Hybrid Machine Learning Framework for Hydrological Assessment

Jungho Kim[a,b], Heechan Han[c*]

Lynn E. Johnson[a,b], Sanghun Lim[d], Rob Cifelli[b]

a Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, Colorado, U.S.A.

b NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, Colorado, U.S.A.

c Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, Colorado, U.S.A.

d Water Resources Research Division, Korea Institute of Construction Technology, Ilsanseo-gu, Goyang-si, Gyenggi-do, South Korea

*Corresponding author: Heechan Han (postal address: Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, 80523, Colorado, U.S.A ; e-mail: heechan@colostate.edu)

Submitted to *Journal of Hydrology*

June, 2019

Abstract

This study introduces a novel hydrological assessment tool (HAT) based on hybrid machine learning (HML) framework. The HML framework combines an unsupervised clustering technique and a supervised classification technique, to determine reasonable performance ratings (unsatisfactory, satisfactory, good, and very good) and build a practical assessment tool. Hydrologically significant error indices are used to cluster the performance rating groups and train the HAT. The HAT was applied to the National Water Model (NWM), which is operated in real time for the continental United States (CONUS). For establishing, training, and validating the HAT, data from October 2013 to February 2017 were used, and a performance assessment was conducted on the NWM in the San Francisco Bay Area. As a result, the HAT determined the performance ratings that were reliable in terms of the statistics and hydrograph. It was confirmed that the HAT could perform an accurate hydrograph assessment as the concordance rate of the performance ratings was 98%. The NWM was evaluated against 57 USGS streamflow gauges using the HAT and was found to perform with 46% on average, good and very good ratings. The HML framework, an integral part of the HAT, is expected to be useful not only in hydrological analysis but also across all geophysical fields that deal with physical processes.

Keywords: Hydrological assessment, Hybrid machine learning, National water model, Streamflow evaluation, Performance ratings

1. Introduction

Identifying and predicting the response of hydrologic systems by using a simulation model are very important for reducing damages from natural disasters (Abbott et al., 1986; Dutta et al., 2003; Rozalis et al., 2010; Yoo et al., 2012; Kim et al., 2018a; 2018b). This is because a hydrologic model can identify in advance the potential occurrence of various water-related natural disasters as it estimates and predicts the flow and volume from surface to groundwater runoff in time and space (Henderson and Wooding, 1964). Moreover, by virtue of advanced remote sensing techniques, quantitative precipitation estimation schemes (Kim et al., 2015), and correction methods (Yoo et al., 2014; Kim and Yoo, 2014) to improve accuracy of meteorological inputs (e.g. precipitation), hydrological products from models will play a role in a wide range of disciplines. Many types of hydrologic models have advanced from the basic lumped approach that combines characteristics across an entire watershed to provide forecast information at an outlet point to distributed hydrologic models that account for spatially varying characteristics across the watershed and can be used to simulate a local-scale flood (Liang et al., 1994; Arnold et al., 1998; Singh et al., 2002). In contrast to the evolution and improvement of hydrologic modeling, general hydrological evaluation methods have remained simple, most relying on a few error indices. A hydrological evaluation method is not simply to determine whether there are many or few errors; it should reasonably determine the reliability of outputs and present objective indices understandable to users. The limitations of current hydrological evaluation methods must be overcome, and a new assessment tool is required that can objectively evaluate any hydrologic model performance.

There are many potential and important uses for the hydrological evaluation method in hydrology. Its main purposes include calibrating the model, evaluating its performance, and communicating with stakeholders. The hydrologic model, which has a complex structure and various parameters, requires a calibration process depending on the status of outputs, and the evaluation of its results determines the necessity, strategy, and extent of calibration (Moriasi et al., 2007). As the model's performance differs depending on the status of inputs arising from various meteorological forcings and geographical characteristics and the status of calibration, the hydrological evaluation method is useful for evaluation of its performance (Beven, 1993; Freer et al., 1996). Furthermore, the hydrological evaluation method serves as to provide guidance on the model's reliability to forecasters and operators who use the hydrologic model outputs for decision-making flood warnings and mitigation (Al-Sabhan et al., 2003).

For current hydrological evaluation, the graphical and statistical methods are commonly used (Green and Stephenson, 1986; Legates and McCabe, 1999; Coffey et al., 2004). The graphical method is used for a qualitative evaluation by comparing observations and simulated hydrographs, and the statistical method is used for a quantitative evaluation based on statistics for various error indices (ASCE, 1993). In general, the statistical method is based on an evaluation method that statistically divides the error index range and determines outputs in terms of various ratings (Santhi et al., 2001; Moriasi et al., 2007). Such an evaluation framework relatively straightforward process, and hence, its advantage is that it is readily applied. Nevertheless, its limitation is that it cannot present standardized ratings for various error indices. More importantly, the evaluation framework based on a single error index cannot reflect the complementary interaction between different error indices. It is also questionable how reasonably the error index range defined statistically represents the performance of a hydrologic model (Donigian et al., 1983; Ramanarayanan et al., 1997; Gupta et al., 1999; Singh et al., 2004).

Several requirements must be satisfied in developing a robust hydrological assessment tool. First, a statistical meaningful index, including error indices should be sought to ensure the objectivity of an evaluation framework. Second, a combination of complementary error indices, not a single error index, must be considered (Green and Stephenson, 1986; Coffey et al., 2004). Furthermore, the outputs of a hydrologic model suitable for the application should be used for evaluation. For example, a long-term complex hydrograph without separating single events should be avoided when evaluating a flood forecasting model as some period with no rain could play a role in generating noise that leads calculating inadequate error indices, for the purpose of hydrological assessment in flood forecasting (Ramirez, 2000). It is also important to consider the significance of the rising and recession limbs of a hydrograph as each limb represents a meaningful response of hydrological process. The rising limb is mainly formed by concentration of direct runoff which determines peak flow and time-to-peak. Since the recession limb is formed by all types of runoff, it is dominant over the rising limb in determining total runoff volume related to the water budget (Boyle et al., 2000).

Machine learning could be the alternative to overcome the shortcomings of a general evaluation method described above. Machine learning utilizes algorithms that detect patterns and relationships inherent to inputs and outputs, and is used across many areas with the development of various new algorithms and more powerful computers (Hong, 2008; Sahoo et al., 2017). Owing to an increase in the amount of data in hydrology, the use of machine learning is becoming increasingly important. More specifically, it is expected to serve as a supplementary solution in physics-based deterministic hydrology as many studies are being performed on physical factors such as surface runoff from rainfall, groundwater, and soil moisture (Coulibaly and Anctil, 1999; Tokar and Johnson, 1999; Shortridge et al., 2016).

Machine learning that can combine two or more methods for effective data analysis is referred to as Hybrid Machine Learning (HML). In general, the HML uses two machine learning techniques suitable for most application and can complement the limitations of a single technique and deliver improved outcomes (Tsai and Chen, 2010). HML has been used widely in financial applications. Hsieh (2005) combined the K-means clustering technique and the neural network technique and developed a credit scoring model based on a hybrid mining approach. Huysmans et al. (2006) used a framework that combined an unsupervised self-organizing maps technique and supervised multi-layered perception technique to obtain a new credit scoring method. Tsai and Chen (2010) reviewed various combinations of clustering machine learning techniques and classification machine learning techniques, and demonstrated a high applicability of HML in developing credit rating systems. Tsai (2014) developed a novel hybrid financial distress model based on clustering and classification machine learning for supporting financial decisions. These studies that coupled clustering and classification machine learning techniques to establish a HML framework demonstrated better results than a single machine learning technique. The HML framework is considered an attractive approach for hydrological evaluation using various error indices. A HML framework could secure a stable performance assessment by employing a big data and has an advantage to determine a composite rating metric.

This study aims to develop a novel hydrological assessment tool (HAT) by adopting HML framework based on a combination of clustering and classification techniques and a composite of error indices. National Oceanic and Atmospheric Administration (NOAA) National Water Model (NWM) is used to develop the HAT since it has enough simulation data for over 5 years for training and testing the HAT. The NWM has been operated in real time since 2016 for the continental US (CONUS) (Han et al., 2019). The performance test is conducted on rising and recession limbs in a single hydrograph as well as the total hydrograph. To build, train, and validate the model, NWM simulated streamflow from October 2013 to February 2017 is applied at selected USGS streamflow sites across the San

Francisco Bay area. The performance of the HAT is then tested against the NWM simulated streamflow data.

The rest of this paper is organized as follows: Section 2 reviews the hydrological assessment framework in flood forecasting and introduces a HML framework and the HAT used in this study. Section 3 presents data descriptions for this study, the study area, and the HAT assessment results of simulated streamflow, which is estimated by the NWM from 2013 to 2017. Section 4 compares the error-index-based results presented by previous studies for the performance test of a hydrologic model with the results of the new HAT and provides an overall discussion. Section 5 presents the conclusion of the study.

2. Materials and Methods

2.1 Hydrological Assessment Framework in Flood Forecasting Aspect

Various error indices are used for hydrological assessment. An error index is useful as it measures the simulated value against the reference value. Many cases where an error index was applied to hydrological assessment are noted in previous studies (Green and Stephenson, 1986; Legates and McCabe, 1999; Moriasi et al., 2007; Yoo et al., 2016). Table 1 lists the error indices frequently used in hydrology.

Table 1

Error indices can be classified into two types based on their purpose. The first type of index is related to hydrograph characteristic values and includes errors of peak flow, peak time, and total runoff volume. Peak flow is calculated from complex interactions between precipitation, infiltration, and effective rainfall resultant at the watershed outlet or measurement point. It is the maximum flow during the period in which direct runoff occurs intensively. Peak time refers to the time at which the peak flow occurs. As these error indices are determined by the rising limb of a hydrograph, they are very useful in assessing the performance for flood forecasting.

