Author/s:
Zheng, F;Maier, HR;Wu, W;Dandy, GC;Gupta, HV;Zhang, T

Title:
On Lack of Robustness in Hydrological Model Development Due to Absence of Guidelines for Selecting Calibration and Evaluation Data: Demonstration for Data-Driven Models

Date:
2018-02-01

Citation:
Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V. & Zhang, T. (2018). On Lack of Robustness in Hydrological Model Development Due to Absence of Guidelines for Selecting Calibration and Evaluation Data: Demonstration for Data-Driven Models. Water Resources Research, 54 (2), pp.1013-1030. https://doi.org/10.1002/2017WR021470.

On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models

Feifei Zheng, Holger R Maier, Wenyan Wu, Graeme C Dandy, Hoshin V Gupta, and Tuqiao Zhang

Feifei Zheng: Corresponding author, Professor, College of Civil Engineering and Architecture, Zhejiang University, China. feifeizheng@zju.edu.cn. Tel: +86-571-8820-6757. **Mail address:** A501, Anzhong Building, ZijinGang Campus of Zhejiang University, 866 Yuhangtang Road, Hangzhou, Zhejiang 310058, China.

Holger R Maier: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia. holger.maier@adelaide.edu.au; Adjunct professor, College of Civil Engineering and Architecture, Zhejiang University, China.

Wenyan Wu: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia. wenyan.wu@adelaide.edu.au

Graeme C. Dandy: School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide, South Australia, 5005, Australia. graeme.dandy@adelaide.edu.au.

Hoshin V. Gupta: Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ 85705, USA. hoshin@email.arizona.edu

Tuqiao Zhang: Professor, College of Civil Engineering and Architecture, Zhejiang University, China. ztq@zju.edu.cn.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of record. Please cite this article as doi:10.1002/2017WR021470.

Abstract: Hydrological models are used for a wide variety of engineering purposes, including streamflow forecasting and flood-risk estimation. To develop such models, it is common to allocate the available data to calibration and evaluation data subsets. Surprisingly, the issue of how this allocation can affect model evaluation performance has been largely ignored in the research literature. This paper discusses the evaluation performance bias that can arise from how available data are allocated to calibration and evaluation subsets. As a first step to assessing this issue in a statistically rigorous fashion, we present a comprehensive investigation of the influence of data allocation on the development of data-driven artificial neural network (ANN) models of streamflow. Four well-known formal data splitting methods are applied to 754 catchments from Australia and the US to develop 902,483 ANN models. Results clearly show that the choice of the method used for data allocation has a significant impact on model performance, particularly for runoff data that are more highly skewed, highlighting the importance of considering the impact of data splitting when developing hydrological models. The statistical behaviour of the data splitting methods investigated is discussed and guidance is offered on the selection of the most appropriate data splitting methods to achieve representative evaluation performance for streamflow data with different statistical properties. Although our results are obtained for data-driven models, they highlight the fact that this issue is likely to have a significant impact on all types of hydrological models, especially conceptual rainfall runoff models.

Keywords: model evaluation bias, hydrological models, calibration and evaluation, artificial neural networks (ANN), data allocation, data splitting.

1. Introduction

[1] Traditionally, hydrological models have been classified as either black box, physical or conceptual models [*Beven, 2006; Li et al., 2015*]. More recently, *Mount et al.* [*2016*] have argued that it would be beneficial to think of different model types in terms of a continuum from strongly hypothesis-based to almost entirely data-based, rather than belonging to distinct categories. For example, in data-driven artificial neural network (ANN) models, the degree of data influence is high, with the available data being used to infer both model structure and values for the model parameters, while the role of scientific hypotheses is confined mainly to the selection of candidate model inputs. In contrast, fully coupled surface water – groundwater models tend to be strongly hypothesis-based, with model structures established based on a combination of assumptions and physical knowledge regarding the nature of the hydrological processes to be represented, while data are used mainly for estimating values for the model parameters. However, even in these models, the process parameterization equations employed are often empirical generalizations based on the analysis of observational data [*Mendoza et al., 2015*]. A variety of model types fall between these extremes. For example, in classical conceptual rainfall runoff (CRR) models, the structural representations serve as hypotheses regarding how key processes can be represented in a conceptual manner, and available data are used primarily for parameter estimation. In the recent class of CRR models designed to provide flexibility in structure selection [*Fenicia et al., 2008, 2011; Clark et al., 2015*], data are used both to evaluate alternative model structures (from a set of pre-defined alternatives) and to estimate parameter values.

[2] Regardless of which model type is used, it is generally considered good practice to conduct an independent assessment of the performance of a model, using data that were not used for model development [e.g. P*ower, 1993; Biondi et al., 2012; Humphrey et al., 2017*]. This practice is also commonly used to perform comparative evaluations of the performance of different types of models (i.e. performance on the independent evaluation data is used to infer whether the performance of a particular model can be considered to be superior to that of another) [e.g. *Dibike and Coulibaly, 2005; Valipour et al., 2013; Li et al., 2015; Humphrey et al., 2016*].

[3] Evaluation of model performance on an independent data set requires the available data to be split into model calibration and evaluation subsets [*Klemes, 1986*]. Consequently, the results of both model calibration and evaluation depend on which subset of the available data is used for the former and which is used for the latter. It follows that the method used to decide which data subset is used for calibration and evaluation can have a significant impact on the results – influencing both the nature of the model obtained and also the conclusions regarding model adequacy/performance arrived at through the out-of-sample assessment. For example, if a significant number of 'extreme' events is included in the calibration data, the model development process will be influenced by the information that is contained in those data about system behaviour under such conditions. However, if such events are not included in the evaluation data, the evaluation process will be unable to provide an independent assessment of how well the model is likely to perform under similar conditions in the future. In addition, given that models generally tend to perform worse under more extreme conditions than under normal conditions, the metric used to quantify simulation error during the evaluation period is likely to indicate better model performance than if the evaluation data contained more extreme events, thereby providing an *optimistic* assessment of the magnitude of the errors that can be expected once the model is deployed in practice. Conversely, if extreme events are not included in the calibration

data, the evaluation period errors are likely to provide a *pessimistic* assessment of the magnitude of the errors that could be expected in practice.

[4] The impact of which subset of the available data is used for calibration and evaluation is likely to be dependent on the modelling approach used. Take, for example, CRR models that have pre-specified structures. Model calibration, in this case, will cater to any model structural inadequacies by producing a particular (conditional) set of parameter estimates, and these will then result in correspondingly biased simulations (that are particular to the specific nature of the structural inadequacy) during both the calibration and evaluation periods. During the calibration period the metrics used will typically act to minimize such biases "on average"; but of course, this does not guarantee that such biases will remain minimized for out of sample periods [*Gupta et al., 2009*]. Further, state errors that accumulate during normal periods will also influence performance during extreme events. One should also remember that the model parameterization equations *may not actually contain* information about system behaviour under extreme conditions, since the theory may not be robust under such conditions.

[5] Conversely, with flexible data-based modelling approaches, such as ANNs, the development of the model equations is informed by the nature of the structure in the data. In this case, the main problem is that there may or may not be sufficient information in the data regarding how the system will behave under out of sample conditions. And, since the more recent flexible-structure type CRR models are an attempt to take advantage of the strengths of both the hypothesis-driven and data-based approaches, one might expect such models to suffer (to some degree) from both the structural bias and data information content issues [*Beven and Smith, 2015*].

[6] Intuitively, one would expect the ability of a model to extrapolate outside the range of the data used for calibration to increase with the degree of physical system understanding (scientific hypotheses) built into the model, thereby resulting in improved out-of-sample performance. However, our informal survey of many hydrologists and search of the literature reveals that the validity of this commonly held assumption has not actually been subject to rigorous testing thus far. Meanwhile, a number of studies using CRR models have shown that, while such models have parameterization equations and parameters that are conceptually associated with a variety of physical processes, the degree to which calibrated parameters correspond to these processes is questionable [*e.g. Sorooshian and Gupta, 1983; Troutman, 1985; Beven, 1993; Gan and Biftu, 1996; Ferket et al., 2010; Shin et al., 2013; Li et al., 2015*]. In addition, several studies have highlighted that the evaluation performance of CRR models is worse if the types of events that are present in the calibration and evaluation data are different [e.g. *Hartmann and Bardossy, 2005; Chiew et al., 2009; Coron et al., 2012; Thiel et al., 2015a, b; Fowler et al., 2016*] and that model performance improves if model parameters are updated over time in non-stationary environments [e.g. *Merz et al., 2011; Bowden et al. 2012; Brigode et al., 2013; de Vos et al., 2010; Luo et al., 2012; Vaze et al., 2010; Gibbs et al., 2017]*. This further challenges the perception that the extrapolation ability of more physically based models is superior to that of more data-driven models.

[7] Overall, it seems clear that the issue of how the available data should be used in the model calibration and evaluation process is important regardless of the type of hydrological model being considered. It is therefore surprising that there have been such few studies that have investigated this issue in a systematic fashion [e.g. *May et al., 2010; Wu et al., 2013; Fowler et al., 2016*] and that this issue is generally ignored in the vast majority of papers on hydrological

modelling. There are even fewer studies that have attempted to develop approaches for addressing this issue. For models for which the time structure of the modelling data has to be maintained, the use of multiple calibration and evaluation periods has been suggested [e.g. *Ebtehaj et al.,2010*; *Coron et al., 2012, 2014; McInerney et al., 2017*]. In contrast, for models for which the time structure of the modelling data does not have to be maintained, the use of different sampling approaches has been suggested [e.g. *Bowden et al., 2002; May et al., 2010; Wu et al., 2013*]. However, there has been a lack of systematic assessment of the utility of these approaches over a large number of catchments with different properties.

