the ~~United States~~US Army Corps of Engineers Climate Preparedness and Resilience Programs and the US Department of the Interior Bureau of Reclamation. The authors would like to thank the USGS Modeling of Watershed Systems (MoWS) group, specifically for providing technical support and the national geospatial fabric data to generate all the basin delineations. We would also like to thank Jordan Read and Tom

Kunicki of the USGS Center for Integrated Data Analytics for their help with the USGS Geodata Portal.

7. References

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1988). Crop evapotranspiration: guidelines for computing crop water requirements. Food and Agriculture Organization of the United Nations, Rome, 15 pp.

Anderson, E. A. (1973). National Weather Service River Forecast System – Snow accumulation and ablation model. NOAA Technical Memorandum, NWS, HYDRO-17, US Department of Commerce, Silver Spring, MD, 217 pp.

Anderson, E. A. (2002). Calibration of conceptual hydrologic models for use in river forecasting. NOAA Technical Report, NWS 45, Hydrology Laboratory, Silver Spring, MD.

Anderson, R. M., V. I. Koren, S. M. Reed (2006). Using SSURGO data to improve Sacramento Model a priori parameter estimates. Journal of Hydrology, 320, 103-116.

Andreassian, V., A. Oddos, C. Michel, F. Anctil, C. Perrin and C. ~~:~~Loumange (2004). Impact of spatial aggregation of inputs and parameters on the efficiency of rainfall-runoff models: A theoretical study using chimera watersheds. Water Resources Research, 40(5): W05209, doi: 10.1029/2003WR002854.

Beldring, S., Engeland, K., Roald, L. A., Saelthun, N. R., and Vokso, A (2003). Estimation of Parameters in a distributed precipitation-runoff model for Norway, Hydrology and Earth System Sciences, 7, 304G316.

Beven, K. and I. Westerberg (2011). On red herrings and real herrings: disinformation and information in hydrological inference. Hydrologic Processes, 25, 1676-1680.

Beven, K, P. J. Smith, and A. Wood (2011). On the colour and spin of epistemic error (and what we might do about it). Hydrology and Earth System Sciences, 15, 3123-3133, doi:10.5194/hess-15-3123.

Blodgett, D. L., N. L. Booth, T. C. Kunicki, J. L. Walker, and R. J. Viger (2011). Description and testing of the geo data portal: A data integration framework and web processing services for environmental science collaboration. US Geological Survey, Open-File Report 2011-1157, 9 pp., Middleton WI, USA.

Burnash, R.J.C., R. L. Ferral, R. A. McGuire (1973). A generalized streamflow simulation system conceptual modeling for digital computers, U. S. Department of Commerce National Weather Service and State of California Department of Water Resources.

Burnash, R. J. C. (1995). The NWS River Forecast System – Catchment model. In Computer Models of Watershed Hydrology, edited by V. P. Singh, pp. 311-366, Water Resources Publications, Highlands Ranch, Colo.

Clark, C. O. (1945). Storage and the unit hydrograph. Proc. Am. Soc. Civ. Eng., vol. 9, pp 1333-1360.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrologic models. Water Resources Research, 44, W00B02, doi:10.1029/2007WR006735.

Clark, M. P., D. Kavetski, and F. Fenicia (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47, W09301, doi:10.1029/2010WR009827.

Dooge, J. C. I. (1959). A general theory of the unit hydrograph. Journal of Geophysical Research, 64(2): 241-256.

Duan, Q., S. Sorooshian, and V. K. Gupta (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research, 28(4): 1015-1031.

Duan, Q., V. K. Gupta, and S. Sorooshian (1993). A shuffled complex evolution approach for effective and efficient optimization. Journal of Optimization Theory and Applications, 76(3): 501-521.

Duan, Q., S. Sorooshian, V. K. Gupta (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of Hydrology, 158, 265-284.

Duan, Q., J. Schaake, V. Andreassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, F. Habets, A. Hall, L. Hay, T. Houge, M. Huang, G. Leavesley, X. Liang, O. N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E. F. Wood (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrology*, **320**, 3-17.

Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014). Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research, 50, 2350-2375, doi: 10.1002/2013WR014185.

Falcone. J. A., D. M. Carlisle, D. M. Wolock, and M. R. Meador (2010). GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. Ecology, 91(2), p. 621. A data paper in Ecological Archives E091-045-D1, available at http://esapubs.org/Archive/ecol/E091/045/metadata.htm, (last access: 05 April 2014), 2010.

Falcone, J. A. (2011). GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. Digital spatial data set 2011. Available at: http://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml (last access: 10 Oct 2013), 2011.

Garrick, M., C. Cunnane, and J. E. Nash (1978). A criterion of efficiency for rainfall-runoff models. *J. Hydrology*, **36**(3-4), 375-381.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez-Barquero (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. Journal of Hydrology, 377, 80-91, doi:10.1016/j.jhydrol.2009.08.003.

Gupta, H. V., C. Perrin, R. Kumar, G. Bloschl, M. Clark, A. Montanari, and V. Andreassian (2014). Large-sample hydrology: A need to balance depth with breadth. Hydrology and Earth System Sciences-Earth System Discussions.

Jensen, M. E., R. D. Burman, and R. G. Allen (1990). Evapotranspiration and irrigation water requirements. American Society of Civil Engineers, ASCE Manual and Reports on Engineering Practice, 332 p., New York, NY.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., and Xian, G. 2013. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. *Remote Sensing of Environment*, 132: 159 – 175.

Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. K. Westerberg (2013). Disinformative data in large-scale hydrological modeling. Hydrology and Earth System Sciences, 17, 2845-2857, doi:10.5194/hess-17-2845-2013.

Kavetski, D., and F. Fenicia (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. Water Resources Research, 47, W11511,doi:10.1029/2011WR010748.

Kavetski, D., G. Kuczera, and S. W. Franks (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. Water Resources Research, 42, W03407, doi:10.1029/2005WR004376.

Klemes, V. (1986). Operational testing of hydrological simulation models. *Hydrol. Sci. J.*, **31**(1), 13-24.

Koren, V. I., M. Smith, D. Wang, and Z. Zhang (2000). Use of soil property data in the derivation of conceptual rainfall-runoff model parameters. American Meteorological Society 15th Conference on Hydrology, Long Beach, CA, pp. 103-106.

Kumar, R., L. Samaniego, and S. Attinger (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resources Research, 49, 360-379, doi:10.1029/2012WR012195.

Lins, H. F. (2012). USGS Hydro-Climatic Data Network 2009 (HCDN-2009), U. S. Geological Survey, Fact Sheet 2012-3047, Reston VA, USA.

Livneh, B. and D. P. Lettenmaier (2012). Multi-criteria parameter estimation for the Unified Land Model, Hydrology and Earth System Sciences, 16, 3029-3048, doi:10.5194/hess-16-3029-2012.

Livneh, B. and D. P. Lettenmaier (2013). Regional parameter estimation for the Unified Land Model. Water Resources Research, 49, 100-114, 10.1029/2012WR012220.

Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, D. P. Lettenmaier (2013): A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions. *J. Climate*, **26**, 9384–9392. doi: http://dx.doi.org/10.1175/JCLI-D-12-00508.1

Lohmann, D., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R.T., and Tarpley, J.D. (2004). Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. Journal Geophysical Research, 109, D07S91, doi:10.1029/2003JD003517.

Martinec, J. and A. Rango (1989). Merits of statistical criteria for the performance of hydrological models. *Water Resources Bulletin*, **25**(2), 421-432.

Martinez, G. and H. V. Gupta (2010). Toward improved identification of hydrologic models: A diagnostic evaluation of the “abcd” monthly water balance model for the conterminous United States. *Water Resour. Res.*, **46**, W08507, doi:10.1029/2009WR008294.

Martinez, G. and H. V. Gupta (2011). Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resour. Res.*, **47**, W12540, doi:10.1029/2011WR011229.

Maurer, E. P, A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen (2002). A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States. Journal of Climate, 15(22), 3237-3251.

Merz, R. and G. Bloschl (2004). Regionalization of catchment model parameters. Journal of Hydrology, 287(1-4): 95-123.

Mizukami, N., V. Koren, M. Smith, D. Kingsmill, Z. Zhang, B. Cosgrove, and Z. Cui (2013). The impact of precipitation type discrimination on hydrologic simulation: Rain-snow partitioning derived from HMT-West radar-detected brightband height versus surface temperature data. Journal of Hydrometeorology, 14, 1139-1158, doi:10.1175/JHM-D-12-035.1.

Morton, F. I. (1983). Operational estimates of actual evapotranspiration and their significance to the science and practice of hydrology. Journal of Hydrology, 66: 1-76.

Nash, J. E. (1957). The form of the instantaneous unit hydrograph. International Association of Scientific Hydrology Publication, 45(3), 114-121, Toronto ON, CA.

Nash, J. E., and J. V. Sutcliffe (1970). River flow forecasting through conceptual models. Part I: A discussion of principles. Journal of Hydrology, 10(3), 282-290, doi:10.1016/0022-1694(70)90255-6.

Nathan, R. J., and T. A. McMahon (1990). The SFB model, Part I – Validation of fixed model parameters. Civil Engineering Transactions, CE32, 157-161.

Nester, T., R. Kirnbauer, D. Gutknecht and G. Bloschl (2011). Climate and catchment controls on the performance of regional flood simulations. Journal of Hydrology, 402, 340-356.

Nester, T., R. Kirnbauer, J. Parajka and G. Bloschl (2012). Evaluating the snow component of a flood forecasting model. Hydrology Research, 43(6), 762-779.

Oubeidillah, A. A., S.-C. Kao, M. Ashfaq, B. Naz, and G. Tootle (2013). A large-scale, high-resolution hydrological model parameter dataset for climate change impact assessment for the conterminous United States. Hydrology and Earth System Sciences, 10, 9575-9613, doi:10.5194/hessd-10-9575-2013.

Oudin, L., V. Andreassian, T. Mathevet, C. Perrin, and C. Michel (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resources Research, 42(7): W07410, doi:10.1029/2005WR004636.

Oudin, L., A. L. Kay, V. Andreassian, and C. Perrin (2010). Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research, 46, W11558, doi:10.1029/2009WR008887.

Perrin, C., C. Michel, and V. Andreassian (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrology*, 242(3-4), 275-301, doi:210.1016/S0022-1694(1000)00393-00390.

Pokhrel, P. and H. V. Gupta (2010). On the use of spatial regularization strategies to improve calibration of distributed watershed models. Water Resources Research, 46, W01505, doi:10.1029/2009WR008066.

Priestly, C. H. B. and R. J. Taylor (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. Monthly Weather Review, 100:81-82.

Samaniego, L., A. Bardossy, and R. Lumar (2010). Streamflow prediction in ungauged catchments using copula-based dissimilarity measures. Water Resources Research, 46, W02506, doi:10.1029/2008WR007695.

Schaake, J., S. Cong, Q. Duan (2006). U.S. MOPEX data set. Report UCRL-JRNL-221228, Lawrence Livermore National Laboratory, Livermore CA, USA. Available online at: https://e-reports-ext.llnl.gov/pdf/333681.pdf (last access: 10 September 2014).

Schaefli, B., B. Hingray, M. Niggli, and A. Musy (2005). A conceptual glacio-hydrological model for high mountainous catchments. *Hydrology and Earth System Sciences*, **9**, 157-171.

Schaefli, B., and H. V. Gupta (2007). Do Nash values have value~~?.~~? Hydrological Processes, 21, 2075-2080, doi:10.1002/hyp.6825.

Schlosser, C.A., Slater, A.G., Robock, A., Pitman, A.J., Vinnikov, K.Y., Henderson-Sellers, A., Speranskaya, N.A., Mitchell, K., and the PILPS 2(d) contributors (2000). Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). Monthly Weather Review, 128, 301-321.

Sherman, L. K. (1932). Streamflow from rainfall by the unit graph method. Eng. News Rec., 108, 501-505.

Shi, X, A. W. Wood, and D. P. Letenmaier (2008). How essential is hydrologic model calibration to seasonal streamflow forecasting? Journal of Hydrometeorology, 9, 1350-1363.

Shuttleworth, W. J., and I. R. Calder (1979). Has the Priestly-Taylor equation any relevance to forest evaporation? Journal of Applied Meteorology, 18, 639-646.

Slack, J. R., and J. M. Landwehr (1992). Hydro-Climatic Data Network (HCDN): A U. S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988. U. S. Geological Survey, Open-File Report 92-129, Reston VA, USA.

Sorooshian, S., Q. Duan, and V. K. Gupta (1993). Calibration of conceptual rainfall-runoff models using global optimization: application to the Sacramento soil moisture accounting model. Water Resources Research, 29(4): 1185-1194.