The second type of error index quantifies hydrograph characteristics. Most notably, it includes correlation coefficient (CC), Nash–Sutcliffe efficiency coefficient (NSE), bias and percent bias (PBIAS), and the RMSE-observations standard deviation ratio (RSR). These error indices have significance according to their development background. For instance, CC indicates a trend of simulated results against observations, whereas bias shows only average differences in ratio. As such, a single error index cannot fully represent the accuracy of a simulated hydrograph. Furthermore, even though various error indices are used together to assess a hydrograph, many individual analyses are required along with a wide range of data to reach a unified conclusion owing to the different features and scales of each indices. Fig. 1 shows poor assessment results obtained from the use of a single error index.

Fig. 1

Hydrological assessment should be based on an agile framework that can be applied in conditions appropriate for various purposes such as flood waves, low flows, and regulated flows in a river system. For the purpose of flood forecasting, an independent hydrograph is mainly assessed to test its performance in terms of surface runoff, which determines the peak value and flood risk level. Evaluation results may sharply diagnose the model performance and suggest a direction for calibration. Moreover, when a hydrological assessment is performed on a monthly or seasonal basis, it can assess the overall hydrological process but its results cannot represent the outperformance of a model in terms of flood forecasting. In addition, as long duration simulated results contain multiple peak flows, repeated rising and recession limbs, and many low flows, they can become noise when estimating error indices.

An independent hydrograph can be separated into two limbs: rising and recession limbs. The rising limb is a part of a hydrograph ranging from the initial point of the direct runoff flow to the peak flow. Conceptually, the initial direct runoff flow starts when the precipitation rate exceeds initial losses in a watershed area. In terms of flood forecasting, the rising limb is very significant as it indicates a concentration time of discharge and as it provides the trend and magnitudes of the increasing flow and a peak flow. The recession limb is the part of a hydrograph ranging from the peak flow to the point where the decreasing flow is corresponds to the discharge immediately before the initial direct runoff. In terms of water management, the recession limb is very important as all hydrological runoff components (surface, subsurface, and groundwater) occur during this time. Finally, understandable terminology must be used to allow people across different disciplines to interpret assessment results.

2.2 Hybrid Machine Learning Framework

Machine learning uses the X dataset as an independent variable and the Y label as a dependent variable, and is divided into supervised learning (SL) and unsupervised learning (USL) based on whether it has the Y label (Bishop, 2006). Some of the most widely known SL approaches include the artificial neural network (McCulloch and Pitts, 1943), the random forest (Breiman, 2001). USL approaches include the self-organizing map (Kohonen, 1982) and K-means clustering (MacQueen, 1967). In the past, it was difficult to utilize machine learning owing to the limitations of computer technology; however, machine learning is garnering significant attention with the recent advances in high performance computing.

Many hydrological applications, which generate and handle large amounts of data and information, are also applying machine learning techniques (Shrestha and Solomatine, 2006; Demissie et al., 2009).

2.2.1 Unsupervised Learning for Clustering

USL is a type of machine learning that detects complex relationships between X datasets with no determined Y label. USL is mostly used for clustering, dimension reduction, and anomaly detection. Clustering is the most widely used technique in USL, and it aims to detect similarity between datasets and to cluster similar data points into one group. In addition, it can be used to identify similarity between data points in a cluster or differences with other objects in another cluster (Tsai and Chen, 2010). Some of the most widely known clustering techniques include K-means (MacQueen, 1967), DBSCAN (Ester et al., 1996), and hierarchical clustering (Johnson, 1967).

Fig. 2

K-means clustering, proposed by MacQueen (1967), is based on non-hierarchical clustering and is effective in detecting clusters from extensive large data sets (Hartigan and Wong, 1979; Everitt et al., 2001; Olden et al., 2012). Fig. 2 shows the conceptual diagram of a K-means clustering technique. K-means includes the number of clusters as a parameter, and uses it to begin clustering initial datasets. As many centroids as a set number of clusters are randomly chosen, and centroids are changed repeatedly until the sum of the distances between each centroid and data points reaches the minimum. Finally, a centroid that has the minimum sum of distances is detected to determine the set number of clusters. The advantages of K-means are that its algorithms are simple and fast to calculate, it can obtain very reliable results, and it can be applied in various applications that involve a large amount of datasets.

2.2.2 Supervised Learning for Classification

SL is a type of machine learning that detects a pattern between the X dataset and the Y label and expresses the relationship in a function; it is used widely across disciplines that require data mining. SL can establish a model that estimates and predicts the Y label for a newly input X dataset by learning a training dataset consisting of an X dataset and Y label pair. SL is mainly used for regression and classification based on a causal relationship for datasets. The supervised classification technique is one of the most widely used techniques for statistics and engineering, and it classifies and predicts given X datasets into a suitable Y label. The dependent variable Y label serves as a category and is used for learning together with the independent variable X dataset. Classification techniques includes random forest (Breiman, 2001), support vector machine (Boser et al., 1992), and artificial neural networks (McCulloch and Pitts, 1943).

Fig. 3

Among the classification techniques, the random forest is highly applicable to applications that require the informed decision making based on numerous data, a high speed processing, and high accuracy. This technique also has an advantage that is easy to link with the USL based clustering technique for HML establishment. The random forest, which was introduced by Breiman (2001), is a type of ensemble learning based on multiple decision trees. The random forest applies randomness to not only training sets but also each decision tree's variable to reduce the high probability of overfit of the traditional decision tree method (Chagas et al., 2016). Fig. 3 illustrates the conceptual diagram of the random forest technique. First, *n* sub-training sets are randomly selected from a given total training set. Here, a sub-training set refers to a single decision tree. While the sub-training set processing is the same as that of traditional decision tree processing, available variables are applied considering randomness. The final outcome is chosen based on majority voting determined from *n* decision trees (Ließ et al., 2012; Chagas et al., 2016). As such, the random forest combines prediction results from multiple trees and makes a decision by using a bootstrap of samples similar to the conventional bootstrap aggregating method (i.e. bagging) and can achieve both predictability and stability (Cutler et al., 2007; Wang et al., 2015). In the random forest, a weight of variables is determined through measuring of contribution of the variables to the prediction accuracy and the node impurity used in training process. The descriptions of the detailed method are well documented elsewhere (Louppe et al., 2013).

2.2.3 Hybrid Machine Learning Framework

HML refers to a combination of two or more machine learning techniques (Tsai and Chen, 2010). In general, such techniques include a combination of: 1) USL techniques, 2) SL and USL techniques, or 3) a combination of SL techniques. Different HML frameworks can be established depending on the combination sequence and type of applied techniques. For a combination of SL and USL techniques, the pattern and characteristics of an X dataset can be defined by USL as a Y label, and the HML framework that shares it with SL can be established.

Fig. 4

Fig. 4 shows the conceptual diagram of a HML framework that combines the USL based clustering technique and the SL based classification technique. First, clustering creates groups (i.e. clusters) and provides them as a Y label to classification. The HML framework generates the Y label required for training in the SL technique from unsupervised clustering, and the SL technique takes charge of modeling, which is difficult in the USL technique. By doing so, the limitations of the two techniques can be mutually complemented. The Y label provided from clustering is applied to classification learning along with the X dataset, and the applicability of the model is confirmed through a verification process. The established model estimates and predicts the Y label for a new X dataset.

2.3 A Framework for Hydrological Assessment Tool

This study adopts a HML technique as described in section 2.2.3, and established a HAT that can assess the accuracy of simulated streamflow. The HML framework is configured through a combination of K-means and random forest. One of the key points in the applied HML framework is that the X dataset, an input, is clustered into multiple groups, and the group is used as the Y label required for classification. Accordingly, the representation of the Y label for the clustered X dataset group should be apparent. In the HML framework, the SL plays a role in establishing a practical model that can estimate the Y label for a new X dataset.

The HAT can evaluate rising and recession limbs for an independent hydrograph as well as the total hydrograph. The evaluation results are determined by four ratings: Very Good (VG), Good (G), Satisfactory (S), and Unsatisfactory (US), which are determined by the unsupervised clustering technique. The HAT can evaluate all streamflow hydrographs estimated or predicted using various methodologies such as deterministic and stochastic approaches. Since this HML framework has a relatively simple structure, it could be applied not only for hydrologic modeling but also more broadly for analysis of other geophysical quantities. Fig. 5 is a schematic diagram of the structure and flow of the HAT.

Fig. 5

The HAT consists of three modules. The first module is for pre-processing. This module aims to separate an independent hydrograph, identify rising and recession limbs, and calculate error indices for the independent hydrograph and two limbs. The separation process has four steps as follows:

(1) Smoothing the hydrograph to eliminate the noise due to small fluctuation (i.e. hydrological responses) in observed hydrograph. The smoothed hydrograph is used to determine the beginning and end points. At a smoothing, three points (t-1, t and t+1) arithmetic mean is used.

(2) Eliminating very low flows below threshold value. The threshold value is defined as mean observed runoff over entire period.

(3) The rate of runoff increment is used to identify the rising and recession limbs of a single hydrograph. The rate of increment at each time is defined as (runoff (t + 1) - runoff (t)) / runoff (t). Parts of rising and recession limbs are defined by setting the threshold of the increment rate for each drainage area (small: <163 km^2, medium: $<1,010$ km^2, large:$>$ 1,010 km^2). The threshold is determined by sensitivity analysis. For rising limb, the threshold values are 0.50 for small area, 0.30 for medium area and 0.25 for large area. For recession limb, the threshold values are -0.50 for small area, -0.40 for medium area and -0.20 for large

area. The beginning point at which the rising limb begins, the end point at which the recession limb ends, and the peak point at which the largest runoff occurs in the hydrograph. In the case of Recession limb, the N-days method is used to determine the point of the end point. For complex hydrographs with two or more peak flows, the largest runoff value is defined as the peak point of the hydrograph, and the rising and recession limbs are defined according to the processes previously described.