[8] In this paper, we apply a systematic approach to quantifying the potential impact of different data allocation approaches on the robustness of the hydrologic model development process. As a first step, we consider hydrological models that do not require continuous inputs, as this enables alterative calibration and evaluation subsets to be constructed with the aid of sampling techniques, making it easier to test the impact of different data allocation approaches in a statistically rigorous manner. Specifically, we develop ANN runoff models using different data allocation approaches for a large sample of catchments having a wide variety of runoff skewness values, corresponding to different levels to which similar events can be included in both the calibration and evaluation subsets. Based on these large sample results, we propose general guidelines for the application of different data allocation methods to ANN runoff model development. Although these guidelines are specific to ANN models, the data allocation methods used are applicable to all hydrological models for which continuous inputs are not required (e.g. data-driven models, event-based models). More broadly, the underlying issue of modelling bias associated with data allocation is likely to be applicable to all types of hydrological models (including those that are conceptual- or physics-based) and therefore deserving of further attention.

[9] Specific objectives of this paper are:

1. To determine how the robustness of model evaluation performance is affected by the data allocation method used.
2. To investigate how the performance of different data allocation methods is related to the skewness of the output data used for model development.
3. To develop guidance regarding the selection of data allocation methods to maximize robustness in model development.

The paper is organized as follows. The methodology is outlined in Section 2, followed by the presentation and discussion of the results in Section 3. A summary and conclusions are presented in Section 4.

2. Methodology

[10] To test the impact of different methods for allocating the available data to calibration and evaluation subsets, we use the benchmarking approach of *Wu et al.* [2013], outlined in Section 2.1. Details of the experimental procedure used are given in Section 2.2.

2.1 Benchmarking approach

[11] The benchmarking approach of *Wu et al.* [*2013*] is based on the premise that the allocation of any particular fraction of the available dataset to each of the two subsets results in an expected model error, which is the average model error over all possible splits of the data for the given fractional allocation (e.g. 80% for calibration, 20% for evaluation). So, if the model evaluation error is smaller than this benchmark error, the assessment of model performance can be

considered *optimistic*, because the error over the full range of data in the available data set is expected to be larger. Conversely, if the model evaluation error is larger than this benchmark error, the assessment of model performance can be considered *pessimistic*, because the error over the full range of data in the available data set is expected to be lower.

[12] If a large fraction of extreme events is included in the calibration data, and therefore excluded from the evaluation data, we can (in general) expect that our assessment of evaluation performance will be biased. Because the resulting model has been calibrated over a wide range of small, medium and large events, it is likely that it will perform well in practice, but our expectation regarding its ability to perform well on other periods (e.g., the future) will tend to be optimistic; i.e., the expected error will be smaller than if the model had been evaluated over the full range of events in the available data. Another way to think of this is that the selected data splitting strategy does not facilitate a robust evaluation of model performance, because it does not allow for model performance to be properly (rigorously) tested on the more extreme types of events. Conversely, if a smaller fraction of extremes is included in the calibration data and a larger fraction included in the evaluation data, a pessimistic assessment of expected model performance is likely to occur. Because the resulting model has not been calibrated on the full range of events (larger events have been excluded) it is likely that it will not perform quite so well in practice. Nonetheless, our expectation regarding its ability to perform well on other periods (e.g., the future) will tend to be pessimistic because the expected error will be larger than if extreme events had been included in the calibration data.

[13] Accordingly, the ideal situation will be when the calibration and evaluation data subsets both contain a range of independent events of similar magnitudes, so that the model development *and* evaluation processes can be performed on the full range of events represented in the available data. As discussed above, when this is not the case, our assessment of expected model performance on the evaluation data is likely to be misleading. In practice, however, this ideal situation is difficult to attain due to the relatively low frequency with which extreme events occur in hydrological datasets. For runoff data, in particular, this situation arises relatively often because distributions of runoff are typically highly skewed [e.g., *Wu et al., 2013*], reflecting the rarity of extreme events, which makes it difficult to include representative events in both subsets of a data split. For this reason, it is important to quantify the effect the degree of skewness in the data has on the strategy used to allocate the available data to calibration and evaluation subsets.

[14] To obtain the *true* value of a benchmark error for a given data set would require evaluation of the performance of models developed for every possible data split. This is generally computationally infeasible within practical timeframes due to the large number of possible data split combinations. However, as shown by *Wu et al.* [*2013*], an estimate having a specified level of accuracy and confidence can be obtained by using a sufficiently large number of random samples, where the required number is a function of the statistical properties of the available data. In this context, a "sample" corresponds to the result of a specific allocation of the available data to calibration and evaluation subsets (i.e. a specific data split), the calibration of the corresponding model, and the calculation of the desired evaluation performance metric. The results are averaged over all samples to determine the benchmark error. It should be noted that the benchmark method is based on the assumption that the hydrological model does not require continuous data for calibration and evaluation, as is the case for most data-driven models and event-based models. For models that require continuous data, the number of potential splits can

be significantly lower, but the benchmark method can still be used to ensure that representative data are included in both the calibration and evaluation subsets.

2.2 Experimental procedure

[15] Our overall experimental procedure is illustrated in Figure 1. To obtain catchment data with a wide range of runoff skewness properties, thereby enabling a rigorous assessment of the impact of different data allocation methods, data from 754 catchments in Australia and the USA are used. For each catchment, ANN runoff models are developed following *Maier et al. [2010]* and *Wu et al., [2014]*, including input selection, data splitting, model structure selection, model calibration and model evaluation. Relevant inputs are selected based on an assessment of partial mutual information. To assess the impact of different data splitting approaches, SS, DUPLEX, SBSS with Neyman sampling (SBSS-N), and SBSS with proportional sampling (SBSS-P) are used [*Wu et al. 2013*]. To account for the stochastic nature of different sampling methods, 100 independent trials of each data splitting method are implemented, producing 100 data splits. ANN models are separately developed, calibrated and evaluated for each of the 301,600 data split combinations.

[16] To investigate how the robustness of model evaluation performance is affected by the data allocation method used (Objective 1), relative bias and variance metrics (over the 100 data splits) are computed for each of the four data splitting methods for each of the 754 catchments. These metrics are computed based on differences between ANN evaluation errors and the performance benchmarks (Section 2.1). Determination of the performance benchmarks requires the development of 600,833 ANNs to achieve the desired levels of accuracy and confidence, as detailed in Section 2.3.1. The development of these performance benchmarks for the 754 catchments represents a significant contribution in itself, as these can be used in future studies to assess the performance of a range of ANN runoff models for any of these catchments in an unbiased fashion.

[17] To investigate the relationship between the performance of different data allocation methods and the skewness of the runoff data (Objective 2), each method is ranked (1 to 4 from best to worst for each of the 754 catchments) in terms of relative bias, variance and estimation efficiency (an equal balance between relative bias and variance). We then investigate how these ranks vary with different ranges of skewness. These results are used in the development of guidelines for selecting data allocation methods that will tend to maximize the representativeness of model evaluation performance. Details of each of these steps are provided in the following sub-sections.

2.2.1 Rainfall-runoff data

[18] Rainfall-runoff data from 754 catchments are used in this study (Figure 2), with 322 of these from Australia [*Zhang and Chiew 2009*] and the remaining 432 from the US [*Duan et al*., 2006]. These catchments represent a wide diversity in catchment properties, such as precipitation, evaporation, and catchment area (Figure 3). The Australian catchments (grey dots in Figure 3) are generally smaller in size than those from the US (black dots in Figure 3), while the relative magnitudes between annual precipitation and annual potential evaporation of these catchments are generally similar in both countries (Figure 3).

[19] Daily rainfall and runoff observations are used to develop ANN models at a daily resolution. Rainfall is recorded in millimeters per day (mm/day). Runoff is recorded in mm/day (based on catchment area) in Australia, and cubic meters per hour (m^3/h) in the US. The length of available record varies, ranging from 10 to 40 years. For an unbiased comparison, a time period of 10 years is selected for each catchment. For catchments with more than 10 years of record, the first 10-

years of observations are used, as these runoff data are less likely to be affected by urbanization relative to those from more recent years.

[20] The distribution of skewness of the daily runoff data from the 754 catchments (Figure 4) reveals a wide range of values from 1.12 to 52.03. This distribution is similar to that of the original (complete) datasets (not shown), suggesting that the selected data are adequately representative of the entire datasets in terms of data skewness.

2.2.2 Development of ANN models

2.2.2.1 Input selection

[21] For each rainfall-runoff dataset, the potential inputs are the rainfall and runoff data at various lagged time steps [*Wu et al. 2013*], with the most appropriate lags depending on catchment properties, such as area and slope. It is noted that only the rainfall and runoff data at previous time steps are considered as the candidate inputs in the present study as they have by far the biggest influence to the output (runoff predictions) as demonstrated in *Li et al.* (*2015*). In this study, a maximum lag of 10 days is used for both rainfall and runoff data, leading to a total of 20 potential inputs for each catchment. A nonlinear input variable selection algorithm based on partial mutual information (PMI), combined with the Akaike Information Criterion for stopping [*May et al., 2008*], is used to identify the significant inputs for each catchment [see *Galelli et al. 2014; Li et al. 2015*]. The results show that the number of the significant inputs for each catchment is smaller than 10, suggesting that selection of the 10-day lag is sufficient to account for the relationship between rainfall and runoff data.

2.2.2.2 Data splitting

[22] For each input-output dataset, 80% of the data are used for model calibration and the remaining 20% for evaluation, a ratio used commonly in previous studies [e.g. *May et al., 2010; Wu et al., 2013*]. Only brief details of the data splitting methods and their implementation are given below [see *Wu et al., 2013* for details]. For each different data splitting method (SS. DUPLEX, SBSS-N and SBSS-P), the data splits are generated as follows.