Thornton, P. E., S. W. Running, and M. A. White (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology, 190: 214-251. http://dx.doi.org/10.1016/S0022-1694(96)03128-9.

Thornton, P. E., and S. W. Running (1999). An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity and precipitation. Agriculture and Forest Meteorology, 93:211-228.

Thornton, P. E., H. Hasenauer, and M. A. White (2000). Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: An application over complex terrain in Austria. Agricultural and Forest Meteorology, 104:255-271.

Thornton, P. E., M. M. Thornton, B. W. Mayer, N. Wilhelmi, Y. Wei, and R. B. Cook (2012). Daymet: Daily surface weather on a 1 km grid for North America, 1980-2012. Acquired online (http://daymet.ornl.gov/) on 15/07/2013 from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. http://doi:10.3334/ORNLDAAC/Daymet_V2

Viger, R. J, and A. Bock (2014). GIS Features of the Geospatial Fabric for National Hydrologic Modeling, US Geological Survey, http://dx.doi.org/doi:10.5066/F7542KMD

Viger, R. J. (2014). Preliminary spatial parameters for PRMS based on the Geospatial Fabric, NLCD2001 and SSURGO, US Geological Survey, http://dx.doi.org/doi:10.5066/F7WM1BF7

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research, 44, W00B09, doi:10.1029/2007WR006720.

~~Wood, A. W., Hopson, T., Newman, A.J., Arnold, J.R., Brekke, L., and Clark, M.P.: A variational ensemble streamflow prediction assessment approach for quantifying streamflow forecast skill elasticity. In preparation, 2014.~~

Xia, Y., K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo, C. Alonge, H. Wei, J. Meng, B. Livneh, D. Lettenmaier, V. Koren, Q. Duan, K. Mo, Y. Fan and D. Mocko (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.*, **117**, D03109, doi:10.1029/2001JD016048.

Ylimaz, K. K., H. V. Gupta, and T. Wagener (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research, 44, W09417, doi:10.1029/2007WR006716.

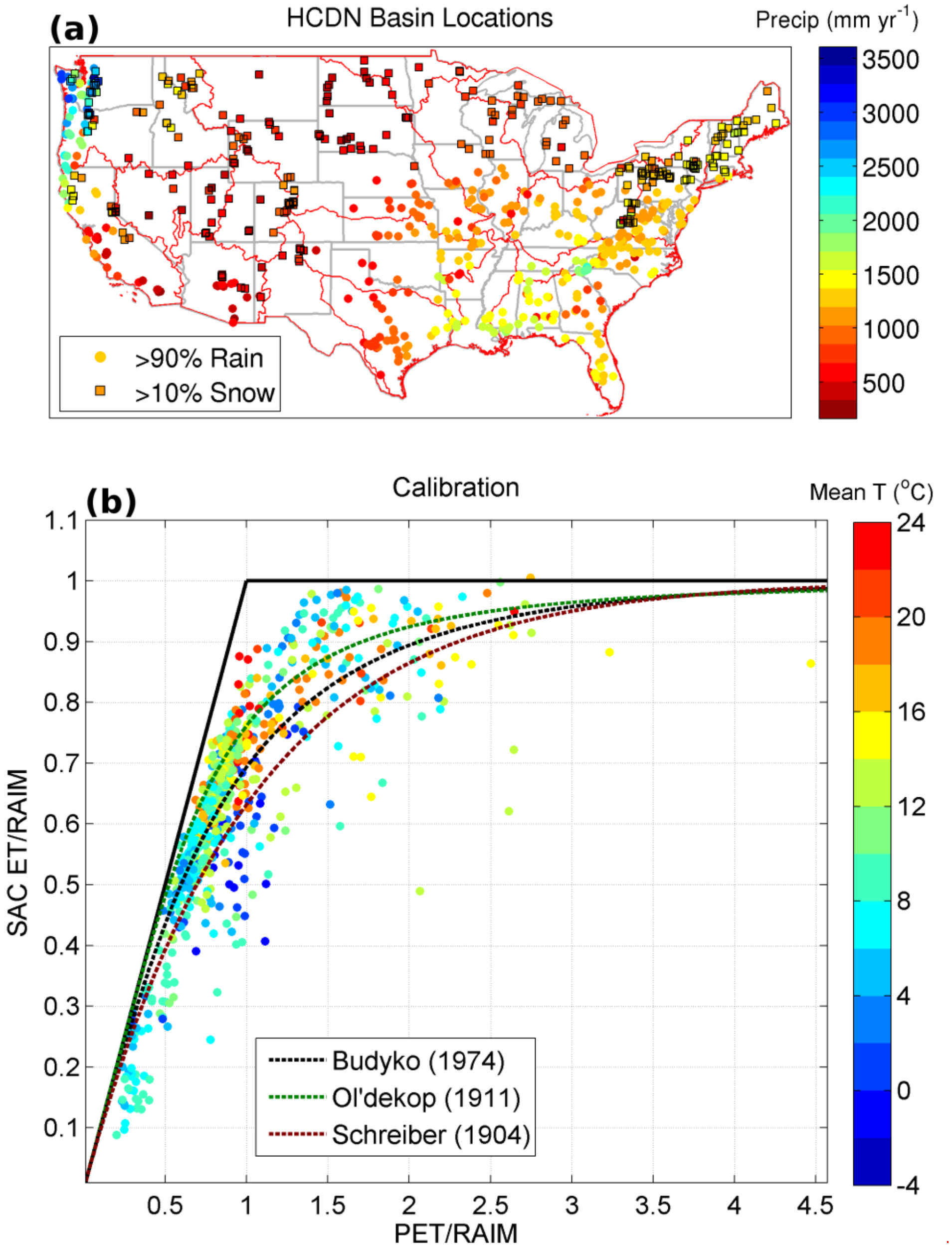
Zhang, Z., V. Koren, S. Reed, M. Smith, Y. Zhang, F. Moreda, and B. Cosgrove (2012). SAC-SMA a priori parameter differences and their impact on distributed hydrologic model simulations. Journal of Hydrology, 420-421 (2012), 216-227.

Zotarelli, L., M. D. Dukes, C. C. Romero, K. W. Migliaccio, and K. T. Morgan (2009). Step by step calculation of the Penman-Monteith Evapotranspiration (FAO-56 Method). University of Florida Extension, AE459, http://edis.ifas.ufl.edu (last access: 01 April 2014), 10 pp.

Table 1. Table describing all parameters calibrated and their bounds for calibration.

Parameter	Description	Units	Calibration Range
	Snow-17		
MFMAX	Maximum melt factor	mm ºC^{-1} 6-hr^{-1}	0.8 – 3.0
MFMIN	Minimum melt factor	mm ºC^{-1} 6-hr^{-1}	0.01– 0.79
UADJ	Wind adjustment for enhanced flux during rain on snow	km 6-hr^{-1}	0.01– 0.40
SI	SWE for 100% snow covered area	mm	1.0 – 3500.0
SCF	Snow gauge undercatch correction factor	-	0.1 – 5.0
PXTEMP	Temperature of rain/snow transition	ºC	-1.0 – 3.0
	SAC-SMA		
UZTWM	Upper zone tension water maximum storage	mm	1.0 – 800.0
UZFWM	Upper zone free water maximum storage	mm	1.0 – 800.0
LZTWM	Lower zone tension water maximum storage	mm	1.0 – 800.0
LZFPM	Lower zone free water primary maximum storage	mm	1.0 – 1000.0
LZFSM	Lower zone free water secondary maximum storage	mm	1.0 – 1000.0
UZK	Upper zone free water lateral depletion rate	day^{-1}	0.1 – 0.7
LZPK	Lower zone primary free water depletion rate	day^{-1}	0.00001 – 0.025
LZSK	Lower zone secondary free water depletion rate	day^{-1}	0.001 – 0.25
ZPERC	Maximum percolation rate	-	1.0 – 250.0
REXP	Exponent of the percolation equation	-	0.0 – 6.0
PFREE	Fraction percolating from upper to lower zone free water storage	-	0.0 – 1.0
	Others		
USHAPE	Shape of unit hydrograph	-	1.0 – 5.0
USCALE	Scale of unit hydrograph	-	0.001 – 150.0
PT	Priestly-Taylor coefficient	-	1.26 – 1.74

Figures


(a)
HCDN Basin Locations
Precip (mm yr^{-1})
3500
3000
2500
2000
1500
1000
500
● >90% Rain
■ >10% Snow
(b)
Calibration
Mean T (°C)
24
20
16
12
8
4
0
-4
SAC ET/RAIM
PET/RAIM
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Budyko (1974)
Ol'dekop (1911)
Schreiber (1904)

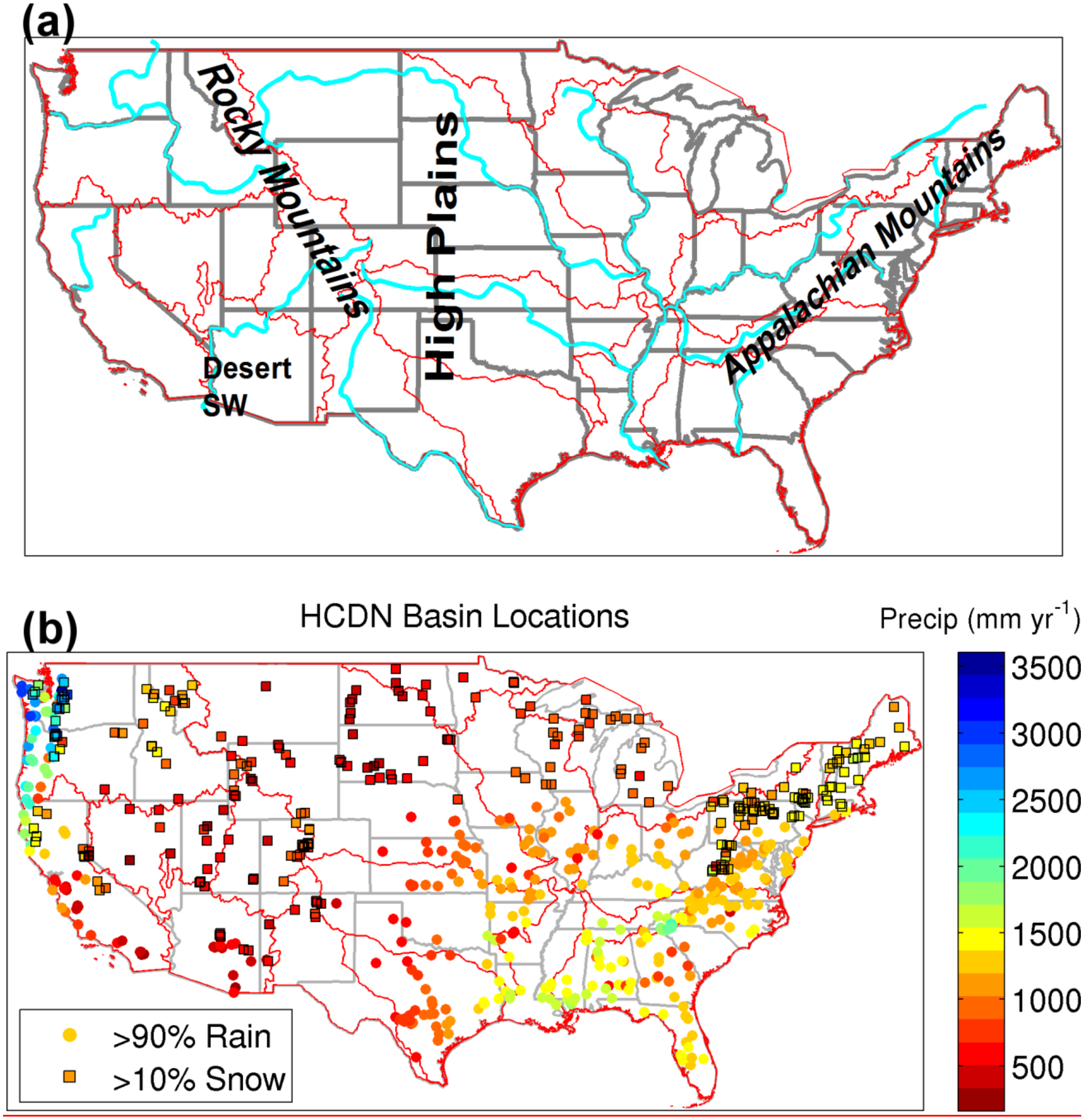
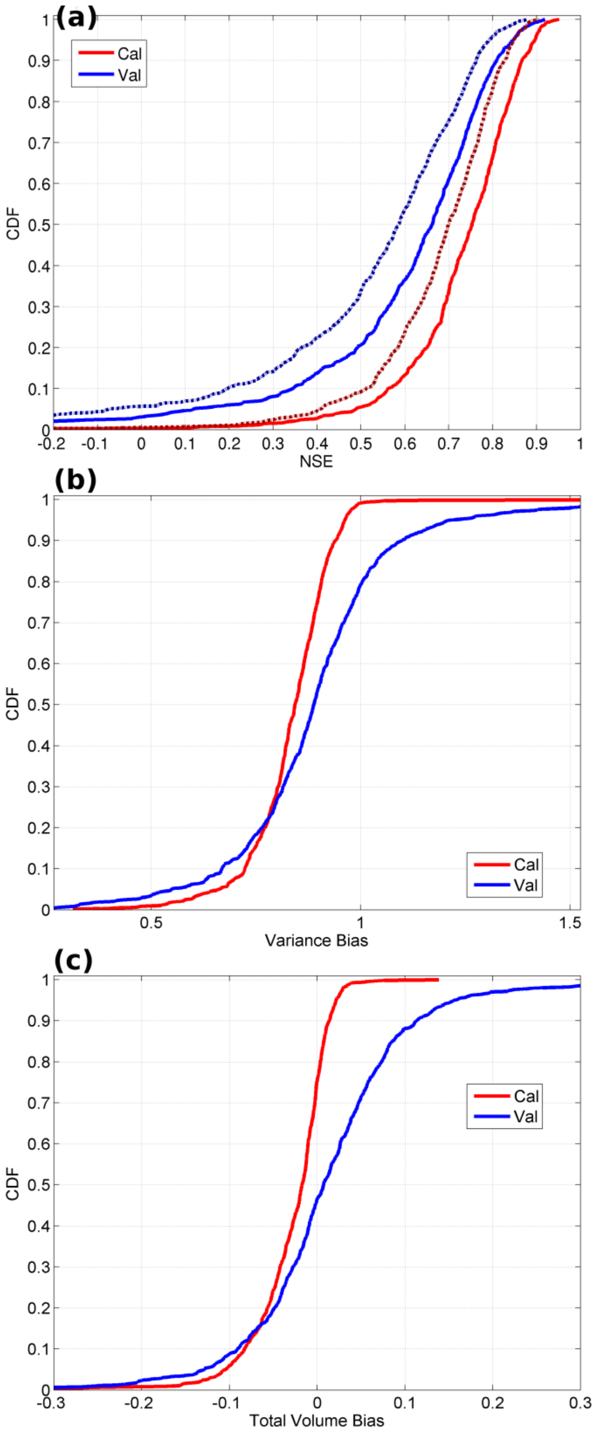