In this study, five indices to evaluate the performance of the NWM hydrologic model are used. Within the error indices shown in Table 1, this study used three (CC, NSE, PF) of them and modified two (modified PBIAS and TP) of them, to build the clustering module. The combination of the five error indices demonstrated better performance in the clustering module than the other combinations. For example, using NSE and RSR together was not as good as using only NSE as statistical meanings of the two error indices are similar (see Table 1). Each error index used in this study has a different role in determining clusters. PF and the modified TP were used as hydrograph characteristic values, and CC, NSE, and Mod-PBIAS, which quantified the characteristics of a hydrograph from various aspects, were applied as error indices. The CC shows the trend of a hydrograph and NSE shows the variance of simulated errors against observations. Mod-PBIAS refers to modified PBIAS and aims to consider errors in runoff volume. Mod-PBIAS considers the cancellation effect of the runoff volume error, which cannot be reflected by the existing PBIAS, and overcomes the limitations of the conventional method, which estimates errors only based on the observed runoff volume (Eq. (1)). Furthermore, the modified TP (hereinafter referred to as Mod-TP) was used instead of the existing TP so that the peak times that have different error directions but the same scale can be clustered in the same group (Eq. (2)). These estimated error indices are used as the X dataset in the clustering and classification modules.

$$\text{Mod-PBIAS} = ABS(\sum(Q_{obs} - Q_{sim})) \div \sum(Q_{obs} + Q_{sim}) \times 100\ (\%) \quad (1)$$

$$\text{Mod-TP} = ABS(T_{obs} - T_{sim}) \quad (2)$$

The second is the clustering module. This module determines ratings, which indicate the performance level of a hydrologic model, based on the error indices described above and provides the Y label required for training and testing in the classification module. CC, NSE, and Mod-PBIAS are applied to rising and recession limbs, and the PF (%) and the Mod-TP (hr) are used in addition to these three indices in the total hydrograph.

In the clustering process, it is necessary to determine the appropriate k as k (i.e. the number of clusters) of K-means is an important parameter that affects the reliability of the clustering result. This study implements sensitivity analysis using k values (from 4 to 30) and compares the observed and simulated hydrographs to verify clustering results in the four ratings. The sensitivity analysis consists of two steps to determine an initial k and final k. To determine the initial k, statistics (e.g. mean and variance) of error indices are used to rank in order of superiority. In order to confirm the final k, R-square value between the simulated and observed hydrographs was used as another statistics. In this study, the initial k is determined to 20. When more than 20 of k is used, it was difficult to distinguish clustered groups due to similar statistics of the groups. Conversely, when smaller than 20 of k is used, mean value of error indices was not representative of each group as variance of error indices was too wide. Final k was determined to 4 of the clustered groups referring to VG, G, S, and US.

The third module is the classification. This module is responsible for modeling, training, and testing the HAT. The range of the five error indices of clusters from the clustering module has a limitation to represent the relationship between the clusters and the ranges since it indicates only a degree of distance between a centroid of clusters and error indices. To overcome this limitation, this study employs the third module, the classification. The classification module aims to model the range of the error indices and to help understanding of the clusters from the clustering module. The classification module identifies the algorithm between the clusters and the range of error indices and builds up the knowledge for modeling the range through training. In addition, the classification module provides weights of the error indices so that it is able to analyze the contribution of the indices to clustering. The error indices are used as the X dataset and four ratings determined in the clustering module are used as the Y label. The HAT training is performed using a large amount of streamflow data, and the performance of the trained HAT can be verified from the X dataset and Y label for verification. The verified HAT can be implemented by using observed and simulated time series streamflow data, and the four ratings can be determined for the rising and recession limbs and a total.

2.4 National Water Model

The NWM is a fully distributed hydrologic model that aims to enhance flood forecasting capability of the NOAA hydrologic prediction system (Han et al., 2019). The NWM simulates the water cycle with mathematical representations of different physical processes and their interactions. This complex representation of physical processes such as rainfall rate and spatial distribution, snowmelt and infiltration and movement of water through the soil layers varies significantly with the change in terrain, soils, vegetation types, and various other variables (Cosgrove et al., 2018). The NWM is based on the community WRF-Hydro modeling system, which produces various hydrological analysis and prediction products, including gridded fields of surface runoff, soil moisture, snowpack, shallow groundwater levels, inundated area depths, and evapotranspiration; as well as estimates of river flow and velocity for approximately 2.7 million river reaches defined by the seamless National Hydrography Dataset (NHD) Plus v2.0 hydrography dataset.

The NWM ingests atmospheric forcings (e.g. temperature, humidity and precipitation rate) into a Noah-MP Land Surface Model (LSM) to simulate land surface processes at a 1-km resolution; then once exfiltration from the soil column is calculated, a diffusive wave overland routing scheme moves water horizontally across the landscape at 250 meters. Catchment aggregation occurs and distributes the water into the channel network at the end of each modeling time step, and flow is routed according to a modified Muskingum-Cunge scheme along a modified version of the NHDPlus, where waterbodies (lakes and reservoirs) are encountered on the network and store/release water according to a level pool routing scheme (https://water.noaa.gov/about/nwm).

2.5 Data

This study is performed in the nine county regions surrounding the San Francisco (SF) Bay area, California. The SF is an area of diverse topography with regions near sea level juxtaposed with mountains rising in excess of 1,000m. The SF Bay area is a flood-prone region owing to orographic rainfall occurring in steep terrain (Cifelli et al., 2018). The orographic rainfall is often produced from moisture plumes over the Pacific Ocean known as atmospheric rivers (ARs, Ralph et al., 2012). As an example, an AR event starting on December 29, 2005 brought more than 20 inches of rain across the SF Bay region. Urban areas such as the city of San Francisco recorded 24-hour rainfall totals of 5 inches on December 31 alone. There was major flooding in the Napa and Russian River basins, with 10 counties declaring federal disaster areas. Over 1,000 homes were flooded in Napa, costing over $300 million in damages. The geographic diversity and resulting flooding events in the SF Bay area provides a challenging testbed to evaluate the performance of the NWM.

Fig. 6

Fig. 6 shows the locations of the SF Bay area and stream gages that are currently operated by the USGS. A total of 91 USGS gages were identified across the nine counties in the SF Bay area. Upon review on the USGS's observed data, a subset of 57 USGS gages were selected in this study, excluding those that observed low-quality streamflow data associated with reservoir operations and diversions. The watershed for these 57 gages varies from 11.5 to 3,425.3 km^2. This study used the NWM to conduct a retrospective streamflow simulation using NLDAS forcing data (Cosgrove et al., 2003) as inputs. The HAT is developed and tested using long time period data from October 2013 to February 2017. The performance of HAT and NWM for the SF Bay area is assessed against the USGS streamflow data.

3. Results

3.1 Clustering of Rating Labels

The ratings of the four clustered groups were categorized into VG, G, S, and US. As a part of the statistical method, the characteristics of error indices for each rating were examined. Fig. 7 illustrates the probability distribution of error indices for each rating group and each error index. The results are for individual rising and recession limbs and total hydrographs. According overall results, the trend of the probability distribution depending on rating group was obvious that all of average error index from VG to US moves toward the direction of negative meaning (i.e. negative infinity for NSE), and variance increases gradually. It was also found that the percentage of a higher rating level was higher as the fraction of each rating approached the ideal error index (i.e. 1.0 for CC and NSE), whereas the percentage of a lower rating level was higher as it was further away. These results were observed in all error indices and in rising and recession limbs and total hydrographs.

Fig. 7

In addition, table 2 lists statistics of error indices depending on the performance rating for the total hydrograph case. From the table, it confirms the range features of error indices by the clustered rating level. It is found that the ranges (minimum to maximum) of the error indices were overlapped since the rating groups have clustered with a composite of the error indices. For example, a range of CC in very good rating level is from 0.74 to 1.00 and in good rating level is from 0.44 to 0.98. This result suggests that the clustered rating levels are very reasonable as there is no absolute range for performance rating. However, a range from Q1 to Q3 of the error indices was barely overlapped in the ratings. The characteristics of mean and variance statistics in the table were very obvious by each rating level, and it supports the results in Fig. 7.

Table 2

Fig. 8

It was confirmed that error indices for each rating were reasonably clustered. Subsequently, an assessment of the quality of simulated hydrologic model results for each rating was conducted. Fig. 8 shows scatter density plots between USGS observations and simulated NWM values for each rating. To remove the variability of different streamflow scales by various watershed areas and rainfall events, the observed and simulated streamflow were normalized by a peak flow so that it did not exceed 1.0. The results showed that the distribution trend of the scatter plot of each rating was distinct, and the observed trend was consistent each rating's meaning. According to the results for the total hydrograph, VG's coefficient of determination was 0.86 and was the highest, and the data points tended to cluster around the X=Y line. G showed a similar distribution trend to that of VG, but its density for the X=Y line was relatively lower and more scattered. G's coefficient of determination was 0.66. S showed a more scattered distribution trend than G, and its coefficient of determination was 0.49. For US, most of the data points were located around the X or Y axis, indicating simulated values were largely underestimated or overestimated compared with observed ones. As a result, US's coefficient of determination was 0.01, which was the lowest. The scatter plot trend for each rating was observed to be identical in the results of rising and recession limbs. In addition, Fig. 9 shows samples of the comparison results between the observed and simulated hydrographs in the four ratings (clustered groups). Runoff (Y-axis) and duration time (X-axis) are normalized using a maximum value. The results present a degree of quality of hydrograph in accordance with each rating.

Fig. 9

The clustering module determined the ratings that were reliable both statistically and graphically. The determined ratings were then used as the Y label in the classification module and served as a link between two machine learning techniques.

3.2 Classification and Verification

The classification module was built based on the supervised random forest technique, and aims to detect the hidden pattern between error indices (X dataset) and ratings (Y label). Since the random forest technique includes the SL process for modeling the evaluation tool, the classification module in which all processes were completed became the HAT that can perform a hydrological assessment on new X datasets.