[23] In the semi-deterministic SS approach, the data are ordered along the output variable dimension in increasing order, and the calibration data are formed by selection of every k^{th} (e.g. k = 2, 3 or 4) sample from a random starting point. The remaining data are allocated to the evaluation subset [*Wu et al., 2013*]. The sampling interval k=5 is determined according to the specified allocation percentages between the calibration and evaluation subsets in the present study.

[24] In the deterministic DUPLEX method, samples are drawn based on Euclidean distances. The two points with the largest Euclidean distance are assigned to the calibration set, and the next pair of points that are farthest apart in the remaining list is assigned to the evaluation set. This process is repeated until the evaluation set is filled, with the remaining data allocated to the calibration set [*May et al. 2010*]. Consequently, the allocation of data to calibration and evaluation subsets is completely deterministic.

[25] In contrast, the SBSS-N and SBSS-P data splitting methods are stochastic. The SBSS approach involves two steps [*Bowden et al., 2002*]. In the first step, the data are partitioned into K strata (clusters) using a self-organizing map [SOM; *Kohonen, 1990*], which considers the distances between data points. In the second step, data for the calibration and evaluation subsets are obtained by sampling from each of these strata. For SBSS-P, the sampling is done in

proportion to the number of samples in each stratum. For SBSS-N, the sample allocation is increased for strata that contain a larger number of data points, or where the data points within a stratum have a larger variance [*May et al., 2010*].

[26] It is noted that, for the SS (semi-deterministic) and DUPLEX (deterministic) approaches, we also perform the splits 100 times for each dataset, in order to ensure consistency with the other two approaches in terms of the number of ANN models used in the calibration of bias and variance.

2.2.2.3 Model architecture and structure

[27] Following *Wu et al. [2013]*, the ANN model architecture is a general regression neural network [GRNN, *Specht 1991*]. GRNNs are akin to kernel regression methods, which have fixed structure and only require a single model parameter to be estimated – the kernel bandwidth. Because GRNNs have a fixed model structure, variations in model performance due to choices involved in the ANN model structure selection process are avoided; such choices include the appropriate number of hidden layers and nodes, their transfer functions, and the degree of connectivity. This has the desired effect of isolating the effects of the choice of evaluation data (i.e. different data splitting methods) on model performance. Furthermore, *Li et al. [2014]* have demonstrated that GRNNs can provide comparable performance to more complex ANN model architectures.

2.2.2.4 Model calibration and evaluation

[28] The ANN model calibration and evaluation approach follows *Wu et al. [2013]*. As GRNNs have only a single model parameter, the calibrated model parameters are well-defined, enabling the impact of the data splitting approach to be better isolated, rather than being confounded by other uncertainties in the ANN model development process, as mentioned above. Brent's method [*Press et al., 1992*] is used for estimation of the model parameter. The performance metric used for calibration and evaluation is the root mean squared error (RMSE).

2.2.3 Analysis of model outputs

2.2.3.1 Impact of data splitting method on evaluation performance

[29] To objectively assess the impact of different data splitting methods on model evaluation performance, the evaluation of model performance for each data splitting approach is assessed in terms of bias and variance relative to the expected benchmark error obtained using the method of *Wu et al.* [2013].

[30] Bias refers to the average of the differences between the actual evaluation errors (RMSE) and the benchmark evaluation error for the ANN models developed using the 100 data splits. A different value of bias is obtained for each combination of catchment and data splitting method, resulting in a distribution of bias values over the 754 catchments for each of the four data splitting methods considered. To enable objective comparison of the results from the 754 catchments with different skewness values, the *relative* bias (*RB*) is used:

$$RB = E[\frac{M - \overline{M}}{\overline{M}} \times 100\%] \quad (1)$$

where M is the evaluation performance of the ANN model developed for a particular catchment in terms of RMSE, and $\overline{M}$ is the benchmarking RMSE value obtained for that catchment using the

method in *Wu et al.* [2013], representing the *expected* evaluation performance over the entire dataset available for model development.

[31] Variance refers to the spread of the differences between the actual evaluation errors and the benchmark evaluation error for the ANN models developed using the 100 data splits. A different value of variance is obtained for each combination of catchment and data splitting method, resulting in a distribution of variance values over the 754 catchments for each of the four data splitting methods considered. To enable objective comparison of the results from the 754 differently skewed catchments, the *relative* variance (RV) is used:

$$RV = Var[\frac{M - \overline{M}}{\overline{M}} \times 100\%] \quad (2)$$

[32] Ideally, both *RB* and *RV* will be close to zero. The former means that the performance of the evaluated model is in agreement with the performance that would be expected based on the properties of all of the data available for model development. The latter means that the same evaluation performance is obtained for each of the 100 data splits.

[33] If relative bias is positive, the average of the actual evaluation errors obtained is greater than the benchmark error. This suggests that the evaluation errors of the developed models tend to be greater than expected, indicating that the model performance as determined by the evaluation error is pessimistic, as discussed in Section 2.1. Conversely, if relative bias is negative, the average of the actual evaluation errors obtained is less than the benchmark error. This suggests that the evaluation errors of the developed models tend to be smaller than expected, indicating that model performance as determined by the evaluation error is optimistic (see Section 2.1). Small values of relative variance indicate that the model performance is reasonably consistent each time a particular data splitting approach is implemented. This is desirable, as it indicates that using a single or small number of models with different data splits for model development can provide robust results. Large values of relative variance are undesirable, indicating that a large number of models with different data splits would need to be developed to obtain an unbiased assessment of a model's predictive capability.

[34] To obtain the required benchmark RMSE values ($\overline{M}$) an accuracy level of 1% and a confidence level of 95% are used, resulting in sample sizes ranging from 100 to 9,143 as shown in Figure 5. Overall, this requires the development of 600,883 ANN models and, consequently, a total of 902,483 ANNs are developed in this research – 301,600 for determining the *M* values and 600,883 for estimating the $\overline{M}$ values.

2.2.3.2 Relationship between performance of data splitting methods and skewness of runoff data

[35] To relate the performance of different data splitting methods to the skewness of the runoff data, three performance metrics are used, *RB*, *RV* and *REE* (relative estimation efficiency), where *REE* is computed based on an equal balance between *RB* and *RV* [*Zheng et al. 2014*] – this accounts for the fact that there is generally a trade-off between bias and variance when different data splitting methods are compared [*May et al., 2010; Wu et al. 2013*]:

$$REE = \{RB^2 + RV\}^{1/2} \quad (3)$$

[36] The relative performance of different data splitting methods is determined for each of the 754 catchments by calculating its rank based on the above three metrics. These ranks are then

averaged for catchments with runoff data that have skewness values within certain predetermined ranges (see Table 1).

[37] Finally, we investigate the relationships between the relative performance of different data splitting methods, in terms of average rank, using either *RB*, *RV* or *REE* as the performance measure, and the skewness of the runoff data. These results are used to develop guidelines regarding which data allocation method to use for model development. Note that *May et al. [2010]* and *Wu et al. [2013]* found that dimensionality of the data (the number of model inputs) also has an impact on the suitability of different data splitting methods, and that this is especially the case for the SS approach where the data are ranked only on the output variable (making it sensitive to dimensionality of data – i.e. the number of model inputs). However, as only rainfall-runoff relationships are considered here, the dimensionality of the problem is relatively low and so this is not a factor in our study. In addition, there is a high degree of correlation between the inputs and the output, making the ranking based on the runoff data representative of that of the input-output data [*Wu et al., 2013*].

3 Results and Discussion

3.1 Impact of data splitting method on evaluation performance

[38] The distributions of relative bias and relative variance obtained for the four data splitting methods are shown in Figure 6. Each distribution consists of 754 values corresponding to the different catchments considered, where each value is computed for a particular catchment over 100 different data splits. It is observed that the relative variance of DUPLEX is zero, indicating that the ANN parameter estimates for the same data splits are identical. This highlights that the ANN type used in the present study is able to eliminate any variability caused by model parameter estimation for the same data split, and can hence isolate the impacts of different data splits on evaluation performance.

[39] Overall, the results show that the method by which the data are allocated to calibration and evaluation data subsets can have a substantial impact on the assessment of model performance, with relative bias values ranging from approximately -90% (substantial over-estimation of model evaluation performance), to approximately +50% (substantial under-estimation of model evaluation performance). Clearly, the way the data are divided into calibration and evaluation subsets can have a substantial impact on the perceived predictive ability of a particular model. For example, two different modellers might develop models using exactly the same data and model development processes (e.g. same model type and structure, same percentage of data used for calibration and evaluation, same model calibration methods and procedures, same error measures), but the resulting assessments regarding model evaluation performance could be vastly different, depending on how the available data are allocated between calibration and evaluation subsets.

[40] At a finer level of granularity, we see that; a) while the overall spread of relative bias is quite large, the degree of spread in the results varies considerably with data splitting method; b) different data splitting methods either result in predominantly optimistic or pessimistic results; and c) there are substantial differences in relative variance, depending on which data splitting method is used.

[41] In terms of bias, the stochastic SBSS-P method performs extremely well, with a comparatively small range in relative bias values and an expected value of approximately zero

over the 754 catchments considered. However, the variability in performance associated with different randomly selected data split realizations is very high. This result is in agreement with *May et al. [2010]* and *Wu et al. [2013]* and can be explained by the fact that samples in this method are drawn at random from different strata / clusters so that results from individual trials can be highly variable (even though the average over a sufficiently large number of trials, here 100, tends to provide an unbiased indication of a model's predictive performance). The ability of SSBS-P to produce unbiased results if the number of data split realizations is sufficiently high is due to the random nature of the sampling approach, which ensures that even sparse areas of the input-output space can be represented in the evaluation set.