Figure 1. (a) Contiguous United States (US) with states (gray), rivers (blue) and major hydrologic regions (red). Text indicates major geographic regions discussed in text. (b) Location of the 671 HCDN-2009 basins across the contiguous ~~United States~~US used in the basin dataset with precipitation shaded. Circles denote basins with > 90% of their precipitation falling as rain, squares with black outlines denote basins with > 10% of their precipitation falling as snow as determined by using a 0°C daily mean Daymet temperature threshold. State outlines are in thin gray and hydrologic regions in thin red. ~~(b) Model derived Budyko analysis for the 671 basins with basin mean temperature shaded (colored dots) and three derivations of the Budyko curve (dashed lines).~~


(a)
Cal
Val
CDF
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
NSE
(b)
Cal
Val
CDF
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1
1.5
Variance Bias
(c)
Cal
Val
CDF
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Total Volume Bias

~~Figure 2.~~

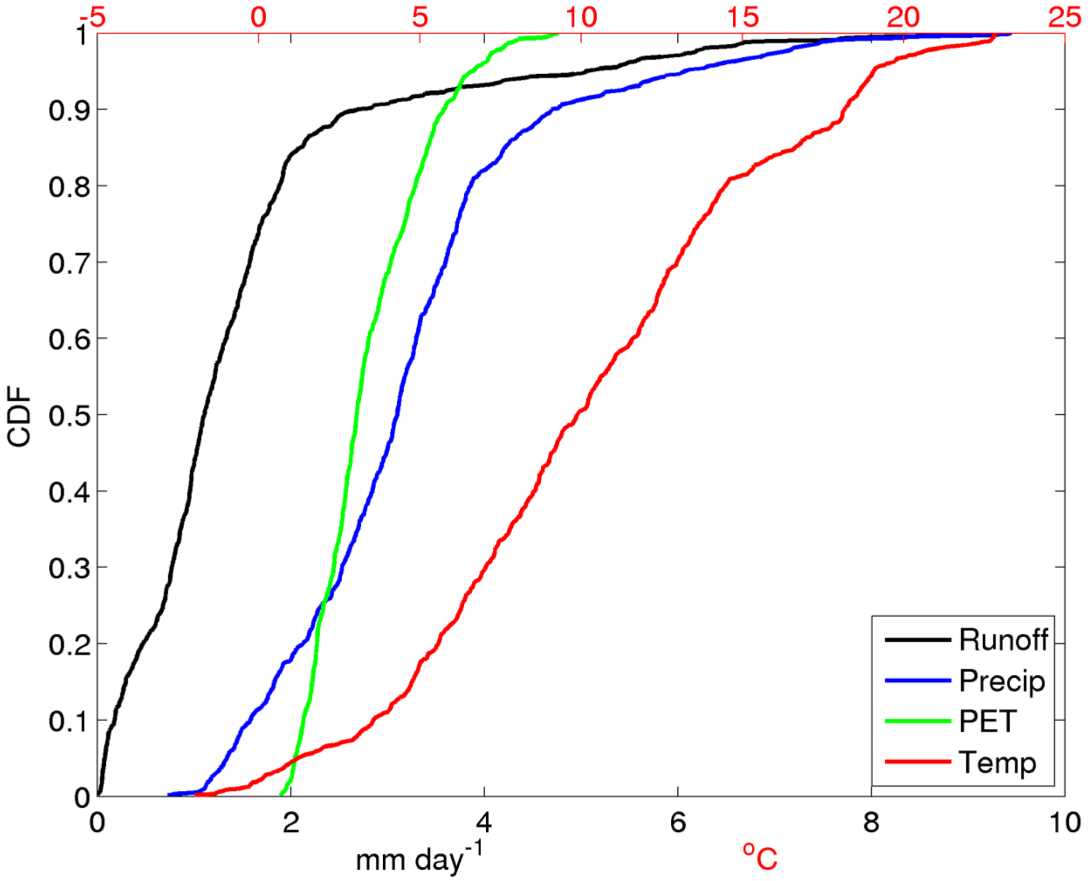


Figure 2. Annual cumulative density functions (CDFs) of runoff (mm day^{-1}) (black, bottom X-axis), precipitation (mm day^{-1}) (blue, bottom X-axis), potential evapotranspiration (mm day^{-1}) (green, bottom X-axis), and temperature (°C) (red, top X-axis).

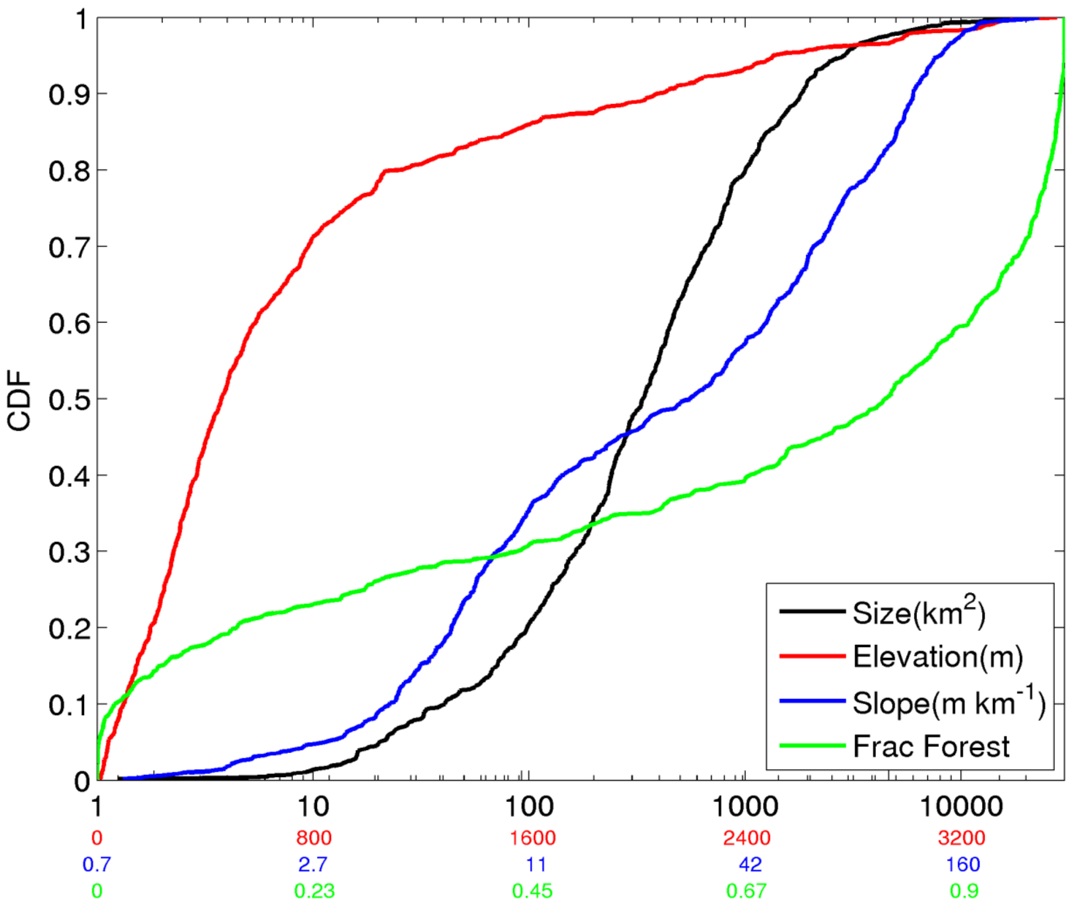


Figure 3. Cumulative density functions of basin size (km^2) (black), basin mean elevation (m) (red), mean slope (m km^{-1}) (blue), and fractional forest cover (green) for the basin set.

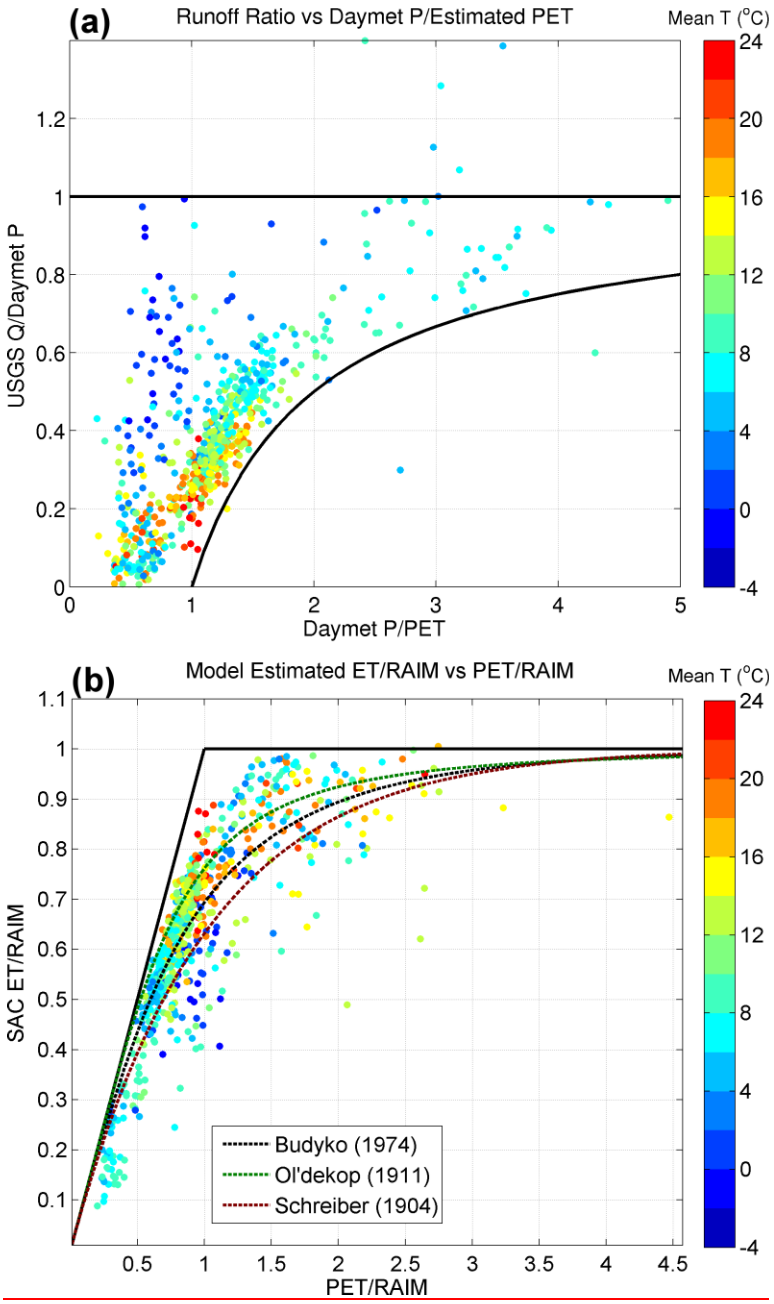


Figure 4. (a) Runoff ratio of observed runoff to Daymet estimated precipitation versus ratio of Daymet estimated precipitation to Priestly-Taylor estimated potential evapotranspiration (PET). (b) Model derived Budyko analysis using model evapotransipiration (ET), PET and total surface water input (rain plus melt, RAIM) for the 671 basins and three derivations of the Budyko curve (dashed lines). Basin mean temperature shaded (coloring) in both panels.