Table 3

The trained classification module evaluates the performance of the model through a verification process. Table 3 lists the verification results for the trained classification module. Here, 80% of the training data was used for training whereas 20% was used for verification. The verification was performed by comparing the ratings previously determined by the clustering module and the ratings determined through the HAT. According to the results, the concordance rates of the HAT ratings were 98% (Rising), 99% (Recession), and 97% (Total), which confirms that the HAT could perform an accurate hydrograph assessment. The concordance rate for each rating was also observed to be similar to the above.

Table 4

Table 4 lists a weight of error indices determined in the classification module. In overall, Mod-PBIAS is the most important error index to assign the ratings, and CC and NSE are the next higher in order. In the case of total hydrograph, the weights of TP and PB are similar to NSE. It is speculated that the accuracy of baseflow played a role in determining the weight as the evaluation subject of the HAT is total runoff flow consisting of baseflow and direct runoff flow, not only for direct runoff flow. That could describe the main reason of why Mod-PBIAS is considered as the most important weight.

3.3 Test to Evaluate the NWM

The HAT tested the performance of the NWM for the SF Bay area through adopting a concept of leave-one-out-cross (LOOC) validation method (Efron, 1983) which is widely used in Machine Learning technique. The LOOC validation method leaves one set of total available data sets as a test set, trains the HAT using the remaining data sets except the one set and tests the NWM performance using the one set, and repeat this process as many times as needed. In this study, the entire simulation period (October 2013-February 2017) is equally divided into 10 sub-periods as the data sets by sequence of date, and the LOOC validation method is applied to each sub-period. The entire simulated results by the LOOC validation method are analyzed at various points of view. Table 5 shows the validation result of the LOOC validation method using a fraction of incorrect ratings. A range of the fractions is from 1.9 to 4.4 % on average, which confirms that the HAT is properly built and performs an accurate hydrograph assessment. However, 'Overrated' and 'Underrated' results did not show significant proportional differences.

Table 5

First, the results by drainage size are presented in Fig. 10. Overall, the performance of the NWM for the SF Bay area was rated VG or G by the HAT for at least 46% of the simulated hydrographs regardless the limbs and total hydrograph. The occurrence of VG and G increased with drainage area. For the total hydrograph results, the ratings of small areas

VG and G accounted for 42% or more, medium areas 50% or more, and large areas 58% or more. Similar trends were identical across the limbs of the hydrograph.

In this study, training of the HAT was implemented for each hydrograph limb. Thus, the distribution of the four labels could be different depending on the hydrograph limbs (i.e. rising and recession). For example, a fraction of US in the rising limb is 5% on average while a fraction of US in the recession limb is 28% which is 5 times higher.

Fig. 10

Fig. 11 shows a map representing the average ratings of total hydrograph at USGS gages and the fraction of ratings for each county. From US to VG, the model performance score ranges from 0.0 to 3.0, and the arithmetically averaged score is indicated on the map. According to the results by county, Marin County scored 0.62 points on average and showed the lowest NWM performance among six other counties except three counties whose the observed data properly usable is not found. VG and G accounted for less than 18.5%. Following Marin County, Napa County showed the second lowest performance at 1.11 points. The best performance was shown in Santa Clara County, which scored 1.79 points on average. These VG and G ratings of the county accounted for 66.7% or more. In addition, the overall results demonstrated that the NWM performance for the Southern SF Bay area (San Mateo, Santa Clara, and Alameda) was better than that for the Northern SF Bay area (Marin, Sonoma, and Napa).

Fig. 11

Since the accuracy of simulated streamflows varies with various characteristics of rainfall and watershed, it is necessary to examine how the decision of performance ratings is affected by them. Fig. 12 shows the contribution of four impact factors, complexity of hydrograph with the numbers of peak, runoff duration, drainage size and whether regulated or not, to performance ratings. In the case of complexity of hydrograph, the ratings were assigned equally regardless of the numbers of peak. Multiple peaks case has a large fraction of VG, and it confirms that the performance of the NWM for complex storm events is reliable and comparable to simulation performance for single storm events. In the case of runoff duration, G, S, and US did not show significant proportional differences by a duration length. For VG, the long duration has the largest fraction.

In the case of drainage size, the higher ratings were assigned to a large drainage area. A fraction of a large drainage area was higher at the three ratings except the US, and the small area tended to be the opposite trend of the large drainage area. There are several reasons for that. The HAT assigns the performance ratings for total runoff flows consisting of baseflows and direct flows, and a large drainage area is affected by the accuracy of baseflow, different from small drainage areas commonly located in the upper river basin. Also, Mod-PBIAS among the error indices is highly influenced to determine the performance rating. In the case of whether regulated or not, G, S and US did not show significant proportional differences, and a fraction of unregulated was higher at VG.

Fig. 12

4. Discussion

One of the most notable hydrological evaluation framework studies was conducted by Moriasi et al. (2007) who suggested general hydrological assessment guidelines. Their study determined classification criteria for an error index through the basic framework of decision trees, and tested the performance of a hydrologic model based on the determined classification criteria. However, their evaluation method can only be used for single indices, and it is difficult to draw a comprehensive conclusion from various indices. Fig. 13 compares the results of the HAT and Moriasi et al. (2007). NSE, PBIAS, and RSR error indices were used for the results of Moriasi et al. (2007).

Fig. 13

The results using the Moriasi et al. (2007) methodology are difficult to interpret in terms of an overall performance rating result. Graphically, the scatter plot distributions of the top three ratings (VG, G, and S) are so similar that it was difficult to distinguish them. The US rating showed no trend in the scatter plot distribution. These results could be reaffirmed by the coefficient of determination. In particular, there were few differences in the coefficient of determination between VG, G, and S, and hence, it was difficult to determine which rating shows high accuracy. When PBIAS was applied, the coefficient of determination of the three ratings ranged from 0.75 to 0.77, and the coefficient of determination for the G rating was estimated to be higher than that of the VG rating. In NSE and RSR, the coefficient of determination for the three ratings ranged from 0.85 to 0.92 and from 0.84 to 0.92, respectively, which was similar to that in PBIAS. While the coefficient of determination of the US rating was estimated to be much lower than those of the top three ratings, it was difficult to conclude that US was assessed well, given that there was no trend in the scatter plot distribution. The advantage of the HAT is, that by objectively combining the indices into an objective algorithm, an overall assessment of the model performance is easier to obtain. In addition, table 6 shows the comparison results of ranges of error indices derived from the HAT and Moriasi et al. (2007). It confirms that the absolute ranges of error indices used in the general evaluation may not reasonable to evaluate simulation results.

Table 6

The HAT showed a high accuracy of over 98% in the verification results. To further improve the performance of the HAT, we believe that a model that uses more training data than those used in this study should be established. For 2%, the ratings were underestimated compared with the actual ratings in all cases. These results may be obtained owing to the use of the random forest, apart from whether the amount of data is simply large or small. The random forest is a machine learning technique that supplements flexibility, which decision trees do not have, and determines classification criteria between the given X dataset and Y label from various decision trees. This technique, however, cannot implement perfect classification criteria without infinite training data owing to the fundamental problem of decision trees-discontinuous classification criteria-even if the optimized classification criteria are determined based on multiple decision trees. Nevertheless, 98% accuracy achieved by the HAT can be considered acceptable, and we believe that the ratings for the hydrologic model determined via the HAT established based on such a performance are reliable.

Understanding uncertainties in the procedures needs for meaningful quantification of the results. In case of this study, uncertainties may arise from two parts: the hydrograph separation and the four ratings assignment. The hydrograph separation is the important process as it determines an independent hydrograph as well as two limbs (i.e. rising and recession) which is the source for evaluation criteria of the hydrologic model performance.

Thus, the results could be slightly varied with the separation methods, especially in determining the end point of a hydrograph. However, it is speculated that the uncertainties from the hydrograph separation are not big enough to change the results as the error indices were barely changed depending on the lengths of a hydrograph. On the other hands, since the rating assignment is a key to evaluate a hydrograph whether it is good or not, the parameter k in the cluster module is very important. In this study, k is determined by the sensitivity analysis method so that the result may include a subjective point of view.

5. Summary and Conclusions

This study describes the HAT based on the HML technique. The HML technique was established by a combination of clustering and classification techniques, and ratings were reasonably determined from a composite of various error indices. The HAT was applied to retrospective simulations of the NWM in the SF Bay area. Conclusions from this study include:

1) A novel assessment tool, HAT, has been developed. Four ratings determined by the HAT accompanied apparent statistical and graphical characteristics and could accurately diagnose outputs for each rating. Accordingly, it could define the status of the model for each rating objectively, and the HAT is expected to be applied to determine the necessity, strategy, and extent of calibration.
2) Through the training and verification processes, we confirmed the reliability of the HAT, and showed that HAT could assess a single hydrograph from three aspects, the rising and recession limbs and total hydrograph. Moreover, easy-to-understand terms were used to define ratings and help understand the assessment results.

3) The HAT assessed the performance of the NWM for the SF Bay area using a limited training and verification data set. The NWM was shown to perform G-VG for at least 46% of the hydrographs examined during from October 2013 to February 2017, regardless of the watershed size.

The new evaluation framework is extensively applicable. The HAT is able to rate for additional performance levels (e.g. super-very-good and super-unsatisfactory) by adding new groups, as it is very flexible. If sub-hourly evaluation is needed like a flash flood, the HAT could implement that through training the HAT based on sub-hourly time step data. Also, the HAT can be applied to not only a flood forecasting model but also any geophysical data that are driven by physically pulsed phenomena. For instance, it can be applied to the indices that represent precipitation, soil moisture content, underground water, pollution load, and natural disasters.

Acknowledgments

This work was supported by the California Department of Water Resources and the NOAA Physical Sciences Division. And this research was supported by Basic Science Research Program through the National Research Foundation of Korea

References

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., Rasmussen, J., 1986. An introduction to the European Hydrological System—Systeme Hydrologique Europeen,“SHE”, 1: History and philosophy of a physically-based, distributed modelling system. *Journal of hydrology*, 87(1–2), 45–59. doi.org/10.1016/0022-

1694(86)90114-9.