[42] In contrast, the stochastic SBSS-N method results in highly biased, optimistic assessments of model performance. This is again in agreement with *Wu et al. [2013]* for skewed data sets, and is due to the Neyman rule allocating all of the data in the sparse extreme regions of the input-output space to the calibration set. As a result, the evaluation set lacks representation of extreme events, so that prediction is relatively easy, causing evaluation errors to be smaller than expected (introducing a positive bias). Further, allocation of most of the extreme data points to the calibration set each time means that the sampling variability due to repeated implementation of the method is relatively low.

[43] For the deterministic DUPLEX method, the results in Figure 6 show that the performance of the models developed using this approach is quite pessimistic, in agreement with *Wu et al. [2013]*. In DUPLEX, data that are furthest apart in the input-output space are alternately allocated to the different subsets until only data in the largest subset remain, which are all allocated to the largest subset [*May et al.*, 2010]. Consequently, as 20% of data are allocated to the evaluation data in the present study, 40% of the most extreme data points are allocated to the calibration and evaluation subsets first, with the remaining 60% of less extreme data allocated to the calibration set. Accordingly, there is a higher proportion of extreme data points in the evaluation set and so the resulting evaluation performance can be expected to be pessimistic.

[44] Finally, the performance of models developed using the semi-deterministic SS method is slightly optimistic, in agreement with the rainfall-runoff case study in *Wu et al. [2013]*, mainly because the most extreme events are placed in the calibration data. Because the number of different starting positions from which the method can commence is limited, variability is quite low. As mentioned in Section 2.3.2, the SS method can be expected to perform well for runoff modelling, because the dimensionality of the problem is low and the ordering of the output has strong correlation with that of the inputs.

3.2 Relationship between performance of data splitting methods and skewness of runoff data

3.2.1 Ranking based on relative bias

[45] The relationships between average rank of the different data splitting methods based on relative bias and skewness of runoff are shown in Figure 7, and the effects of runoff data skewness on distributions of relative bias are shown in Figure 8. As can be seen from Figure 7, SBSS-P clearly performs the best in terms of relative bias, and is relatively insensitive to skewness of the runoff data. This is in agreement with the finding that the range of relative bias is small and centred on zero for *all* catchments (Figure 6a) for this technique. The lack of relative bias sensitivity of SBSS-P to skewness in runoff data is further confirmed by Figure 8d. As discussed previously, this is because the sampling strategy enables the full distribution of

patterns to be included in the evaluation set, even for highly skewed data, resulting in good average performance as long as the number of data splits considered is sufficiently large.

[46] From Figures 7 and 8c it is clear that SBSS-N performs poorly regardless of the skewness of the runoff data. In contrast to the SBSS-based methods, the relative bias rankings of the DUPLEX and SS methods are strongly affected by the skewness of the runoff data. SS outperforms DUPLEX for skewness values less than 20, and vice versa. For smaller skewness values, SS results in models having slightly optimistic assessment of performance, whereas the DUPLEX method results in models that are moderately pessimistic. Overall, for lower levels of skewness, SS provides better performance than DUPLEX, due to smaller absolute values of relative bias.

[47] Interestingly (see Figures 8a and 8b), the optimism of the models developed using both SS and DUPLEX increases with increasing skewness in the runoff data. The performance (measured by relative bias) of DUPLEX-based models is pessimistic for lower values of skewness, and relative bias levels decrease with increasing skewness of the runoff data. In other words, the performance of models developed using DUPLEX improves with increasing skewness, and for runoff data having skewness of 20 to 30, the expected value of the relative bias of DUPLEX-based models is close to zero (Figure 8b). In contrast, SS-based models are optimistic at low skewness values, and increased skewness in the runoff data results in a deterioration in model performance, as the resulting bias moves further away from the benchmark value (Figure 8a). This results in the switch in ranking between SS and DUPLEX observed in Figure 7. Even though DUPLEX and SS mechanisms used for allocating data to the calibration and evaluation subsets are different, both result in increasing relative proportions of more extreme data points in the calibration set as skewness increases, therefore increasing the degree of optimism, for reasons explained in Section 3.1.

3.2.2 Ranking based on relative variance

[48] The average ranks of the different data splitting methods based on relative variance and skewness of runoff are shown in Figure 9 and the effect of runoff data skewness on the distributions of relative variance is shown in Figure 10. Clearly, the relative average ranking of the data splitting methods is insensitive to relative variance (Figure 9). In all cases, the deterministic DUPLEX performs the best (Figures 6b and 10b) as it has no variance being a deterministic procedure.

[49] The performance of SS and SBSS-N is very similar, with the SS method performing slightly better (Figure 9). This is in agreement with Figure 6b, which shows the combined performance over all catchments (i.e. different levels of skewness), and is due to the fact that the performance of neither method is affected substantially by skewness of the runoff data (see Figures 10a and 10c). Being semi-deterministic, the relative variance of the SS method is relatively low, being only affected by the starting position of the sampling. Similarly, for SBSS-N, the degree of variability is also low, because the majority of extreme cases are always allocated to the calibration data, irrespective of the degree of skewness of the runoff data.

[50] The only data splitting method that is significantly affected by the skewness of the runoff data is SBSS-P, with increasing relative variance as levels of skewness increase (Figure 10d). This is because the increase in sparse regions in the input-output space associated with higher levels of skewness increases the variability in the samples drawn at random from the different

strata. However, this does not affect the relative ranking of this method, as it always performs worse, regardless of the skewness of the runoff data (Figure 9).

3.2.3 Ranking based on relative estimation efficiency

[51] The relationships between average ranks of the different data splitting methods based on relative estimation efficiency and skewness of runoff are shown in Figure 11. Even though SBSS-P performs best in terms of relative bias (Figure 7), it generally performs only third best with regard to relative estimation efficiency (which balances the importance of relative bias and relative variance). In addition, the *REE* performance of SBSS-P deteriorates with increasing skewness (Figure 11), due to the rapid increase in relative variance (Figure 10). So, while SBSS-P is an excellent data splitting approach if a large number of models with different data splits is allowed to be developed, it is not suitable when a single data split is used, especially if the runoff data are highly skewed.

[52] The results in Figure 11 confirm that SBSS-N is not a good data splitting approach for skewed data, such as those used in runoff modelling. This is because it results in highly optimistic evaluation performance for such data (Figure 8), the effects of which are not able to be offset sufficiently by its relatively low relative variance (Figure 10).

[53] Overall, Figure 11 shows that SS provides the best *REE* performance for skewness values up to approximately 20, after which the average rank of DUPLEX is lowest. The reasons for this shift were discussed in Sections 3.2.1 and 3.2.2 above. Overall, the relative bias performance of SS deteriorates with increasing skewness of the runoff data, while the variance of both methods is largely insensitive to skewness, so that relative estimation efficiency of models developed using DUPLEX improves compared to SS for higher skewness levels.

3.2.4 Guidelines for the selection of data splitting methods

[54] Based on the large-catchment-sample study results reported above, Figure 12 summarizes our proposed guidelines for the selection of data splitting methods. If time and computational ability permits model development using a large number of different data split realizations, the SBSS-P method might be the most appropriate option for catchments with various degrees of skewness of the runoff data. This is because this method performs well with close to zero bias for all catchments investigated when performance is averaged over 100 models. However, given the large relative variance of this method over models developed with different data splits, it is not applicable for single or a small number of data splits.

[55] When only a single or limited number of data split realizations is feasible, the best method to be used varies with skewness of the runoff data. If skewness values are less than 20, the SS method should be used as its relative bias tends to be close to zero (slightly optimistic) for lower values of skewness and its relative variance tends to be relatively small. At skewness values above 20, the DUPLEX method should be used (and it has the advantage of having zero variance).

[56] While the above guidelines are developed using results from a particular type of ANN model, they should be applicable to other types of models for which the time order of the data does not have to be strictly preserved. As mentioned in Section 2, the performance benchmarks for each of the 754 catchments, on which the development of the above guidelines is based, are applicable to the assessment of the bias and variance of a broad range of runoff models.

[57] An important caveat is that the above guidelines are developed for an 80%/20% (calibration/evaluation) data split and for 10 years of daily data, and if a different ratio and/or different data length is used then the range of runoff data skewness for which the SS and DUPLEX methods are preferable could change slightly. Specifically, if a greater percentage of the data are used for evaluation and / or a longer period of data are used for model development, it is likely that the number of extreme values in the evaluation data will increase, and we might expect somewhat increased pessimism in the resulting evaluation performance.

4. Summary and Conclusion

[58] Hydrological models have been used in a wide context, with typical application examples including streamflow forecasting, flood risk estimation, and water resources planning [*Westra et al. 2008; Zheng et al., 2015a,b; Cortés-Hernández et al., 2016*]. Typically, the observed data are partitioned into periods used for model development/calibration and performance evaluation, with the data allocation determined through some kind of data splitting approach (either empirical or formal). The similarity (or otherwise) of information content associated with each of these periods can have a very significant impact on both the quality of the model obtained and the quality of the assessment conducted regarding evaluation period performance. Clearly, therefore, the manner in which the data are allocated into calibration and evaluation periods is very important.

[59] The results of this study raise serious concerns regarding the common practice in which data splitting (into calibration and evaluation periods) is performed only once during model development. Unless the data allocation method can be guaranteed to provide data splits that are robust with regard to the information content required for both model development and for model assessment, one can generally expect that statistical variability in evaluation period performance will make it difficult (if not impossible) to arrive at robust conclusions regarding the quality of the model so obtained. The fact that this issue has been largely ignored in the vast majority of previous studies may be one reason why progress in hydrologic model generalization and improvement has been slow [*Gupta et al 2014*].