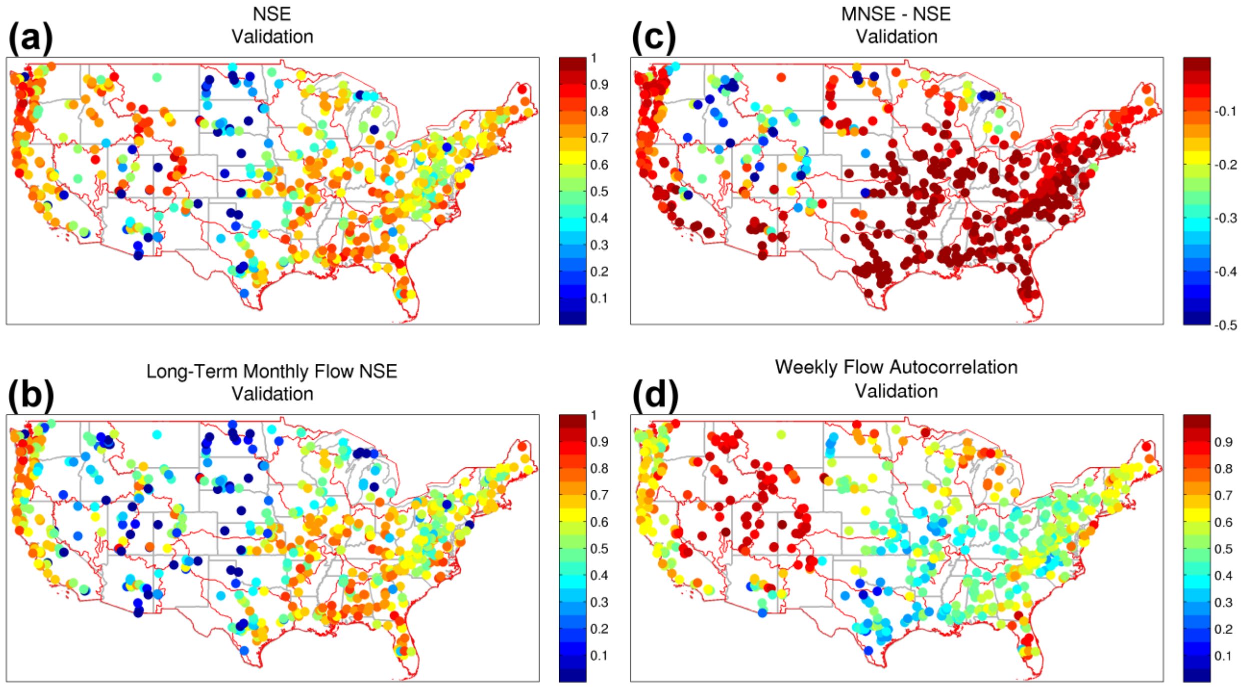


Figure 5. (a) Spatial distribution of Nash-Sutcliffe efficiency (NSE), (b) Nash-Sutcliffe efficiency using long-term monthly mean flows (MNSE) rather than the long-term mean flow, (c) MNSE – NSE for the validation period, (d) weekly flow autocorrelation.

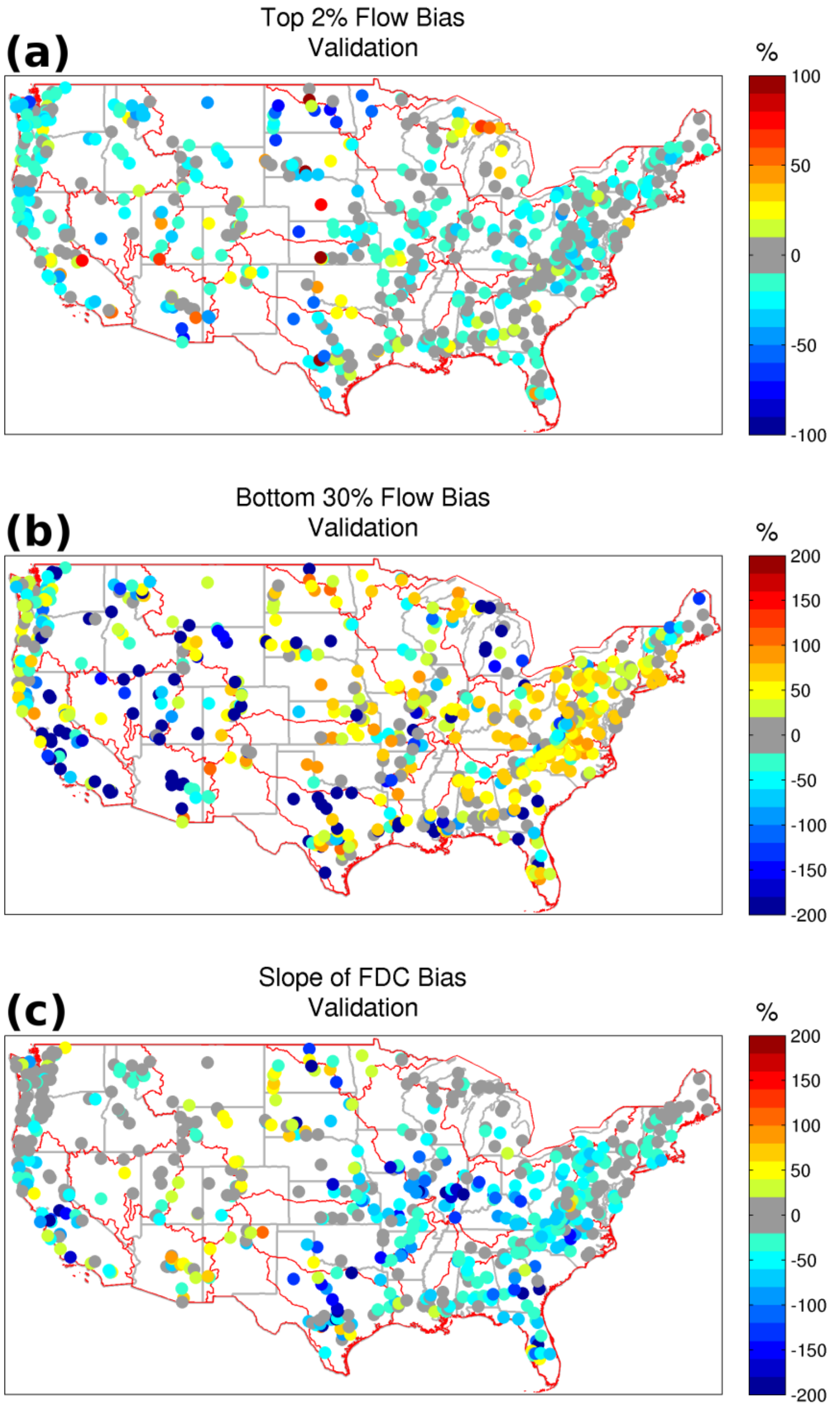


Figure 6. (a) Spatial distribution of the high flow bias, (b) low flow bias, (c) flow duration curve bias for the validation period.

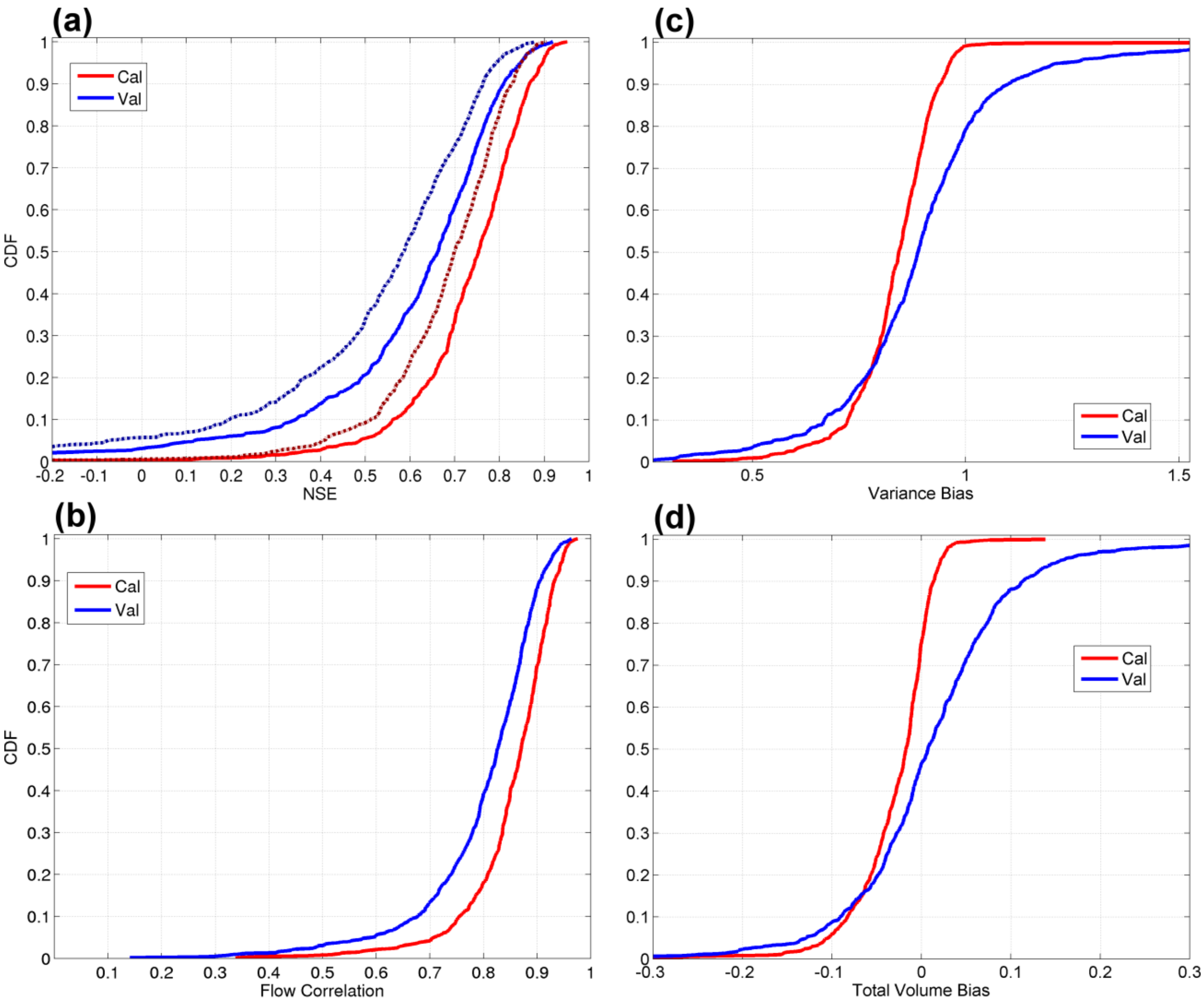
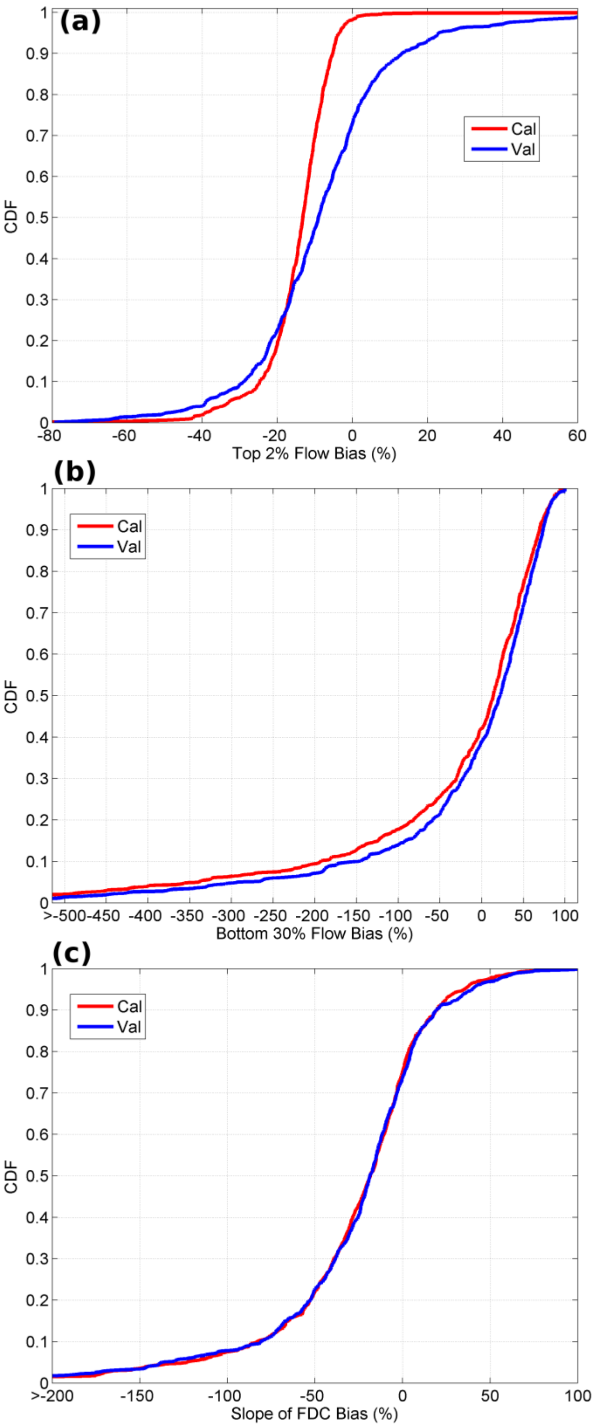