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division., 1993. Criteria for evaluation of watershed models. *Journal of Irrigation and Drainage Engineering*, 119(3), 429–442. doi.org/10.1061/(ASCE)0733-9437(1993)119:3(429).

Al-Sabhan, W., Mulligan, M., Blackburn, G. A., 2003. A real-time hydrological model for flood prediction using GIS and the WWW. *Computers, Environment and Urban Systems*, 27(1), 9–32. doi.org/10.1016/S0198-9715(01)00010-2.

Arnold, J. G., Srinivasan, R., Muttiah, R. S., Williams, J. R., 1998. Large area hydrologic modeling and assessment part I: model development. *Journal of the American Water Resources Association*, 34(1), 73–89. doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Beven, K.,1993. Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in water resources*, 16(1), 41–51. doi.org/10.1016/0309-1708(93)90028-E.

Boser, B. E., Guyon, I. M., Vapnik, V. N.,1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 144–152. doi: 10.1145/130385.130401.

Boyle, D. P., Gupta, H. V., Sorooshian, S.,2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, 36(12), 3663–3674. doi.org/10.1029/2000WR900207.

Breiman, L.,2001. Random forests. *Machine learning*, 45(1), 5–32. doi.org/10.1023/A:101093340.

Cifelli, R., Chandrasekar, V., Chen, H., Johnson, L. E.,2018. High resolution radar quantitative precipitation estimation in the San Francisco Bay area: Rainfall monitoring for the urban environment. *Journal of the Meteorological Society of Japan*. Ser. II, 96, 141-155. doi.org/10.2151/jmsj.2018-016.

C.M. Bishop, Pattern Recognition and Machine Learning. Springer, Aug. 2006.

Coffey, M. E., Workman, S. R., Taraba, J. L., Fogle, A. W.,2004. Statistical procedures for evaluating daily and monthly hydrologic model predictions. *Transactions of the ASAE*, 47(1), 59. doi: 10.13031/2013.15870.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Marshall, C., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., Meng, J.,2003. Real time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research*, 108(D22). doi.org/10.1029/2002JD003118.

Cosgrove, B. A., Gochis, D. J., Graziano, T., Clark, E., Flowers, T.,2018. An update on the NOAA National Water Model and Related Activities. 98th Annual Meeting American Meteorological Society, Austin, 7–11 January 2018.

Coulibaly, P., Anctil, F.,1999. Real-time short-term natural water inflows forecasting using recurrent neural networks. International Joint Conference on IEEE, 6, 3802–3805. 10.1109/IJCNN.1999.830759.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J.,2007. Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792. doi.org/10.1890/07-0539.1.

da Silva Chagas, C., de Carvalho Junior, W., Bhering, S. B., Calderano Filho, B.,2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena*, 139, 232–240. doi.org/10.1016/j.catena.2016.01.001.

Demissie, Y. K., Valocchi, A. J., Minsker, B. S., Bailey, B. A.,2009. Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of hydrology*, 364(3–4), 257–271. doi.org/10.1016/j.jhydrol.2008.11.007.

Donigian Jr, A. S., Imhoff, J. C., Bicknell, B. R.,1983. Predicting water quality resulting from agricultural nonpoint source pollution via simulation: HSPF. *In Agricultural Management and Water Quality*, 200–249.

Dutta, D., Herath, S., Musiake, K.,2003. A mathematical model for flood loss estimation. *Journal of hydrology*, 277(1–2), 24–49. doi.org/10.1016/S0022-1694(03)00084-2.

Efron, B., 1983. Estimating the error rate of a prediction rule: some improvements on cross-validation. Journal of the American Statistical Association, 78, 316–331. doi: 10.2307/2288636.

Ester, M., Kriegel, H. P., Sander, J., Xu, X.,1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, 96(34), 226–231.

Everitt, B. S., Landau, S., Leese, M., Stahl, D.,2001. Cluster Analysis, (4th edn), Arnold: London.

Freer, J., Beven, K., Ambroise, B.,1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, 32(7), 2161–2173. doi.org/10.1029/95WR03723.

Green, I. R. A., Stephenson, D.,1986. Criteria for comparison of single event models. *Hydrological Sciences Journal*, 31(3), 395–411. doi.org/10.1080/02626668609491056.

Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F.,2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. doi.org/10.1016/j.jhydrol.2009.08.003.

Han, H., Kim, J., Chandrasekar, V., Choi, J., Lim, S., 2019. Modeling streamflow enhanced by precipitation from Atmospheric River using the NOAA national water model: a case study of Russian River basin for February 2004. *Atmosphere*, under revision.

Hartigan, J. A., Wong, M. A.,1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*. Series C (Applied Statistics), 28(1), 100–108. DOI: 10.2307/2346830.

Henderson, F. M., Wooding, R. A.,1964. Overland flow and groundwater flow from a steady rainfall of finite duration. *Journal of Geophysical Research*, 69(8), 1531–1540. doi.org/10.1029/JZ069i008p01531.

Hong, W. C.,2008. Rainfall forecasting by technological machine learning models. *Applied Mathematics and Computation*, 200(1), 41–57. doi.org/10.1016/j.amc.2007.10.046.

Hsieh, N. C.,2005. Hybrid mining approach in the design of credit scoring models. *Expert systems with applications*, 28(4), 655–665. doi.org/10.1016/j.eswa.2004.12.022.

Huysmans, J., Baesens, B., Vanthienen, J., Van Gestel, T.,2006. Failure prediction with self organizing maps. *Expert Systems with Applications*, 30(3), 479–487. doi.org/10.1016/j.eswa.2005.10.005.

Johnson, S. C.,1967. Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254. doi.org/10.1007/BF02289588.

Kim, J., Johnson, L., Cifelli, R., Choi, J., and Chandrasekar, V., 2018a. Derivation of soil moisture recovery relation using SCS curve number method. Water, 10(7), 1-21. doi.org/10.3390/w10070833.

Kim, J., Lee, J., Song, Y., Han, H., and Joo, J., 2018b. Modeling the runoff reduction effect of low impact development installations in an industrial area, South Korea. Water, 10(8), 1-15. doi.org/10.3390/w10080967.

Kim, J., Yoo, C., 2014. Use of a dual Kalman filter for real-time correction of mean field bias of radar rain rate. Journal of Hydrology, 519(Part D), 2785-2796. doi.org/10.1016/j.jhydrol.2014.09.072.

Kim, J., Yoo, C., Lim, S., Choi, J., 2015. Usefulness of relay-information-transfer for radar QPE. Journal of Hydrology, 531, 308-319. doi.org/10.1016/j.jhydrol.2015.07.006.

Kohonen, T.,1982. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1), 59–69. doi.org/10.1007/BF00337288.

Legates, D. R., McCabe Jr, G. J.,1999. Evaluating the use of “goodness of fit” measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1), 233–241. doi.org/10.1029/1998WR900018.

Liang, X., Lettenmaier, D. P., Wood, E. F., Burges, S. J.,1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, 99(D7), 14415–14428. doi.org/10.1029/94JD00483.

Ließ, M., Glaser, B., Huwe, B.,2012. Uncertainty in the spatial prediction of soil texture:

comparison of regression tree and Random Forest models. *Geoderma*, 170, 70–79. doi.org/10.1016/j.geoderma.2011.10.010.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.,2013. Understanding variable importances in forests of randomized trees. *In Advances in neural information processing systems*, 431-439.

MacQueen, J.,1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1(14), 281–297.

McCulloch, W. S., Pitts, W.,1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115–133. https://doi.org/10.1007/BF02478259.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L.,2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885–900. doi: 10.13031/2013.23153.

Olden, J. D., Kennard, M. J., Pusey, B. J.,2012. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. *Ecohydrology*, 5(4), 503–518. doi.org/10.1002/eco.251

Ramanarayanan, T. S., Williams, J. R., Dugas, W. A., Hauck, L. M., McFarland, A. M.

S.,1997. Using APEX to identify alternative practices for animal waste management (No. 972209). ASAE Paper.

Ralph, F. M., T. Coleman, P.J. Neiman, R. Zamora,, M.D. Dettinger, 2012. Observed impacts of duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal northern California. *Journal of Hydrometeorology*. 14 (2), 443-459. doi:10.1175/jhm-d-12-076.1.

Ramirez, J. A.,2000. Prediction and modeling of flood hydrology and hydraulics. Inland flood hazards: Human, riparian and aquatic communities, Cambridge University Press.

Rozalis, S., Morin, E., Yair, Y., Price, C.,2010. Flash flood prediction using an uncalibrated hydrological model and radar rainfall data in a Mediterranean watershed under changing hydrological conditions. *Journal of hydrology*, 394(1–2), 245–255. doi.org/10.1016/j.jhydrol.2010.03.021.

Sahoo, S., Russo, T. A., Elliott, J., Foster, I.,2017. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. *Water Resources Research*, 53(5), 3878–3895. doi.org/10.1002/2016WR019933.

Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., Srinivasan, R., Hauck, L. M.,2001. Validation of the swat model on a large RWER basin with point and nonpoint sources. *Journal of the American Water Resources Association*, 37(5), 1169–1188. doi.org/10.1111/j.1752-1688.2001.tb03630.x.

Shortridge, J. E., Guikema, S. D., Zaitchik, B. F.,2016. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016.

Shrestha, D. L., Solomatine, D. P.,2006. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2), 225–235. doi.org/10.1016/j.neunet.2006.01.012.

Singh, J., Knapp, H. V., Arnold, J. G., Demissie, M.,2005. Hydrological modeling of the Iroquois river watershed using HSPF and SWAT. *Journal of the American Water Resources Association*, 41(2), 343–360. doi.org/10.1111/j.1752-1688.2005.tb03740.x.

Singh, V. P., Woolhiser, D. A.,2002. Mathematical modeling of watershed hydrology. *Journal of hydrologic engineering*, 7(4), 270–292. doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270).