[60] The goal of this paper is to contribute to improved understanding of this issue, and to advocate for the development of methods that facilitate robust model development. While the issue of information content of data is arguably quite complex [*Gong et al 2013; Gupta et al 2014; Nearing and Gupta 2015; Nearing et al. 2016*], it is intuitively easy to understand the need for data to contain adequate representation of the full range of behavioural patterns for which the model can be expected to perform [*Sorooshian et al 1983; Yapo et al 1996; Bowden et al., 2002; Wu et al. 2013]*. The important questions to be addressed are: 1) how can an assessment of data quality and information content be formalized, and 2) how to achieve statistically robust model development given limitations associated with typically available hydrological data sets.

[61] This study has attempted to contribute to this discussion by focusing on one specific but important aspect of the statistical properties of available catchment data – the *skewness* associated with runoff. We specifically investigated skewness, because with increasing skewness the probability of representation of extreme events in a typical catchment dataset drops. Although it may be possible to reduce the skewness effect through transformations of the data (e.g., by transferring the ANN runoff predictions to log space), previous empirical studies have shown that such transformations were not successful in improving ANN model performance [*Bowden et al., 2003*]. This is therefore an issue that warrants further exploration.

[62] To achieve a relatively comprehensive study that has some degree of generalizability, we investigated the impact of four different formal data splitting methods on the assessed performance of hydrological ANN models, using a relative large number of rainfall-runoff datasets representing 754 catchments in Australia and the US. Our approach is further facilitated by application of the benchmarking method of *Wu et al. [2013]*, requiring the development of a total of 902,483 ANN models, and thereby providing a rigorous assessment of the statistical behaviour associated with the different data splitting methods.

[63] Our major findings can be summarized as follows:

(1) The specific choice of the data used for model evaluation has a significant impact on assessed model performance. Application of different formal data allocation methods can result in a wide range [-90% to +50%] of deviations of predictive performance from the benchmark value, which represents expected achievable evaluation performance over the entire data set. Accordingly, when using a single data split realization, it becomes virtually impossible to draw robust inferences regarding the adequacy of model performance [*Gupta et al 2012*]. This clearly highlights the importance of taking into consideration the variability caused by the data splitting approach used for model development.

(2) Among the four data splitting methods investigated, the stochastic SBSS-P method produced the lowest statistical tendency to bias in model performance, with the bias being generally insensitive to the skewness properties of the data. However, a large degree of performance variability is obtained using this approach, particularly for larger values of skewness, which means that application of SBSS-P must be accompanied by use of a statistically significant number of data splits to ensure that performance of the resulting model is robust. In contrast, the stochastic SBSS-N method consistently gave the worst performance, characterized by an unreliably optimistic assessment of model quality.

(3) When circumstances preclude the use of a large number of data split realizations, the semi-deterministic SS and deterministic DUPLEX methods may provide the most robust results. SS was found to provide better performance for data having low skewness. However, as data skewness increases to above 20, the DUPLEX method tends to provide better results. When using these methods, their particular tendencies to optimism or pessimism (and the significant variability therein) as a function of data skewness must be kept in mind.

[64] Overall, this study can be considered to be a valuable initial investigation into the broader topic of how available data should best be used for hydrological model development. Such data are invariably limited in terms of the quantity and information content relative to the kinds of hydrological behaviours one may hope that the models are capable of simulating with a respectable degree of accuracy and precision. Accordingly, it makes sense that a statistical approach should be brought to bear on the problem of optimal data use. While this study has focused on the single, albeit important, aspect of runoff data, skewness, a more comprehensive investigation into the specific attributes useful for summarizing the information content (with regard to model development) of available data sets is sorely needed. Furthermore, while the large sample approach applied here does help to guarantee some degree of statistical confidence in our results [*Gupta et al., 2014*], we acknowledge that large parts of the world are not represented herein and

in the future such studies should strive to work with a more comprehensive data set (see for example *Newman et al.*, [*2017*]).

[65] Finally, there are several important and interesting questions regarding how data should best be used for model development that remain unanswered. Firstly, the results of this study point to the possibility of designing improved formal data splitting methods that have some of the strengths of each of the four methods tested here, but are designed to minimize their weaknesses. Secondly, there remains the issue of how best to use multiple data split realizations for physical/conceptual hydrological models that require continuous input. One approach to dealing with this has been suggested by *Ebtehaj et al [2010]* based on the application of moving block bootstrap resampling. This could benefit from a more detailed statistically robust large-catchment-sample investigation. Thirdly, it has been suggested by a number of researchers, including the reviewers of this paper, that having used a split-sample approach for model development, one can then "fine tune" the model parameters using all of the data. However, we are unable to find any literature which suggests that this is actually done in practice (or even in research studies). One reason may be that doing so is not, in principle, different from ignoring the split-sample approach and proceeding to a full data calibration, which has the disadvantage of not providing an independent assessment. The paradox of data splitting is that not all of the data are actually used for model calibration. Therefore, the model prediction uncertainties, which depend on the posterior parameter uncertainties obtained during calibration, remain larger than the case when all of the data were used for parameter estimation. A clear benefit of a multi-split-sample approach is that it alerts us to the fact that the results we develop are unavoidably subject to sampling variability and data information content. However, the trade-off between the benefits of data splitting and that of using all of the data for parameter estimation are not at all well understood, and the tantalizing possibility that an optimal way of exploiting the information content of the data exists. A more meaningful approach is likely to be to develop many versions of the model using multiple equally representative splits of the data and to then use the resulting set of models in ensemble mode.

[64] In closing, it is interesting that we ran into a difference of opinion with the reviewers of this paper regarding the broader implications of our results. Although the demonstration in this paper is limited to data-driven ANN models, we believe strongly that the challenges associated with achieving multiple data splits for physically-based hydrological models (see *Ebtehaj et al [2010]*) does not exempt them from the problems discussed here, and that creative work is necessary to address this potentially critical issue. To further advance the understanding of how data allocation impacts model calibration and evaluation, we intend to pursue this line of research using other types of hydrological models (e.g., process-based models). As always, we invite discussion and collaboration on these and related issues of dynamical earth systems model development.

Acknowledgements

Professor Zheng acknowledges funding support from The National Natural Science Foundation of China (grant number 5178491), and Professor Gupta acknowledges partial support from the Australian Research Council through the Centre of Excellence for Climate System Science (grant number CE110001028). We gratefully appreciate Keith Beven, Saman Razavi and the other two anonymous reviewers for their constructive comments, which help us to improve the quality of this paper significantly. We also gratefully acknowledge data for the 432 US catchments provided by Dr. Thibault Mathevet, which can be also accessed through ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/, with details of this dataset given in http://www.nws.noaa.gov/ohd/mopex/mo_datasets.htm. Data for the Australian catchments are

synthesized based on data given in Chiew et al. (2009) and have been submitted as Supplementary document.

References

Beven, K. (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources, 16(1),* 41-51.

Beven, K. (2006), Rainfall - Runoff Modeling: Introduction, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson, pp. 1-12, John Wiley, Hoboken, N. J.

Beven, K., and Smith, P. (2015) Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. *Journal of Hydrologic Engineering, 20(1),* 6916.

Biondi, D., Freni, G., and Iacobellis, V. (2012) Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics & Chemistry of the Earth, 42-44(2),* 70-76.

Bowden, G.J., Maier, H.R., and Dandy, G.C. (2002) Optimal division of data for neural network models in water resources applications. *Water Resources Research, 38(2),* 2-1–2-11.

Bowden, G. J., G. C. Dandy, and H. R. Maier (2003), Data transformation for neural network models in water resources applications, *Journal of Hydroinformatics, 5*(4).

Bowden, G.J., Maier, H.R., and Dandy, G.C. (2012) Real - time deployment of artificial neural network forecasting models: Understanding the range of applicability. *Water Resources Research, 48(48),* 10549.

Brigode, P., L. Oudin, and C. Perrin (2013), Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, *Journal of Hydrology, 476(1),* 410-425.

Chiew, F.H.S., Teng, J., Vaze, J., Post, D.A., Perraud, J.M., Kirono, D.G.C., and Viney, N.R. (2009) Estimating climate change impact on runoff across southeast Australia: method, results, and implications of the modeling method. *Water Resources Research, 45(10),* 82-90.

Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., Brekke, L., Arnold, J., Gochis, D.J., and Rasmussen, R. (2015) A unified approach for process - based hydrologic modeling: 1. Modeling concept. *Water Resources Research, 51(4),* 2498-2514.

Clark, M.P., Nijssen, B., Lundquist, J.D., Kavetski, D., Rupp, D.E., Woods, R.A., Freer, J.E., Gutmann, E.D., Wood, A.W., and Gochis, D.J. (2015) A unified approach for process - based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research, 51(4),* 2515-2542.

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., and Hendrickx, F. (2014) On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. *Hydrology & Earth System Sciences Discussions, 10(9),* 11337-11383.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F. (2012) Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research, 48(5)*, 213-223.

de Vos, N. J., T. H. M. Rientjes, and H. V. Gupta (2010), Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering, *Hydrological Processes, 24(20)*, 2840–2850.

Dibike, Y.B., and Coulibaly, P. (2005) Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *Journal of Hydrology, 307(1)*, 145-163.

Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, F. Habets, A. Hall, and L. Hay (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *Journal of Hydrology*, *320*(1–2), 3-17.

Ebtehaj, M., H. Moradkhani, and H. V. Gupta (2010), Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, *Water Resources Research*, *46*, W07515, doi:10.1029/2009WR007981.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008), Understanding catchment behavior through stepwise model concept improvement, *Water Resources Research*, *44*(1), 186-192.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, *Water Resources Research, 47(11)*, 260-260.