Figure 7. (a) Cumulative density functions (CDFs) for model Nash-Sutcliffe efficiency (NSE) (solid) for the calibration (red) and validation periods (blue) and NSE using the long-term monthly mean flows (MNSE, dark shaded and dashed), ~~(b) ~~CDFs for (b) simulated-observed flow correlation in the decomposition of the NSE, (c) for the variance bias in the decomposition of the NSE, ~~(c~~and (d) total volume bias in the decomposition of the NSE.


(a)
Cal
Val
CDF
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-80
-60
-40
-20
0
20
40
60
Top 2% Flow Bias (%)
(b)
Cal
Val
CDF
>-500
-450
-400
-350
-300
-250
-200
-150
-100
-50
0
50
100
Bottom 30% Flow Bias (%)
(c)
Cal
Val
CDF
>-200
-150
-100
-50
0
50
100
Slope of FDC Bias (%)

~~Figure 3.~~

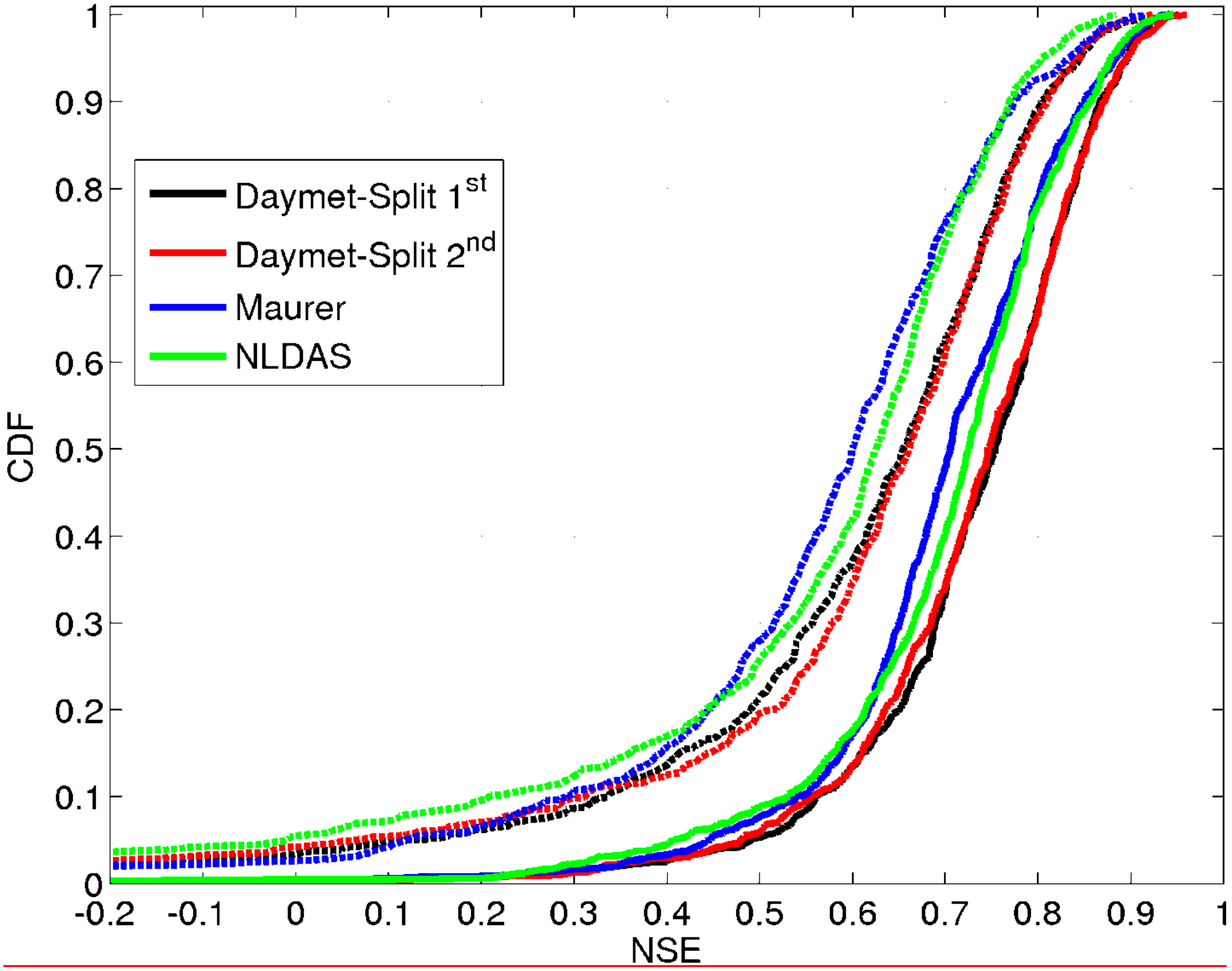


Figure 8. Cumulative density functions for model Nash-Sutcliffe efficiency for the calibration (solid) and validation (dashed) period using three different forcing datasets (Daymet, Maurer, NLDAS). The Daymet dataset was calibrated using the first 15 years (Split 1st) and validated against the remaining data and also calibrated using the last 15 years (Split 2nd) and validated against the initial streamflow data. Maurer and NLDAS calibrations performed using the first 15 years of observed streamflow only.

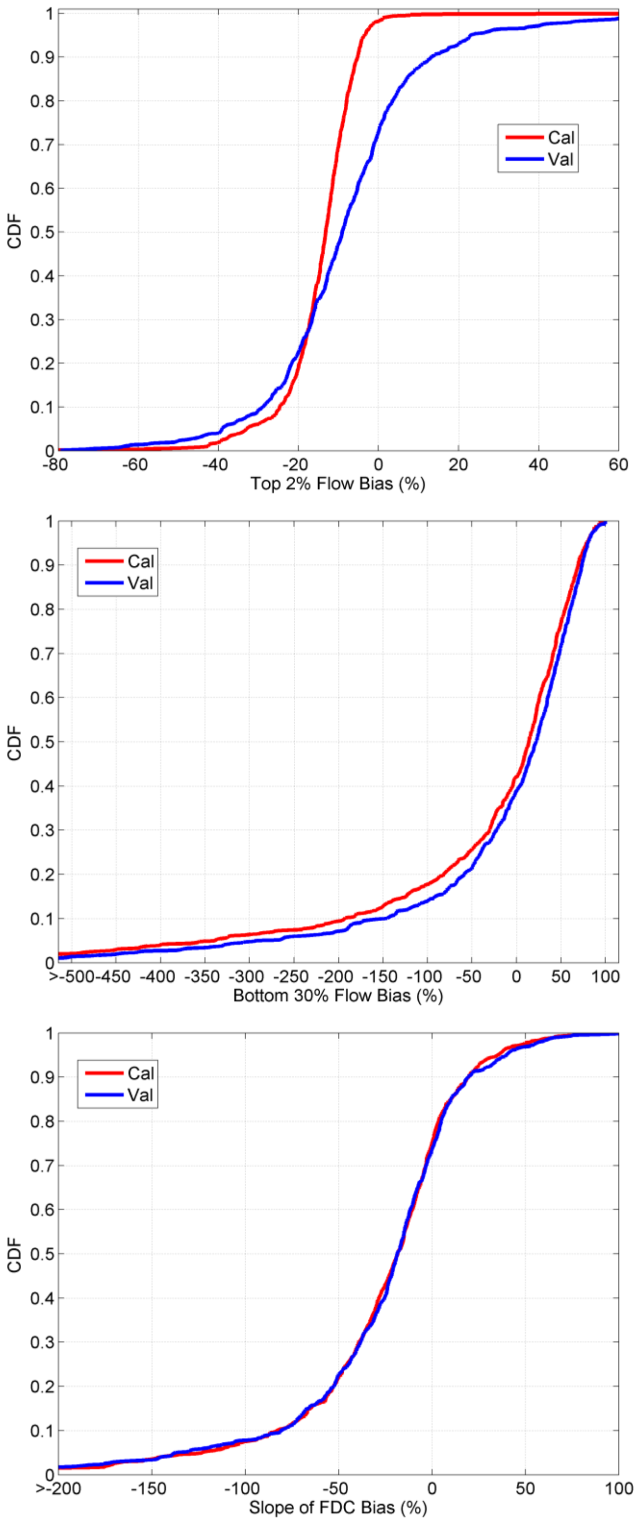
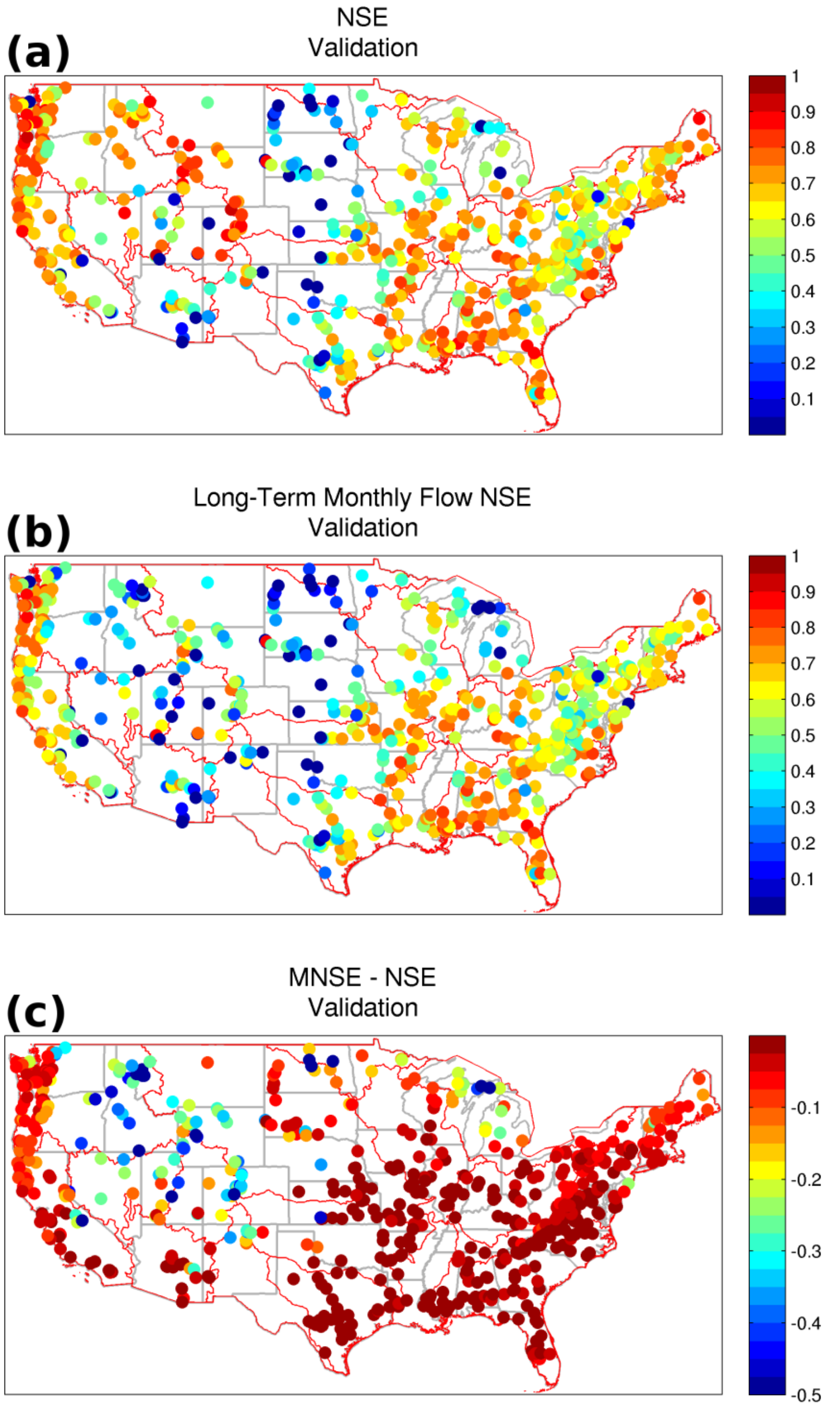
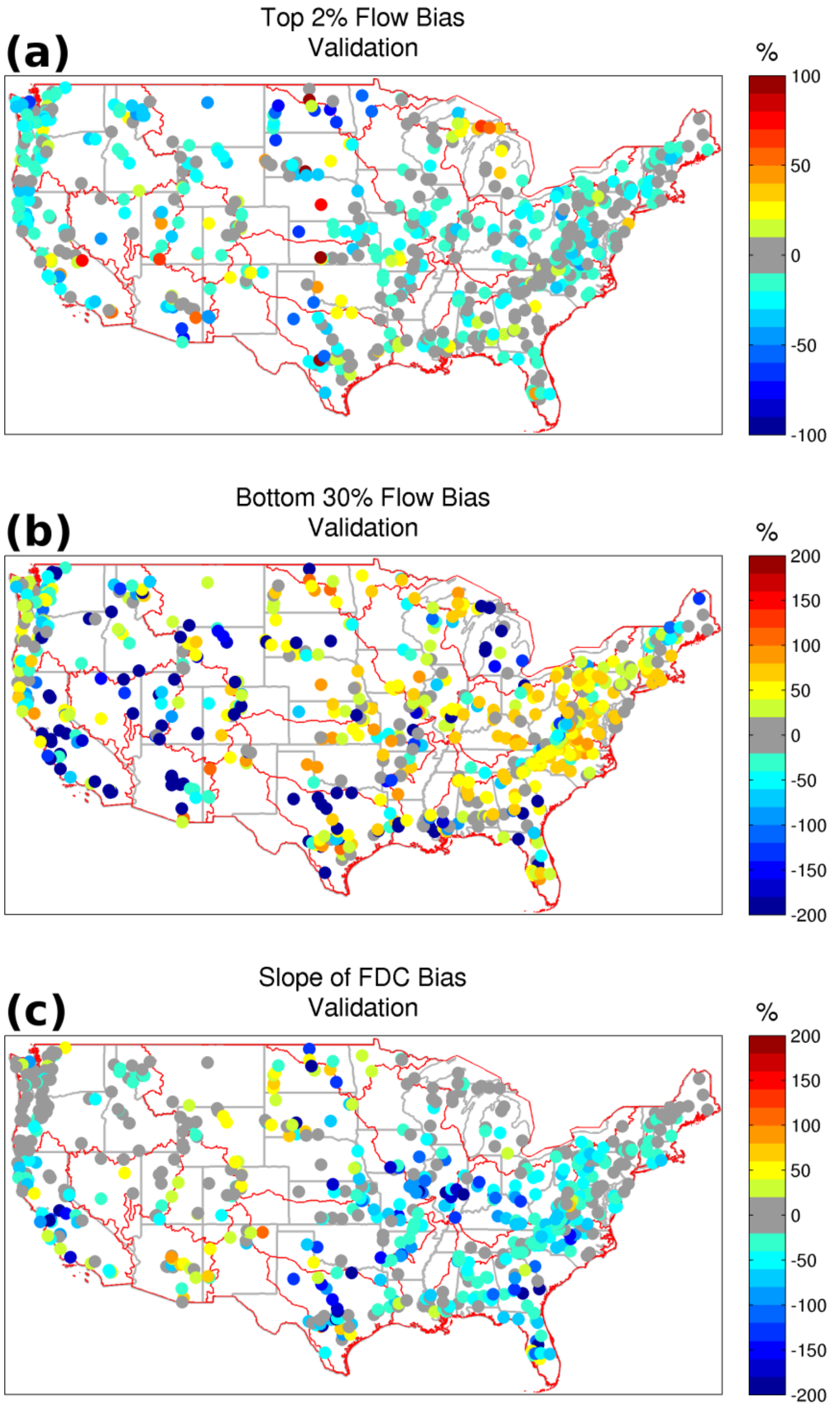