Tokar, A. S., Johnson, P. A.,1999. Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, 4(3), 232–239. doi.org/10.1061/(ASCE)1084-0699(1999)4:3(232).

Tsai, C. F., Chen, M. L.,2010. Credit rating by hybrid machine learning techniques. *Applied soft computing*, 10(2), 374–380. doi.org/10.1016/j.asoc.2009.08.003.

Tsai, C. F.,2014. Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, 16, 46–58. doi.org/10.1016/j.inffus.2011.12.001.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X.,2015. Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527, 1130–1141. doi.org/10.1016/j.jhydrol.2015.06.008.

Yoo, C., Kim, J., Yoon, J., 2012. Uncertainty of areal average rainfall and its effect on runoff simulation: a case study for the Chungju Dam basin, Korea, *KSCE Journal of Civil Engineering*, 16(6), 1085-1092. doi.org/10.1007/s12205-012-1646-x.

Yoo, C., Park, C., Yoon, J., Kim, J., 2014. Interpretation of mean-field bias correction of radar rain rate using the concept of linear regression, *Hydrological Processes*, 28(19), 5081-5092. doi.org/10.1002/hyp.9972.

Yoo. C., Ku, J., Yoon, J., Kim, J., 2016, Evaluation of error indices of radar rain rate targeting rainfall-runoff analysis. *ASCE Journal of Hydrologic Engineering*, 21(9), 1-12. doi.org/10.1061/(ASCE)HE.1943-5584.0001393.

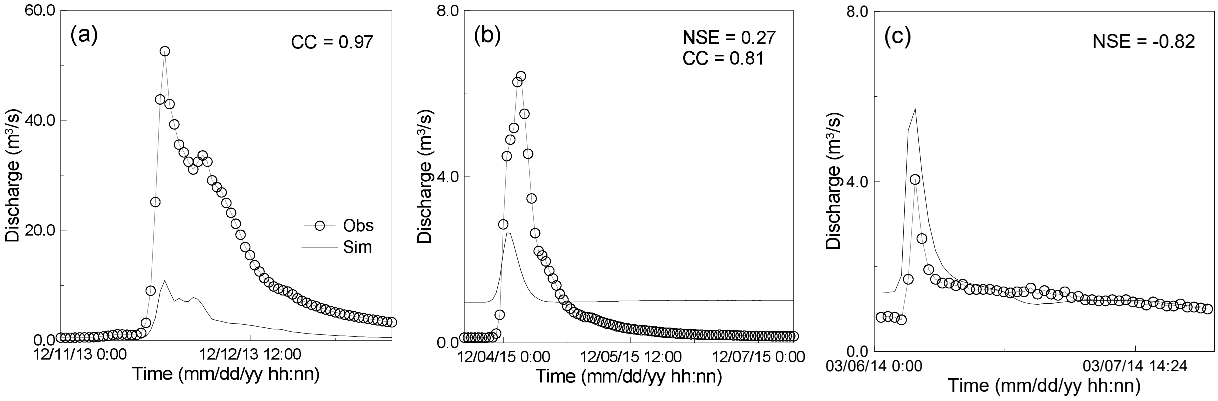
List of Figures

value of the fraction. The results include rising and recession limbs and total hydrograph.

Figure 11 Assessment rating map in San Francisco Bay area with the fraction (%) of the four ratings for counties. The result is for total hydrographs of Feb 2017.

Figure 12 Contribution of impact factors, (a) 'Hydrograph', (b) 'Runoff duration', (c) 'Drainage size' and (d) 'Regulation', to evaluation model performance for training datasets. (a) 'Hydrograph' indicates the complexity of hydrograph based on the numbers of peak, (b) 'Runoff duration' shows the proportions of each rating based on period between beginning and end points of the hydrograph. This includes three periods which are short-term (<36 hr), medium-term (<72 hr) and long-term (>72 hr). (c) 'Drainage size' also shows three drainage sizes which are small (163 km^2), medium (<1,010km^2) and large (>1,010km^2). (d) 'Regulation' means whether the drainage area is regulated by the anthropogenic activities such as reservoir operations and diversions or not.

Figure 13 Same as Figure 8, but for a comparison result of density scatter plots of ratings determined by the HAT and the general evaluation framework with NSE, PBIAS, and RSR.



(a)
CC = 0.97
Obs
Sim
Discharge (m3/s)
60.0
40.0
20.0
0.0
12/11/13 0:00
12/12/13 12:00
Time (mm/dd/yy hh:nn)
(b)
NSE = 0.27
CC = 0.81
Discharge (m3/s)
8.0
4.0
0.0
12/04/15 0:00
12/05/15 12:00
12/07/15 0:00
Time (mm/dd/yy hh:nn)
(c)
NSE = -0.82
Discharge (m3/s)
8.0
4.0
0.0
03/06/14 0:00
03/07/14 14:24
Time (mm/dd/yy hh:nn)

Figure 1

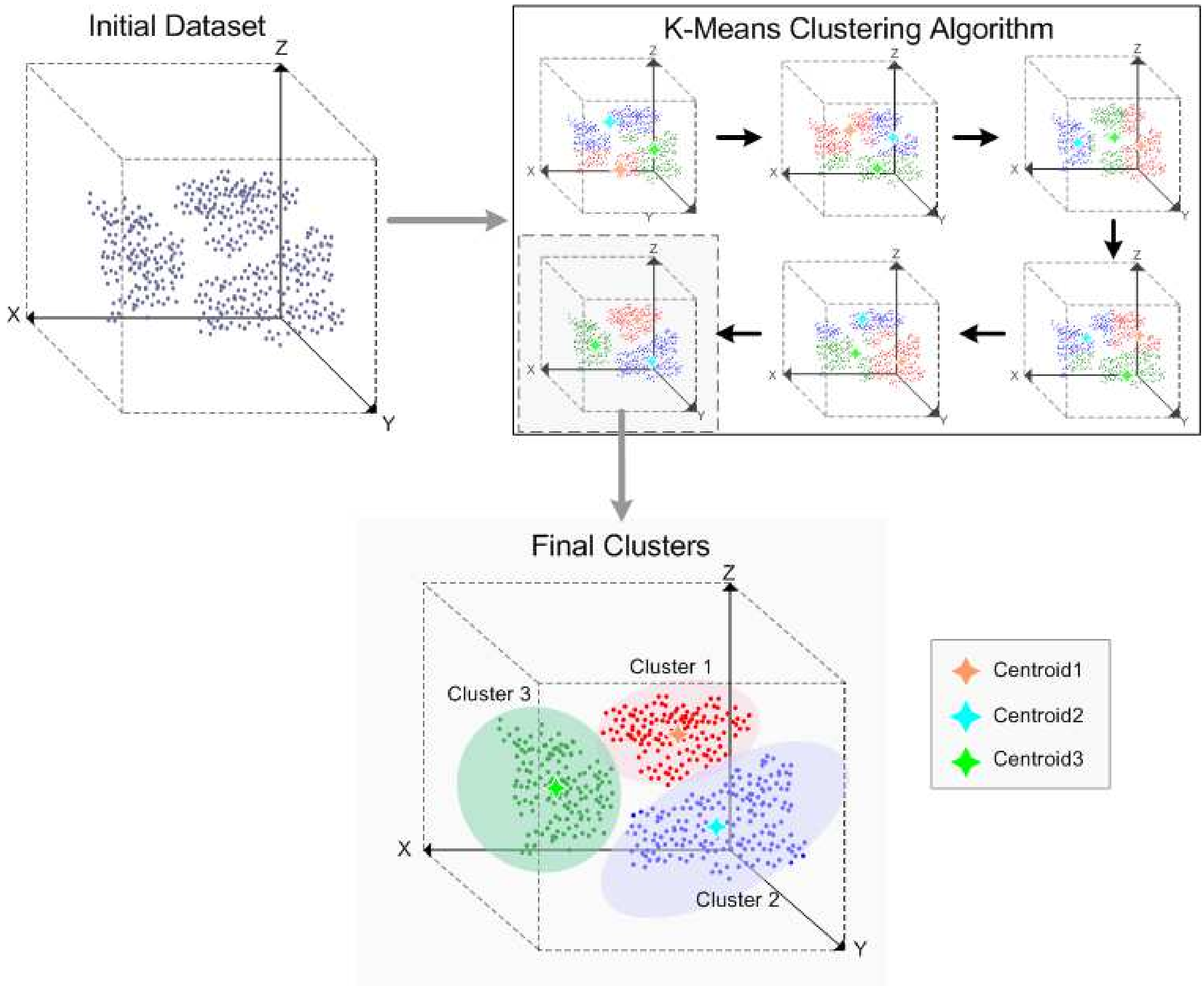

Initial Dataset
K-Means Clustering Algorithm
Final Clusters
Cluster 1
Cluster 2
Cluster 3
Centroid1
Centroid2
Centroid3
X
Y
Z