Ferket, B.V.A., Samain, B., and Pauwels, V.R.N. (2010) Internal validation of conceptual rainfall-runoff models using baseflow separation. *Journal of Hydrology, 381(1)*, 158-173.

Fowler, K.J.A., Peel, M.C., Western, A.W., Zhang, L., and Peterson, T.J. (2016) Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall‐runoff models. *Water Resources Research, 52(3)*, 1820-1846.

Galelli, S., G. B. Humphrey, H. R. Maier, A. Castelletti, G. C. Dandy, and M. S. Gibbs (2014), An evaluation framework for input variable selection algorithms for environmental data-driven models, *Environmental Modelling & Software, 62*, 33-51.

Gan, T.Y., and Biftu, G.F. (1996) Automatic Calibration of Conceptual Rainfall-Runoff Models: Optimization Algorithms, Catchment Conditions, and Model Structure. *Water Resources Research, 32(12)*, 3513–3524.

Gong, W., H. V. Gupta, D. Yang, K. Sricharan, and A. O. H. Iii (2013), Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, *Water Resources Research, 49(4)*, 2253-2273.

Gibbs, M. S., D. Mcinerney, G. Humphrey, M. A. Thyer, H. R. Maier, G. C. Dandy, and D. Kavetski (2017), State Updating and Calibration Period Selection to Improve Dynamic Monthly Streamflow Forecasts for a Wetland Management Application, *Hydrology & Earth System Science*s Discussions, https://doi.org/10.5194/hess-2017-381.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian (2014), Large-sample hydrology: a need to balance depth with breadth, *Hydrology & Earth System Sciences Discussions, 10(7)*, 9147-9189.

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resources Research, 48(8)*, W08301.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology, 377*(1), 80-91.

Hartmann, G. and Bárdossy, A. (2005) Investigation of the transferability of hydrological models and a method to improve model calibration. *Advances in Geosciences, 17*, 83-87.

Humphrey, G.B., Gibbs, M.S., Dandy, G.C., and Maier, H.R. (2016) A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology, 540*, 623-640.

Humphrey, G.B., Maier, H.R., Wu, W., Mount, N.J., Dandy, G.C., Abrahart, R.J., and Dawson, C.W. (2017) Improved validation framework and R-package for artificial neural network models. Environmental Modelling & Software, 92, 82-106.

Klemes, V. (1986) Operational testing of hydrological simulation models. *Hydrological Sciences Journal, 31(3)*, 13--24.

Kohonen, T. (1990), The self-organizing map, Proceedings of the IEEE, 78(9): 1464-1480.

Li, X., A. C. Zecchin, and H. R. Maier (2014), Selection of smoothing parameter estimators for general regression neural networks-Applications to hydrological and water resources modelling, *Environmental Modelling & Software, 59*(8), 162–186.

Li, X., H. R. Maier, and A. C. Zecchin (2015), Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models, *Environmental Modelling & Software, 65*(C), 15-29.

Li, L., M. F. Lambert, H. R. Maier, D. Partington, and C. T. Simmons (2015), Assessment of the internal dynamics of the Australian Water Balance Model under different calibration regimes, *Environmental Modelling & Software, 66*, 57-68.

Luo, J., E. Wang, S. Shen, H. Zheng, and Y. Zhang (2012), Effects of conditional parameterization on performance of rainfall-runoff model regarding hydrologic non-stationarity, *Hydrological Processes, 26(26)*, 3953–3961.

Maier, H. R., A. Jain, G. C. Dandy, and K. P. Sudheer (2010), Review: Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, *Environmental Modelling & Software, 25*(8), 891-909.

May, R. J., G. C. Dandy, H. R. Maier, and J. B. Nixon (2008), Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems, *Environmental Modelling & Software, 23*(10), 1289-1299.

May, R. J., H. R. Maier, and G. C. Dandy (2010), Data splitting for artificial neural networks using SOM-based stratified sampling, *Neural Networks*, *23*(2), 283.

McInerney, D., M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera (2017), Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, *Water Resources Research*, *53*(3), 2199-2239, 10.1002/2016WR019168.

Mendoza, P.A., Clark, M.P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H. (2015) Are we unnecessarily constraining the agility of complex process ‐ based models? *Water Resources Research, 51(1),* 716-728.

Merz, R., Parajka, J., and Blöschl, G. (2011) Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research, 47(2),* 2144-2150.

Mount, N. J., H. R. Maier, E. Toth, A. Elshorbagy, D. Solomatine, F. J. Chang, and R. J. Abrahart (2016), Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, *Hydrological Sciences Journal, 61*(7), 1192-1208.

Nearing, G. S., Y. Tian, H. V. Gupta, M. P. Clark, K. W. Harrison, and S. V. Weijs (2016), A philosophical basis for hydrological uncertainty, *Hydrological Sciences Journal, 61*(9), 1666-1678.

Newman, A. J., N. Mizukami, M. P. Clark, A. W. Wood, B. Nijssen, and G. Nearing (2017), Benchmarking of a physically based hydrologic model, Journal of Hydrometeorology. DOI: 10.1175/JHM-D-16-0284.1.

Power, M. (1993) The predictive validation of ecological and environmental models. *Ecological Modelling, 68(1-2),* 33-50.

Press, W. H., S. A. Tuekolsky, W. T. Vetterling, and B. P. Falnnery (1992), Numerical Recipes: The Art of Scientific Computing, Cambridge Univ. Press, Cambridge.

Shin, M.J., Guillaume, J.H.A., Croke, B.F.W., and Jakeman, A.J. (2013) Addressing ten questions about conceptual rainfall–runoff models with global sensitivity analyses in R. *Journal of Hydrology, 503(11),* 135-152.

Sorooshian, S., V. K. Gupta, and J. L. Fulton (1983), Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall ‐ runoff models: Influence of calibration data variability and length on model credibility, *Water Resources Research, 19(1),* 251-259.

Sorooshian, S., and V. K. Gupta (1983), Automatic calibration of conceptual rainfall ‐ runoff models: The question of parameter observability and uniqueness, *Water Resources Research, 19(1),* 260-268.

Specht, D. F. (1991), A general regression neural network, *IEEE Transaction on Neural Network*, *2*(6), 568-576.

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., and Lerat, J. (2015a) Hydrology under change: an evaluation

protocol to investigate how hydrological models deal with changing catchments. *Hydrological Sciences Journal, 60(7),* 1184-1199.

Thirel, G., Andréassian, V., and Perrin, C. (2015b) On the need to test hydrological models under changing conditions. *Hydrological Sciences Journal, 60(7),* 1165-1173.

Troutman, B.M. (1985), Errors and Parameter Estimation in Precipitation Runoff Modeling: I. Theory. *Water Resources Research, 21(8),* 1195-1213.

Valipour, M., Banihabib, M.E. and Behbahani, S.M.R. (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology, 476(476),* 433-441.

Vaze, J., D. A. Post, F. H. S. Chiew, J. M. Perraud, N. R. Viney, and J. Teng (2010), Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, *Journal of Hydrology, 394(3-4),* 447-457.

Westra, S., A. Sharma, C. Brown, and U. Lall (2008), Multivariate streamflow forecasting using independent component analysis, *Water Resources Research*, *44*(2), W02437.

Wu, W., R. J. May, H. R. Maier, and G. C. Dandy (2013), A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks, *Water Resources Research, 49*(11), 7598–7614.

Wu, W., G. C. Dandy, and H. R. Maier (2014), Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling, *Environmental Modelling & Software, 54*(3), 108-127.

Yapo P, H. V. Gupta, and S. Sorooshian (1996), Automatic Calibration of Conceptual Rainfall-Runoff Models: Sensitivity to Calibration Data, *Journal of Hydrology, 181 (1-4),* 23-48

Zhang, Y., and F. H. S. Chiew (2009), Relative merits of different methods for runoff predictions in ungauged catchments, *Water Resources Research, 45*(45), 4542-4548.

Zheng, F., S. Westra, M. Leonard, and S. A. Sisson (2014), Modeling dependence between extreme rainfall and storm surge to estimate coastal flooding risk, *Water Resources Research*, *50*(3), 2050-2071.

Zheng, F., M. Leonard, and S. Westra (2015a), Efficient joint probability analysis of flood risk, *Journal Of Hydroinformatics, 17*(4), 584-597.

Zheng, F., S. Westra, and M. Leonard (2015b), Opposing local precipitation extremes, *Nature Clim. Change*, *5*(5), 389-390.

Cortés-Hernández, V. E., F. Zheng, J. Evans, M. Lambert, A. Sharma, and S. Westra (2016), Evaluating regional climate models for simulating sub-daily rainfall extremes, *Climate Dynamics*, *47*(5-6), 1613-1628.

Figure Captions

Figure 1. Outline of overall methodology

Figure 2. Catchment locations (red dots) of the 754 rainfall-runoff data from (a) Australia (322) and (b) US (432)

Figure 3. Areas (left) and annual precipitation (AP) /annual potential evaporation (APE) of the Australian (grey) and US (black) catchments

Figure 4. Distribution of the skewness of runoff for the 754 catchments considered

Figure 5. Number of random replicates of the data splits for the Australian (grey) and US (black) catchments determined using the algorithm described in Wu et al. (2013), which is used to identify the benchmarking performance values

Figure 6. (a) Relative bias and (b) Relative variance of the four data splitting methods applied to the 754 catchments. The grey vertical line in the left panel indicates zero bias

Figure 7. Averaged rank of different data splitting methods based on relative bias versus skewness of runoff (rank 1 is the best and rank 4 is the worst)

Figure 8. Relative bias for the four data splitting methods for groupings of catchments with runoff data with different ranges of skewness. The grey vertical line indicates no bias

Figure 9. Averaged rank of different data splitting methods based on relative variance versus skewness of runoff (rank 1 is the best and rank 4 is the worst)

Figure 10. Relative variance for the four data splitting methods for groupings of catchments with runoff data with different ranges of skewness.