Figure 9. (a) Cumulative density functions (CDFs) for model high flow bias for the calibration (red) and validation periods (blue), (b) model low flow bias, (c) model flow duration curve slope bias.



~~Figure 4. (a) Spatial distribution of Nash-Sutcliffe efficiency (NSE), (b) Nash-Sutcliffe efficiency using long-term monthly mean flows (MNSE) rather than the long-term mean flow, (c) MNSE – NSE for the validation period.~~


(a)
Top 2% Flow Bias
Validation
%
100
50
0
-50
-100
(b)
Bottom 30% Flow Bias
Validation
%
200
150
100
50
0
-50
-100
-150
-200
(c)
Slope of FDC Bias
Validation
%
200
150
100
50
0
-50
-100
-150
-200

~~Figure 5.~~

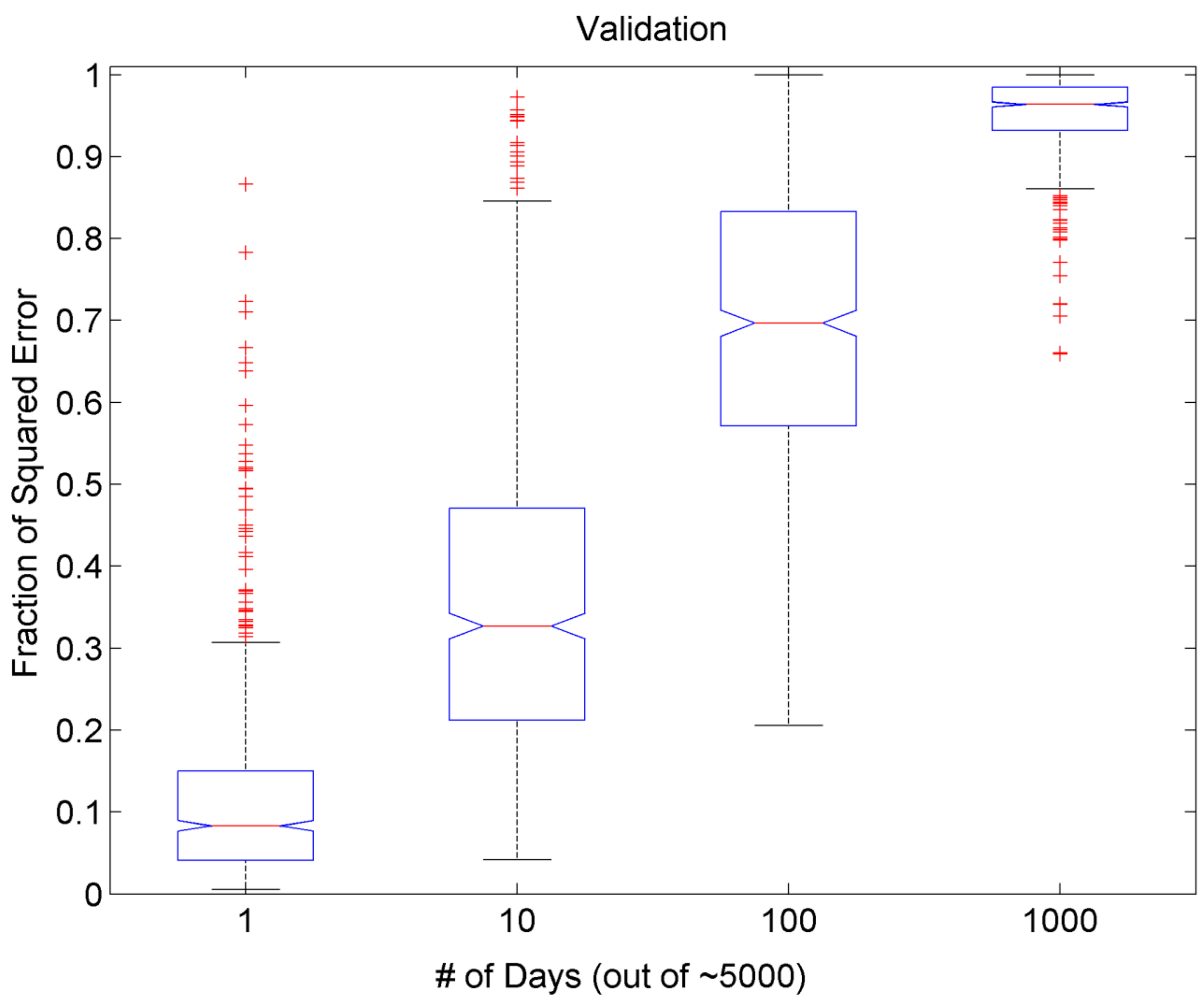
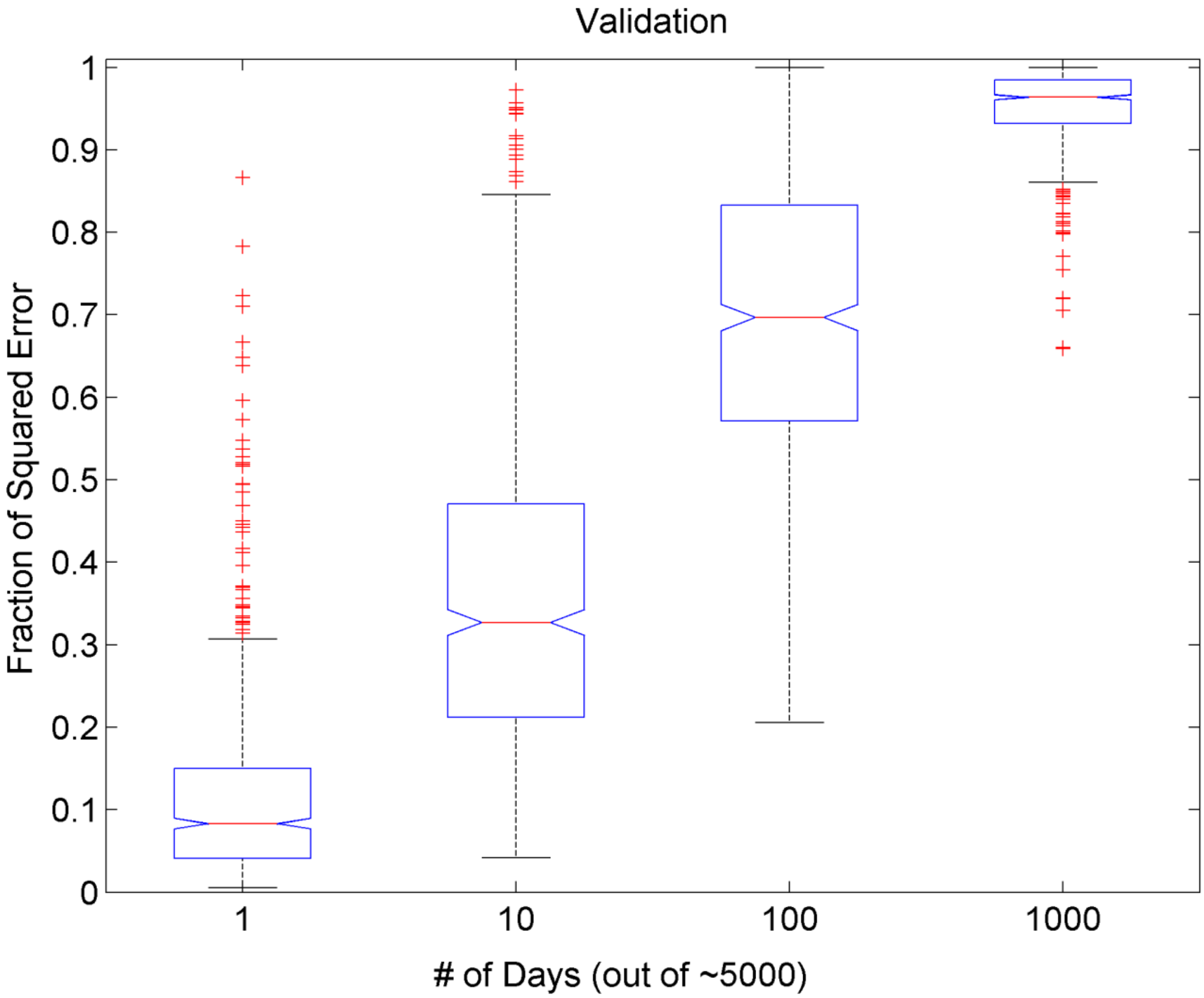
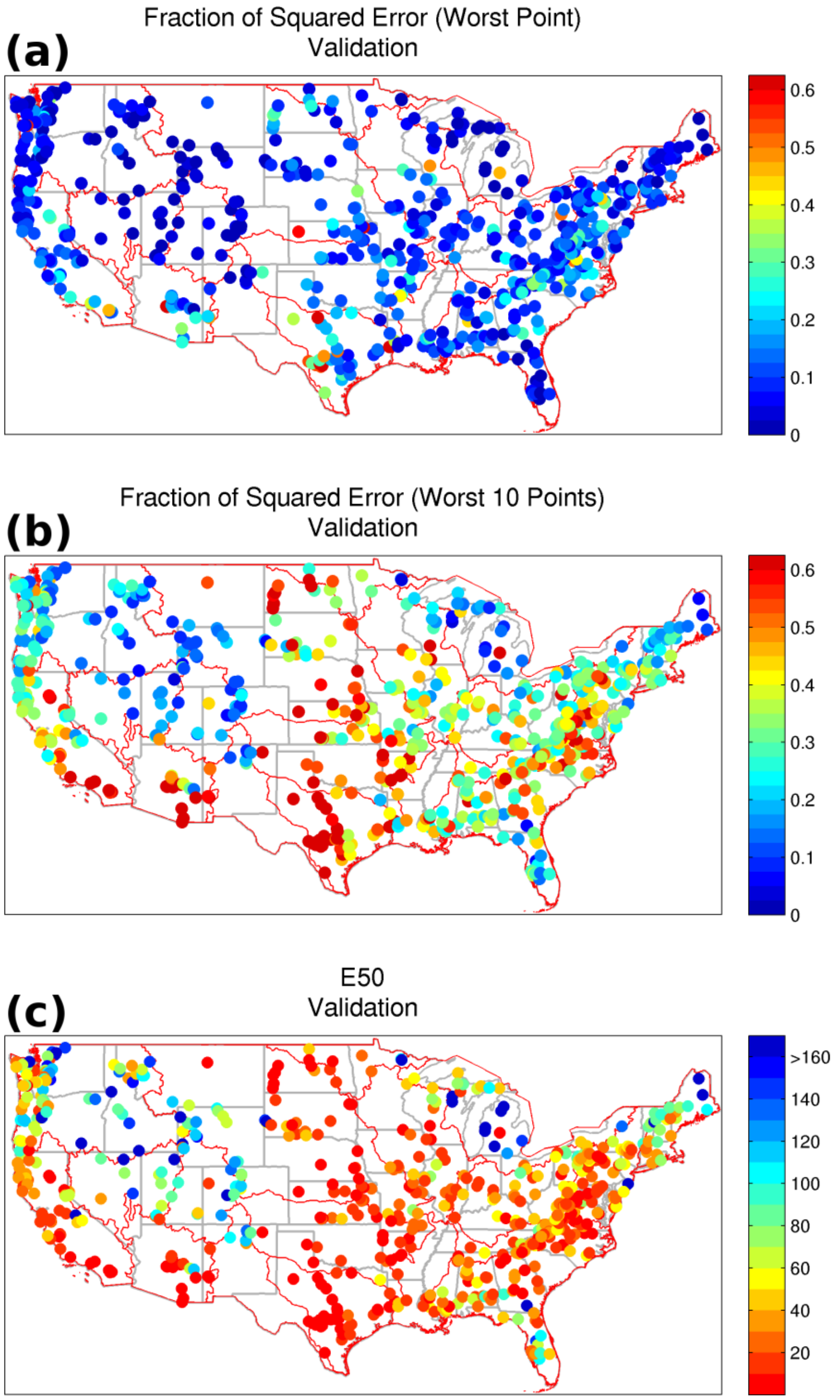