Figure 2

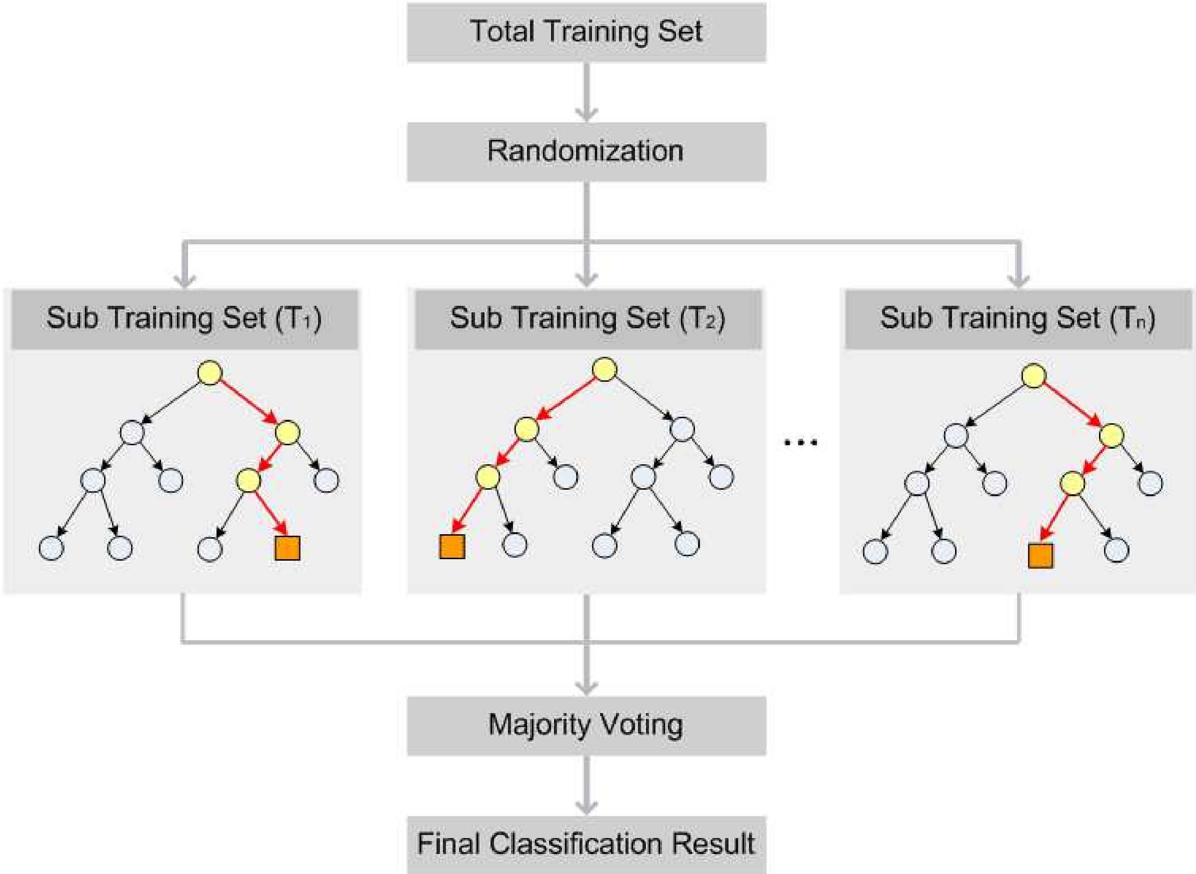

Total Training Set
Randomization
Sub Training Set (T_1)
Sub Training Set (T_2)
...
Sub Training Set (T_n)
Majority Voting
Final Classification Result

Figure 3

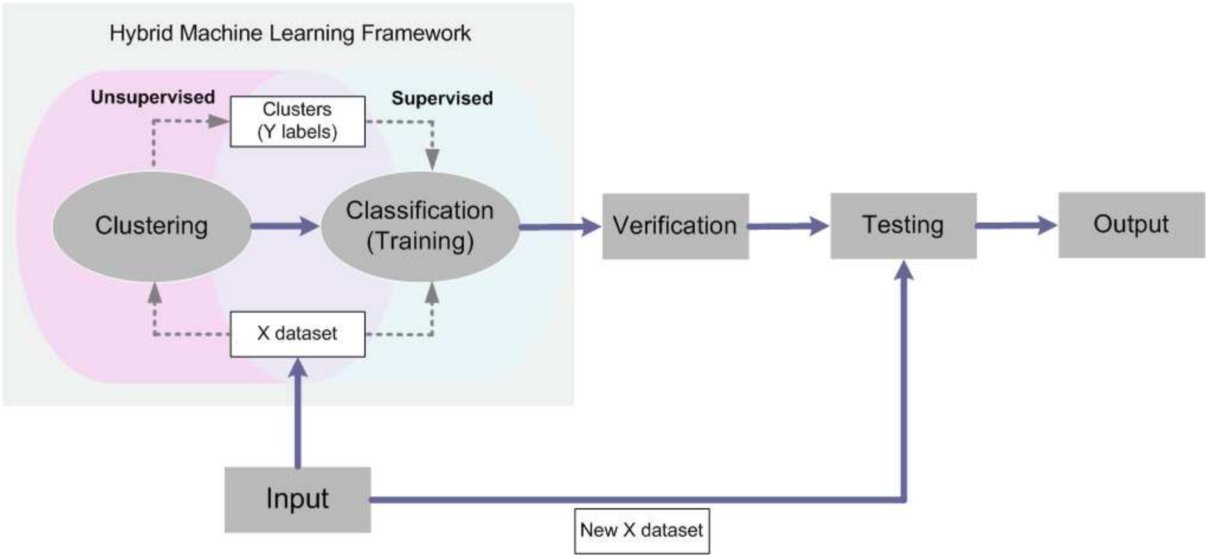

Hybrid Machine Learning Framework
Unsupervised
Clusters
(Y labels)
Supervised
Clustering
Classification
(Training)
Verification
Testing
Output
X dataset
Input
New X dataset

Figure 4

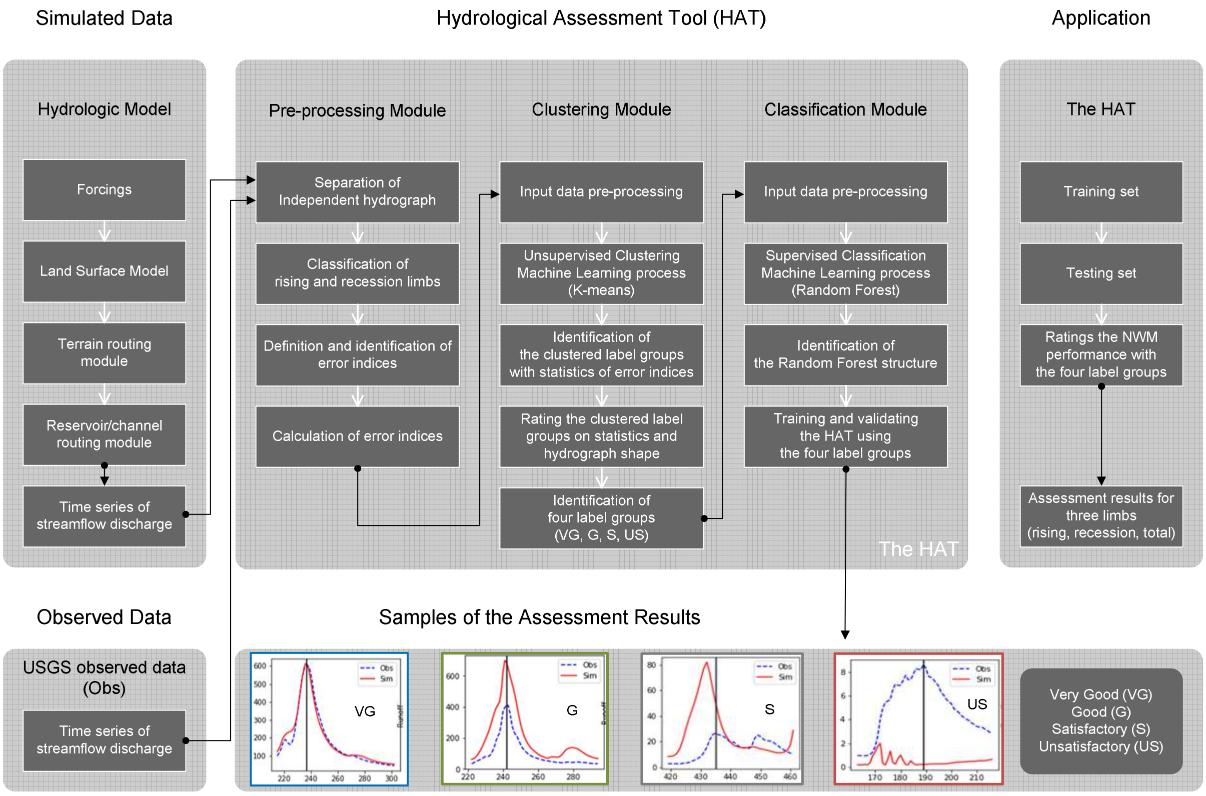

Simulated Data
Hydrological Assessment Tool (HAT)
Application
Hydrologic Model
Forcings
Land Surface Model
Terrain routing module
Reservoir/channel routing module
Time series of streamflow discharge
Pre-processing Module
Separation of Independent hydrograph
Classification of rising and recession limbs
Definition and identification of error indices
Calculation of error indices
Clustering Module
Input data pre-processing
Unsupervised Clustering Machine Learning process (K-means)
Identification of the clustered label groups with statistics of error indices
Rating the clustered label groups on statistics and hydrograph shape
Identification of four label groups (VG, G, S, US)
Classification Module
Input data pre-processing
Supervised Classification Machine Learning process (Random Forest)
Identification of the Random Forest structure
Training and validating the HAT using the four label groups
The HAT
The HAT
Training set
Testing set
Ratings the NWM performance with the four label groups
Assessment results for three limbs (rising, recession, total)
Observed Data
USGS observed data (Obs)
Time series of streamflow discharge
Samples of the Assessment Results
VG
G
S
US
Very Good (VG)
Good (G)
Satisfactory (S)
Unsatisfactory (US)