Figure 11. Averaged rank of different data splitting methods based on relative estimation efficiency versus skewness of runoff (rank 1 is the best and rank 4 is the worst)

Figure 12. Proposed guidelines for selection of data splitting methods

Table 1. Number of catchments belonging to categories with different skewness ranges of daily runoff for assessing the relative performance of different data splitting methods

Skewness Range of Daily Runoff	Number of Catchments
<5	226
[5, 10)	239
[10, 15)	150
[15, 20)	69
[20, 25)	33
[25, 30]	23
>30	14

Figure 1.

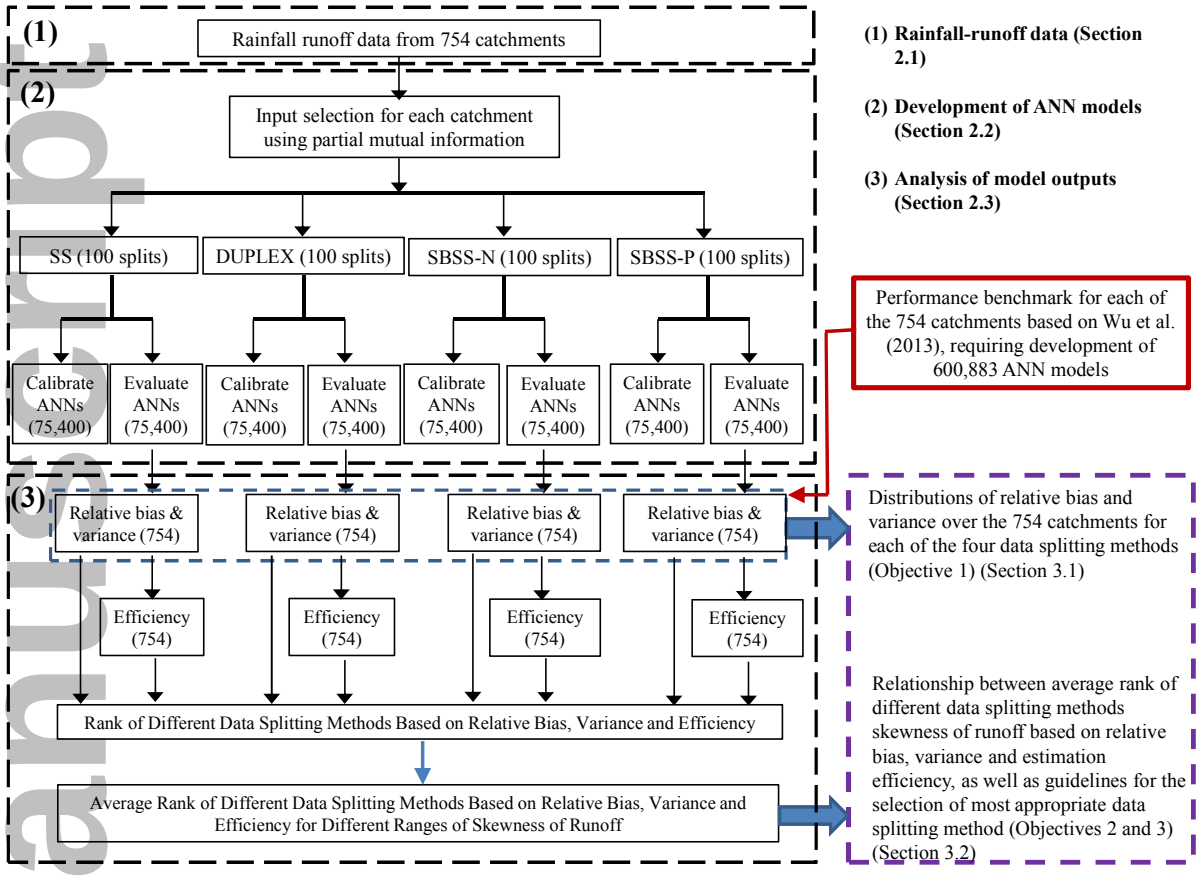
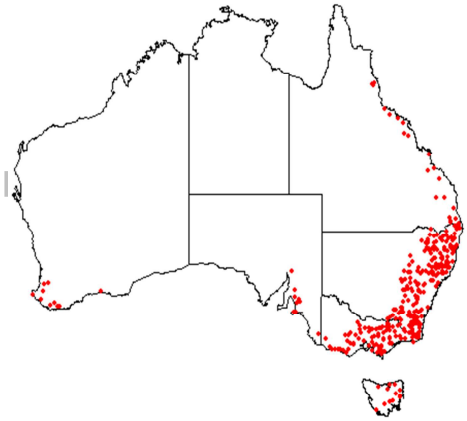
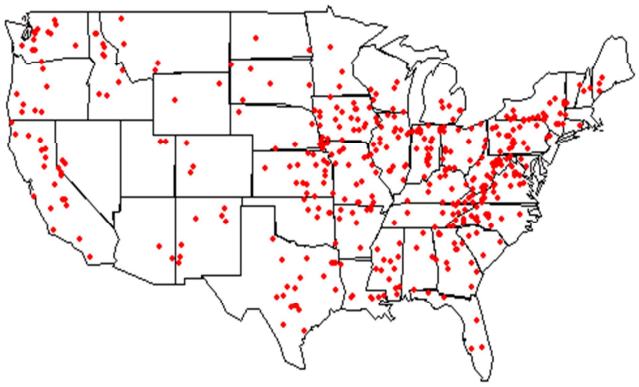


Figure 2.



(a) Australian catchments



(b) US catchments

Figure 3.

Figure 4.

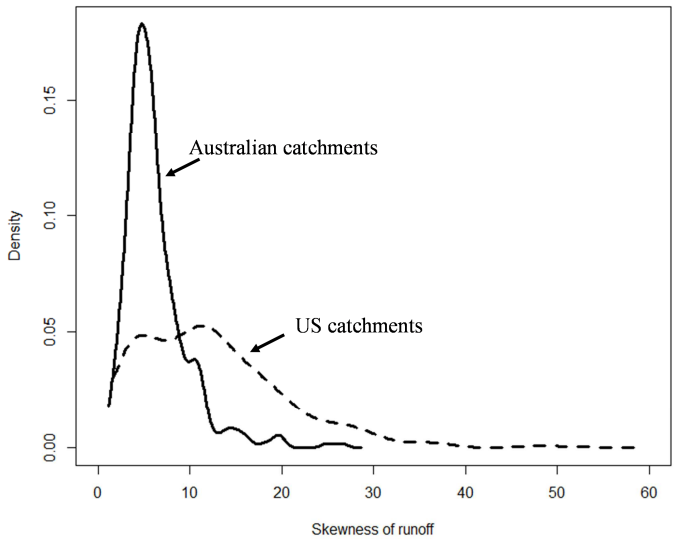


Figure 5.

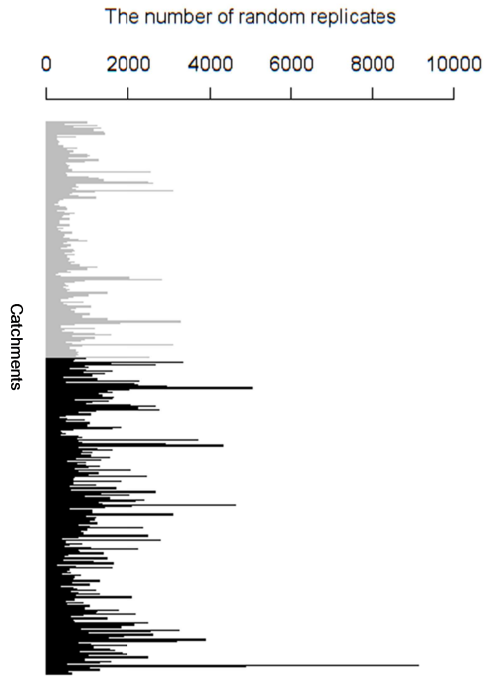


Figure 6.

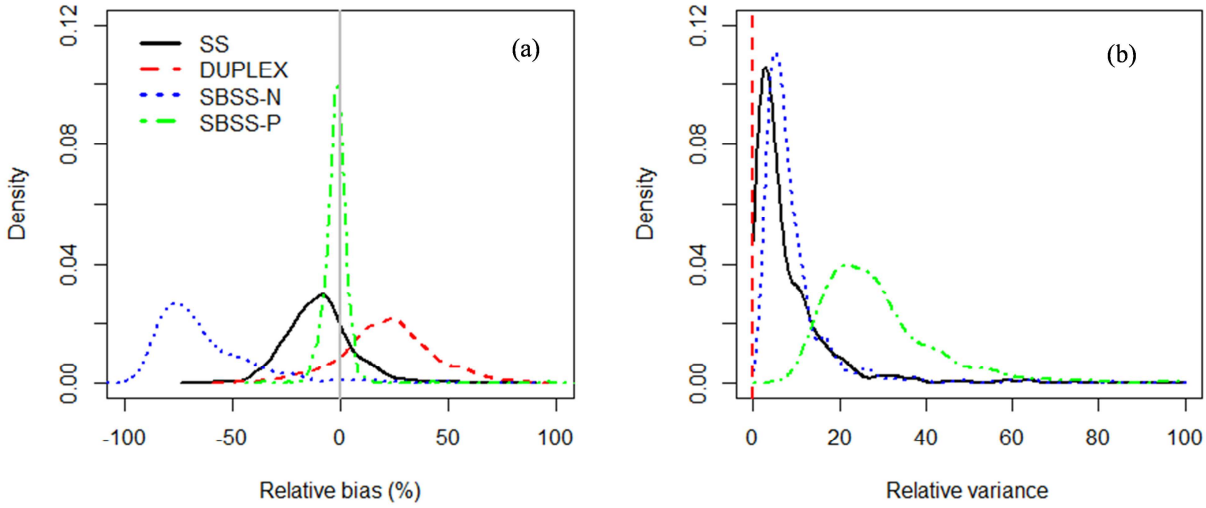


Figure 7.