Figure 10.~~(a) Spatial distribution of the high flow bias, (b) low flow bias, (c) flow duration curve bias for the validation period.~~



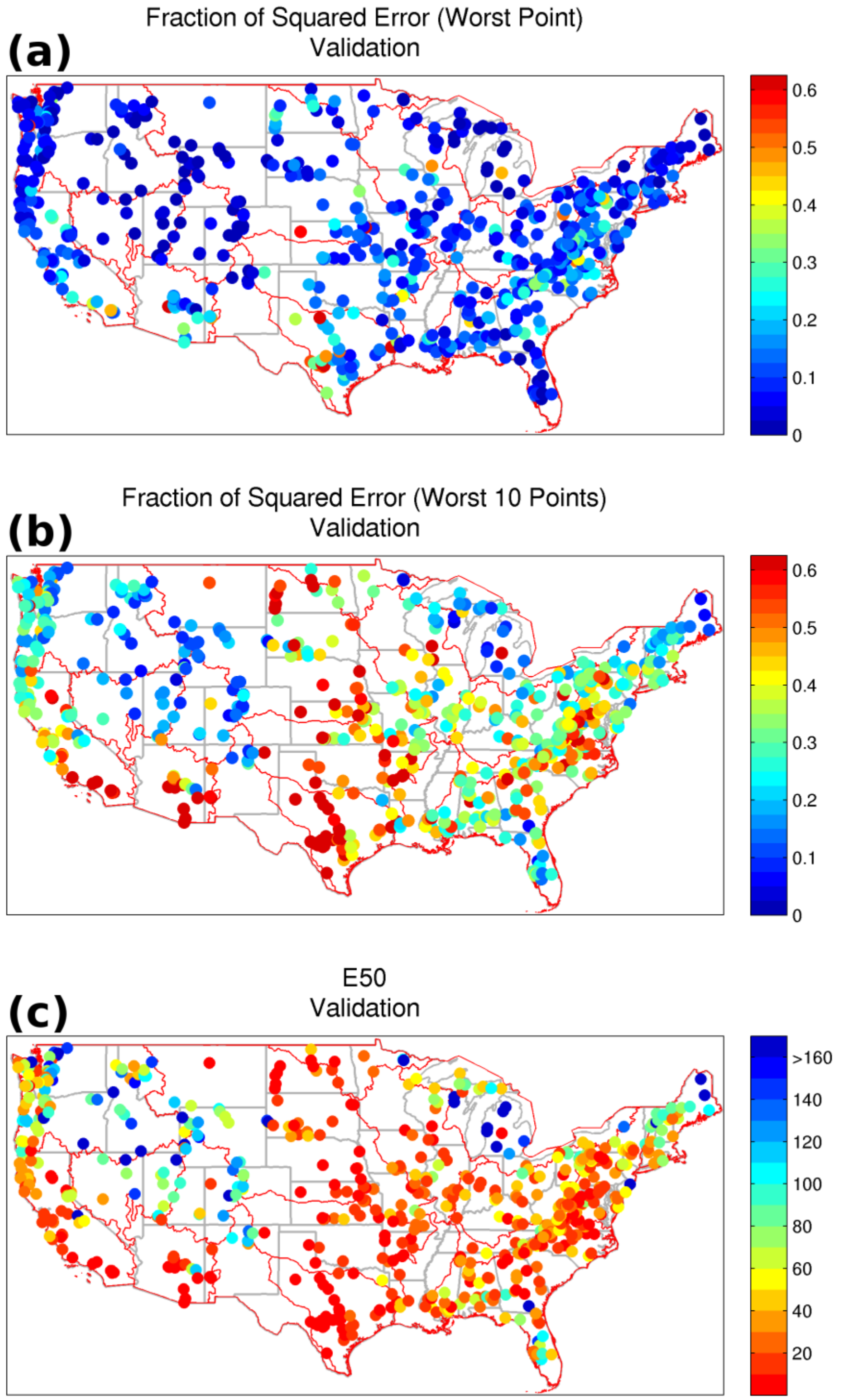
~~Figure 6.~~ Fractional contribution of the total squared error for the 1, 10, 100, 1000 largest error days. The box plots represent the 671 basins with the blue area defining the interquartile range, the whiskers representing reasonable values and the red crosses denoting outliers. The median is given by the red horizontal line with the notch in the box denoting the 95 % confidence interval of the median value.



Fraction of Squared Error (Worst Point)
Validation

Fraction of Squared Error (Worst 10 Points)
Validation

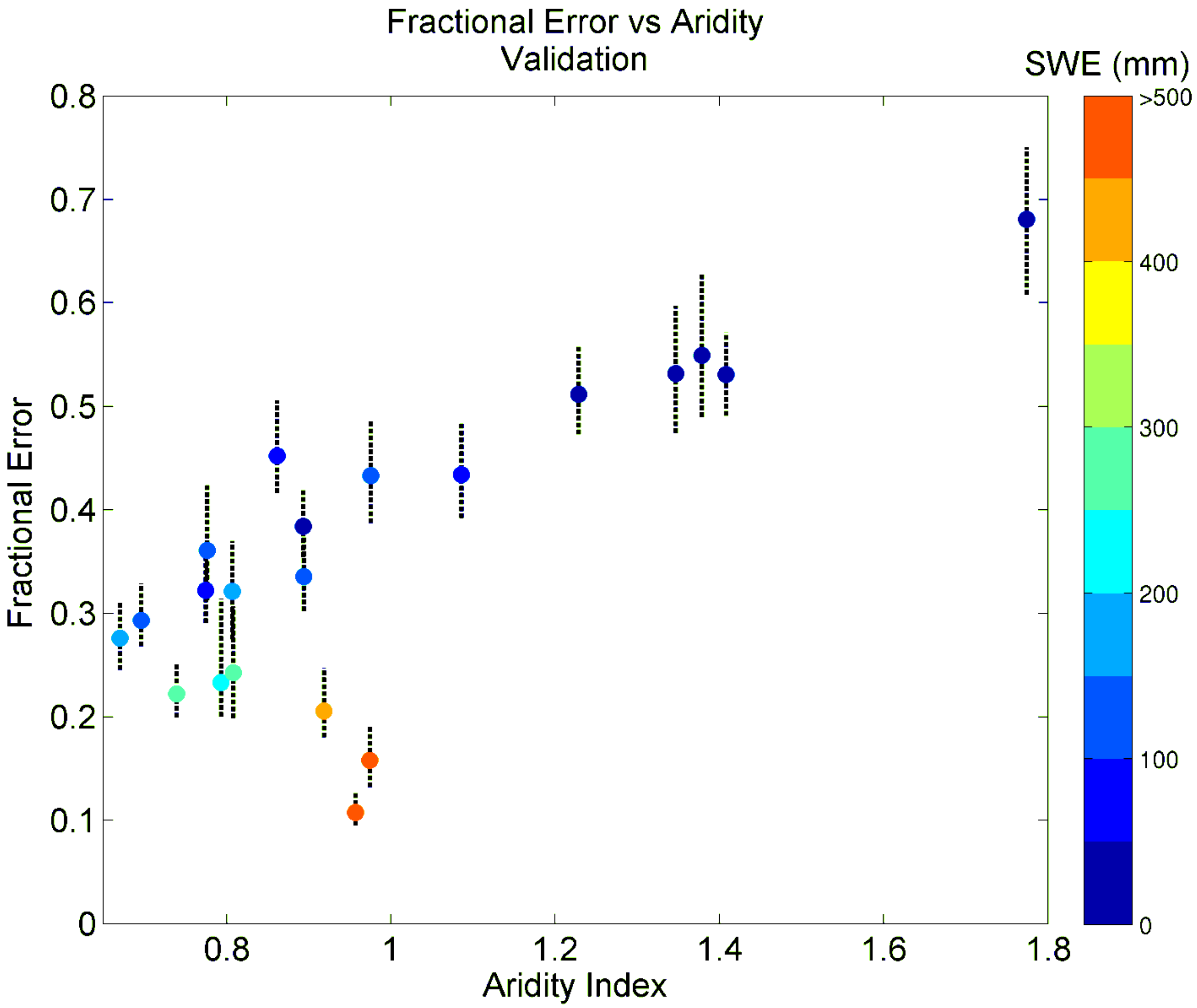
E50
Validation


(a) Fraction of Squared Error (Worst Point) Validation

(b) Fraction of Squared Error (Worst 10 Points) Validation

(c) E50 Validation

Figure ~~7~~11. (a) Spatial distribution of the fractional contribution of total squared error for the largest day during the validation period, (b) 10 largest error days, (c) the number of days contributing 50% of the total objective function error, E50.