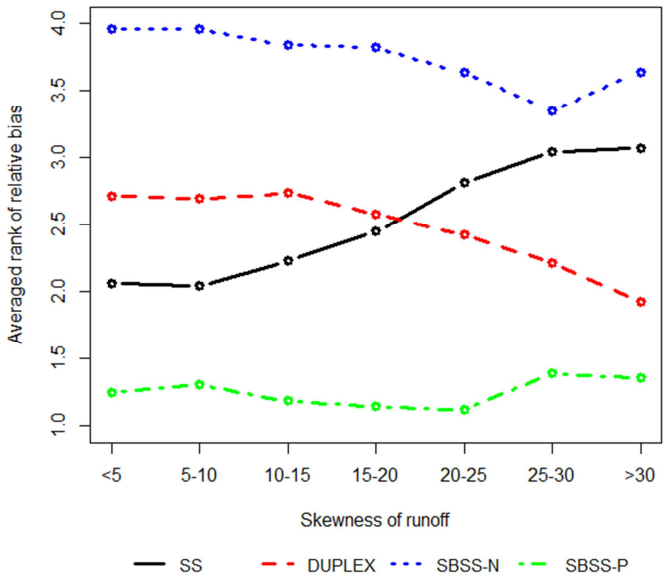


Figure 8.

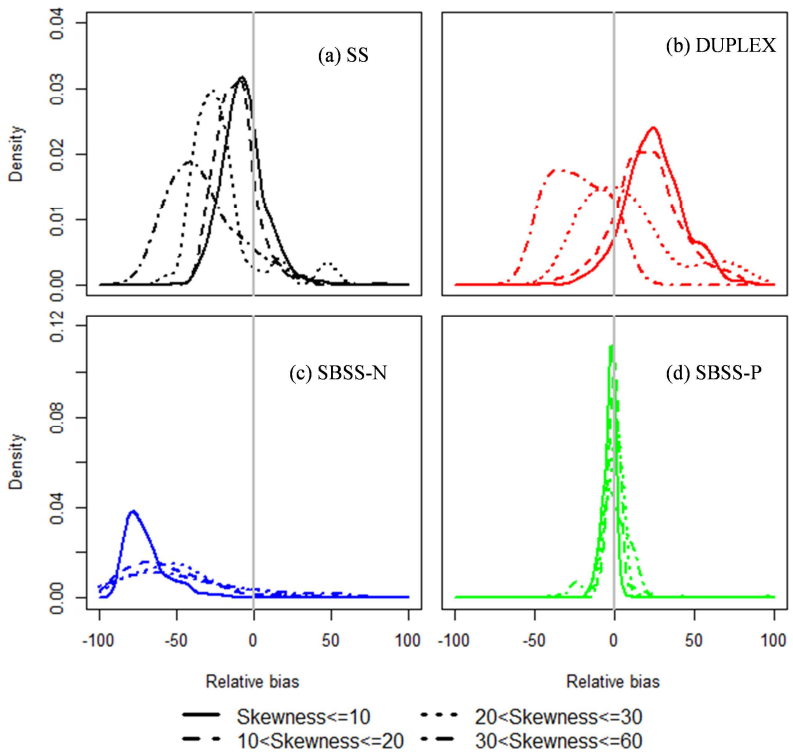


Figure 9.

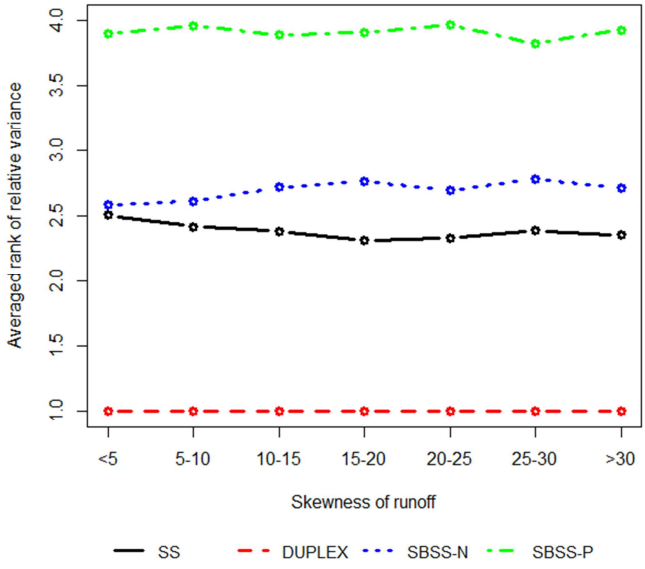


Figure 10.

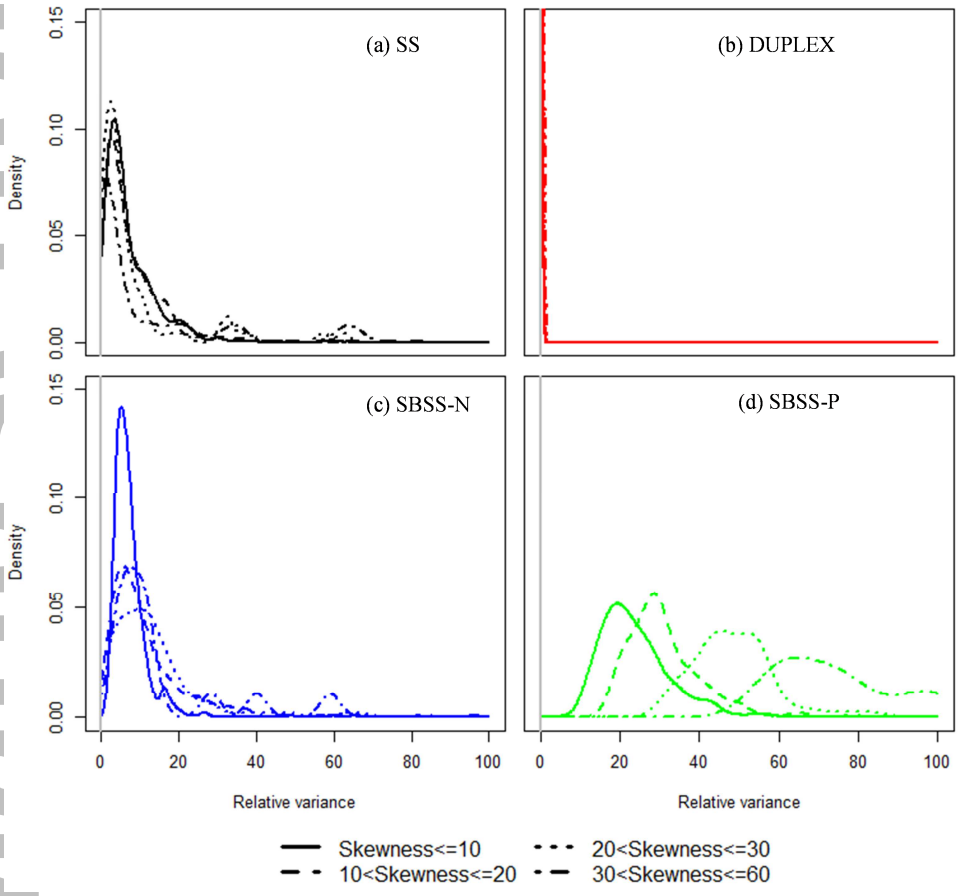


Figure 11.

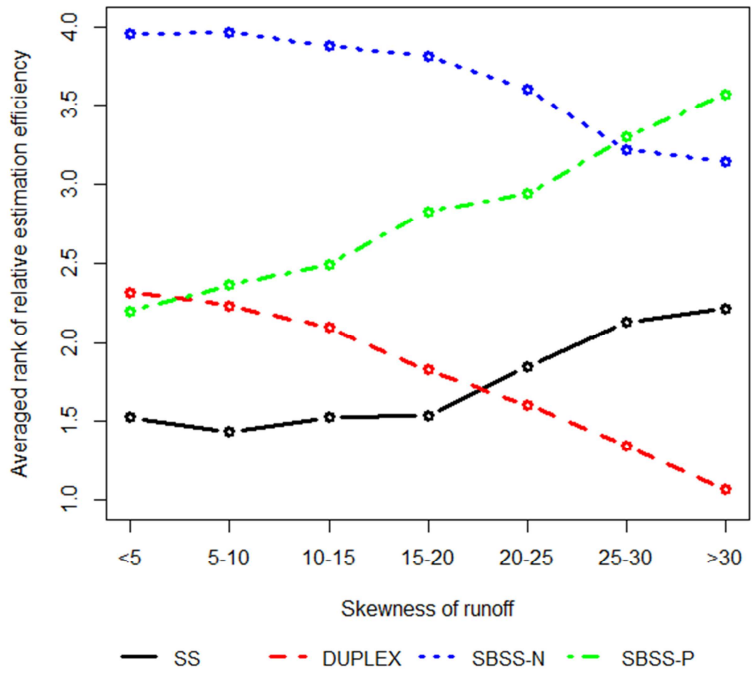


Figure 12.