Fractional Error vs Aridity
Validation
SWE (mm)
>500
400
300
200
100
0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Fractional Error
0.8
1
1.2
1.4
1.6
1.8
Aridity Index

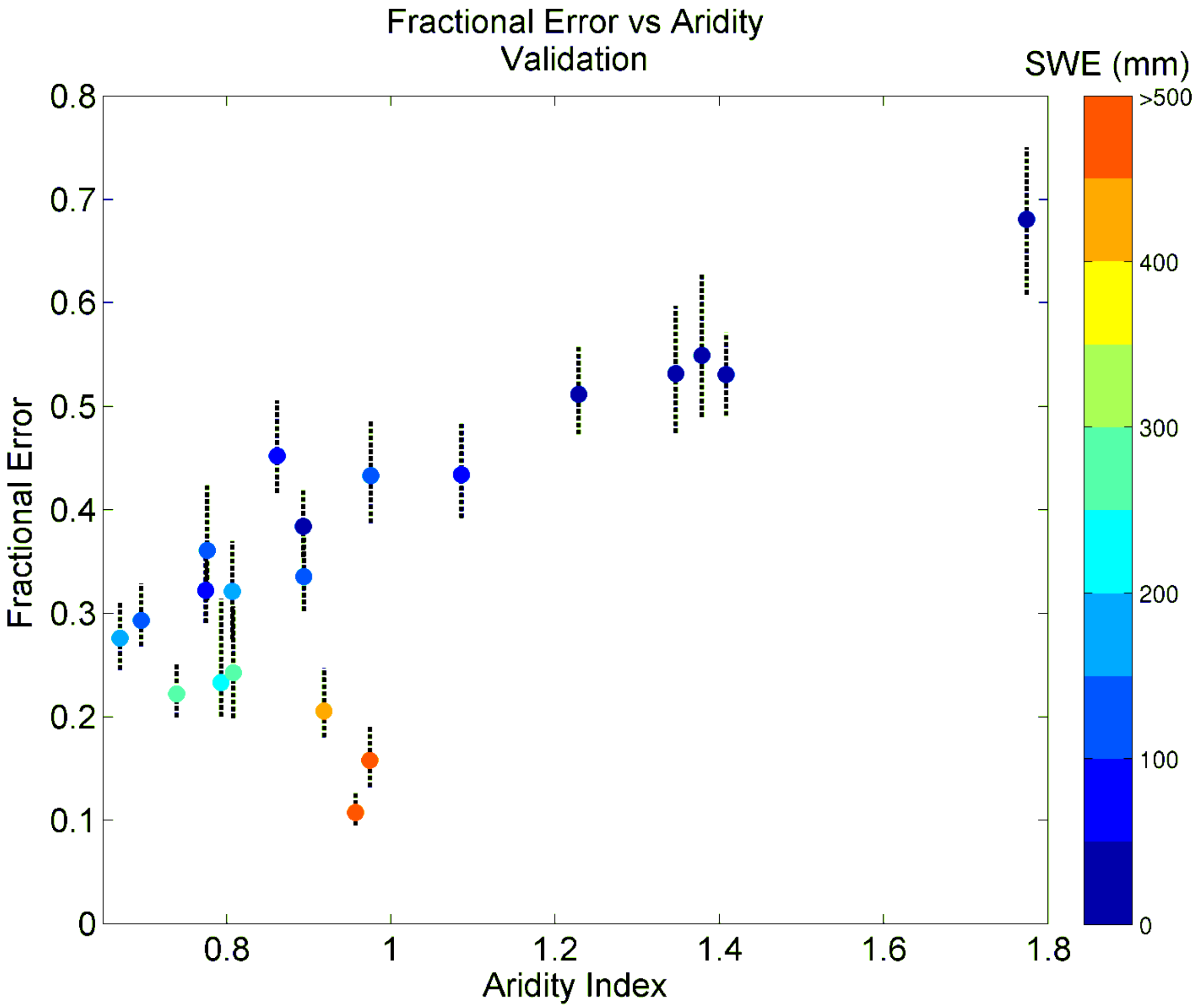
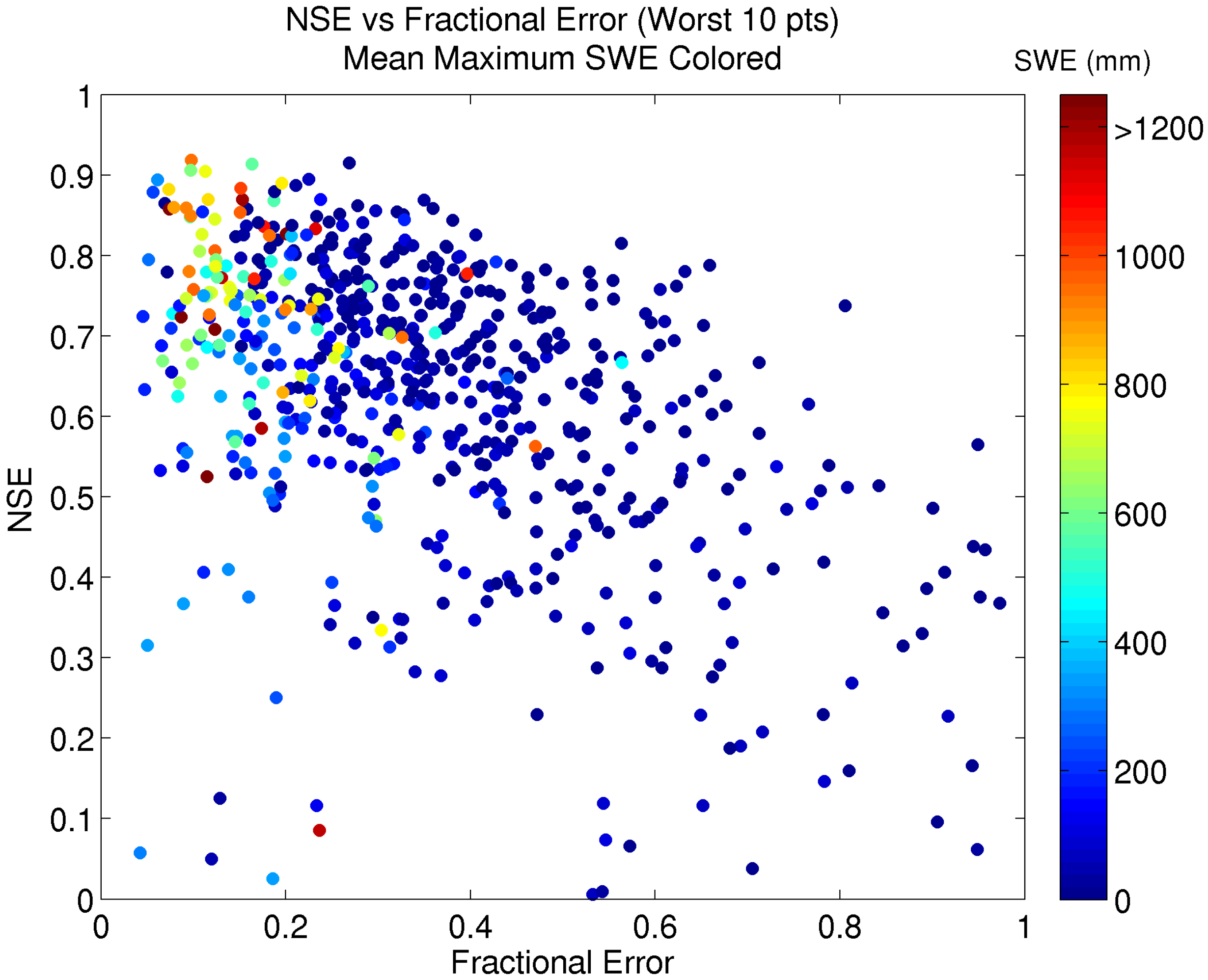


Figure ~~8~~12. Ranked fractional squared error contribution for the 100 largest error days for the 671 basins versus the aridity index with mean maximum snow water equivalent (SWE) shaded. Each dot represents a ~32 basin bin defined by the rank of the fractional error contribution for the 100-largest error days for all basins. The dashed vertical black lines denote the 95% confidence interval for the mean of the fractional error contribution for a given bin.


NSE vs Fractional Error (Worst 10 pts)
Mean Maximum SWE Colored
SWE (mm)
NSE
Fractional Error
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
>1200
1000
800
600
400
200
0

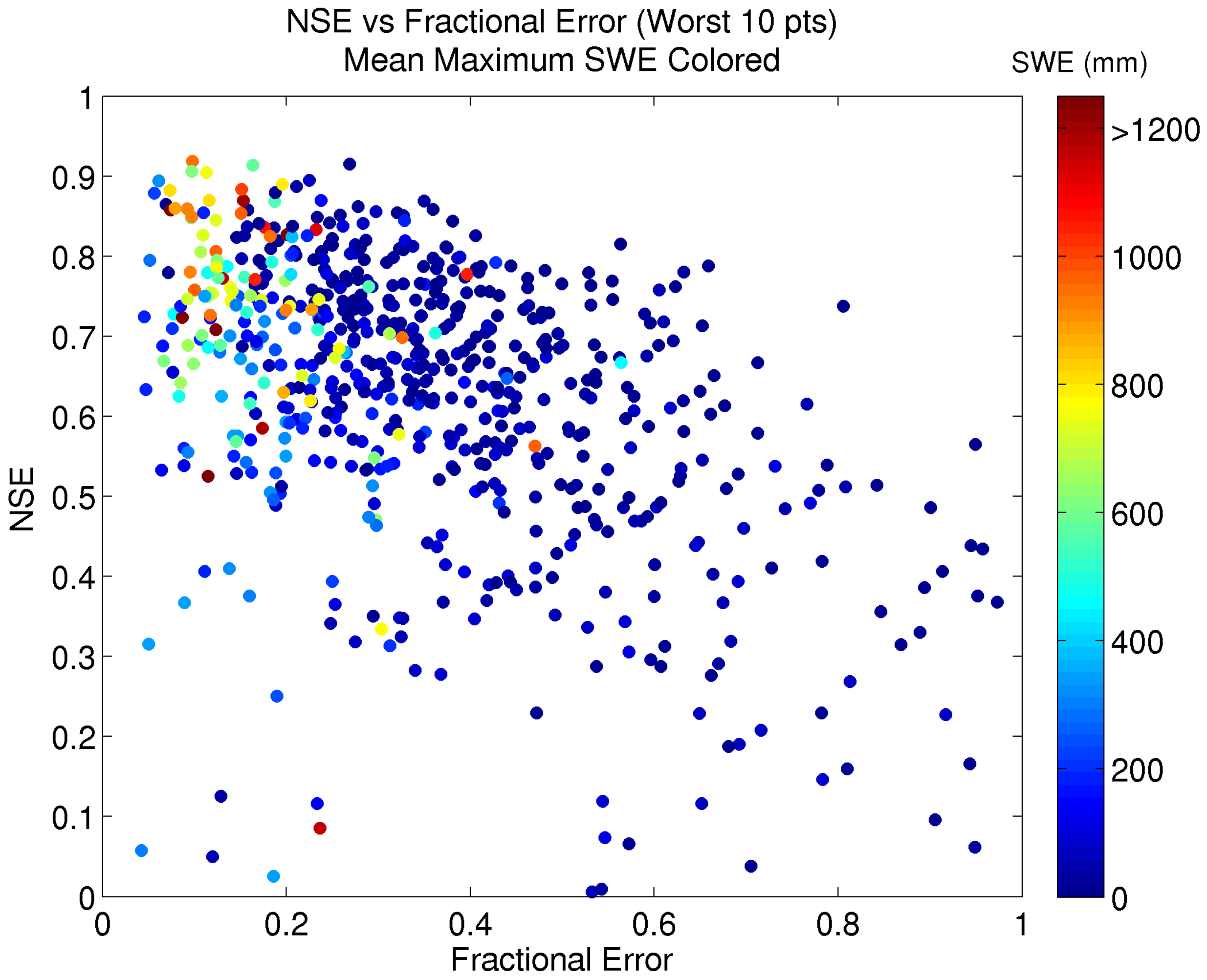


Figure ~~9~~13. Nash-Sutcliffe efficiency versus the fractional error of the 10 largest error days for the validation period for all basins with basin mean peak snow water equivalent (mm) colored.