Figure 5

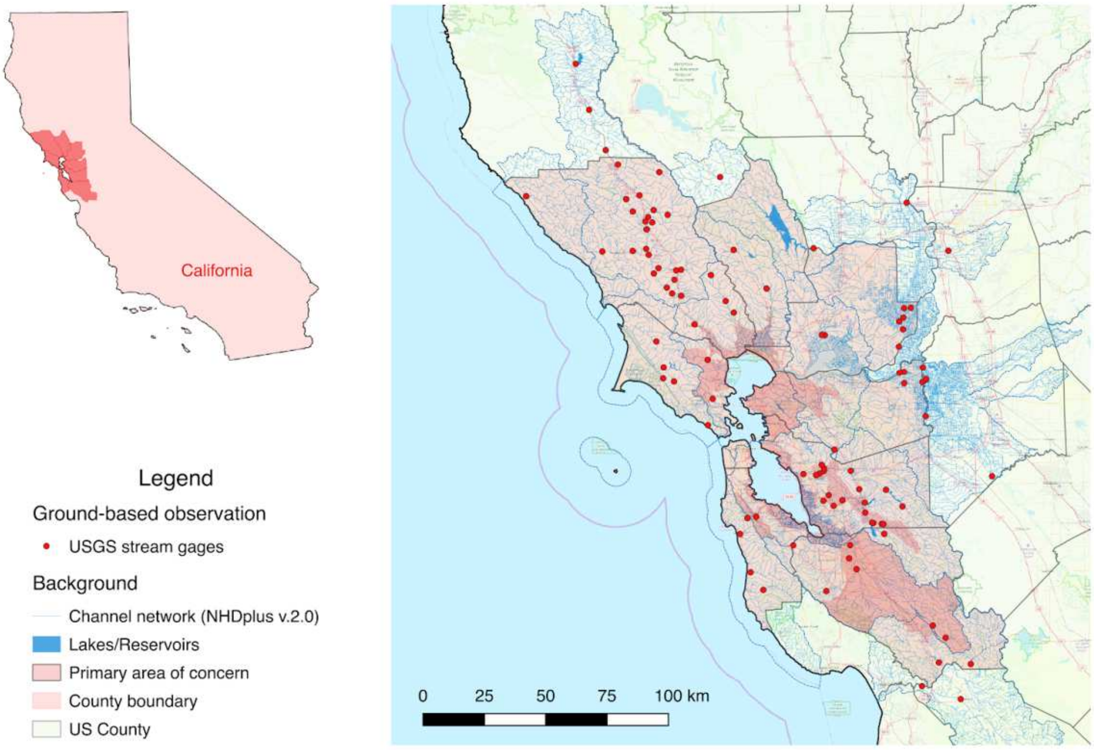

California
Legend
Ground-based observation
USGS stream gages
Background
Channel network (NHDplus v.2.0)
Lakes/Reservoirs
Primary area of concern
County boundary
US County
0 25 50 75 100 km

Figure 6

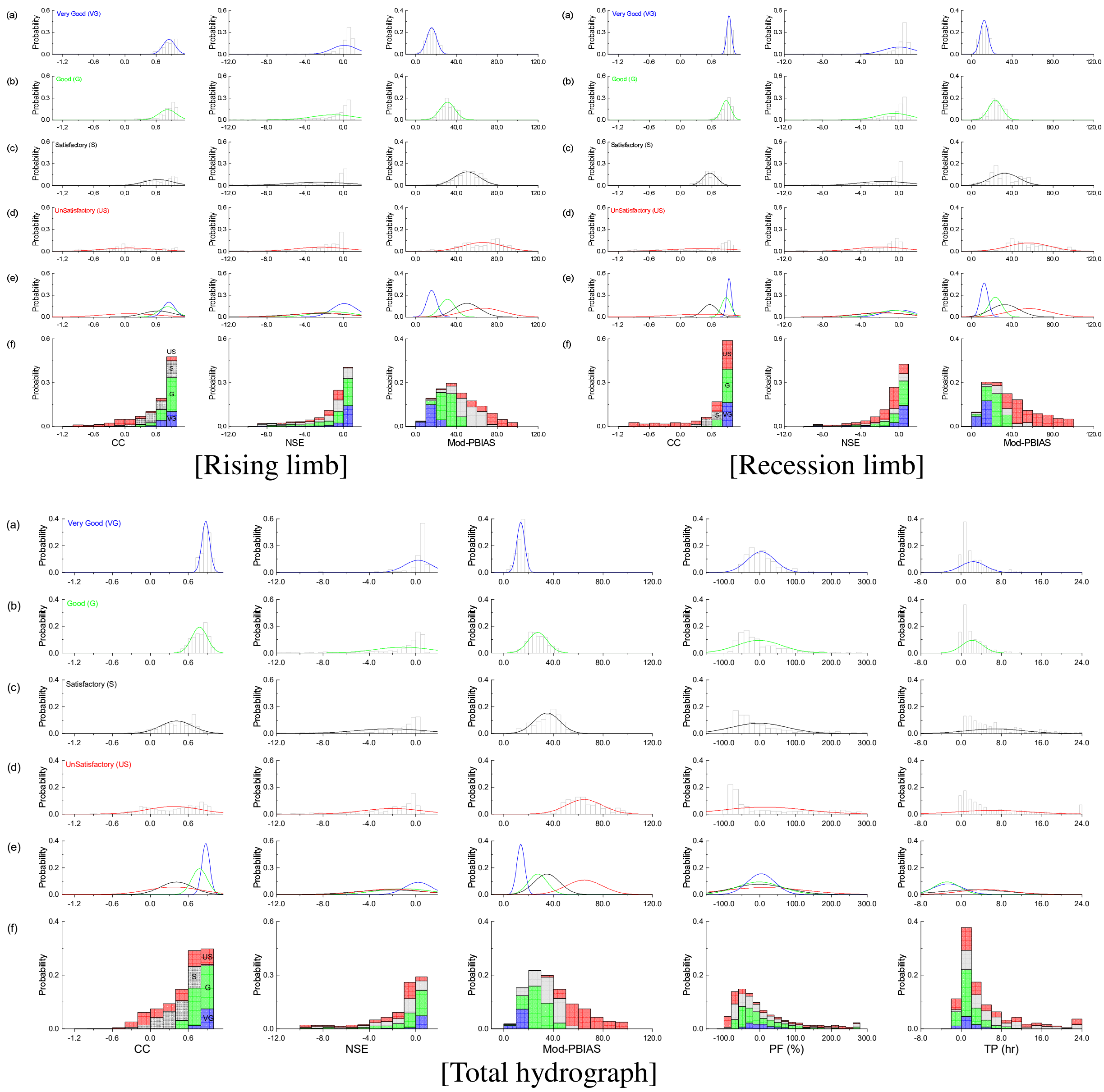
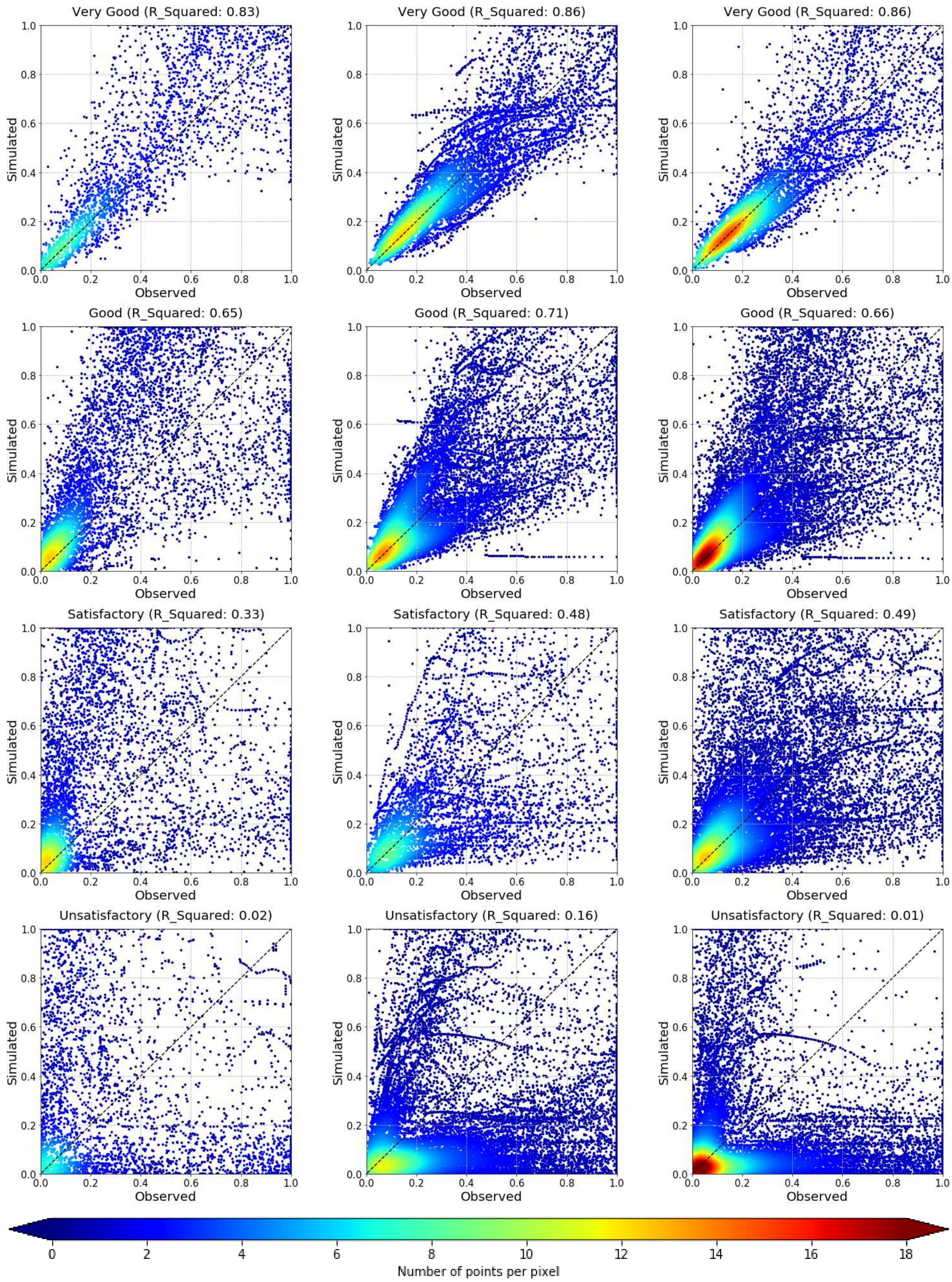
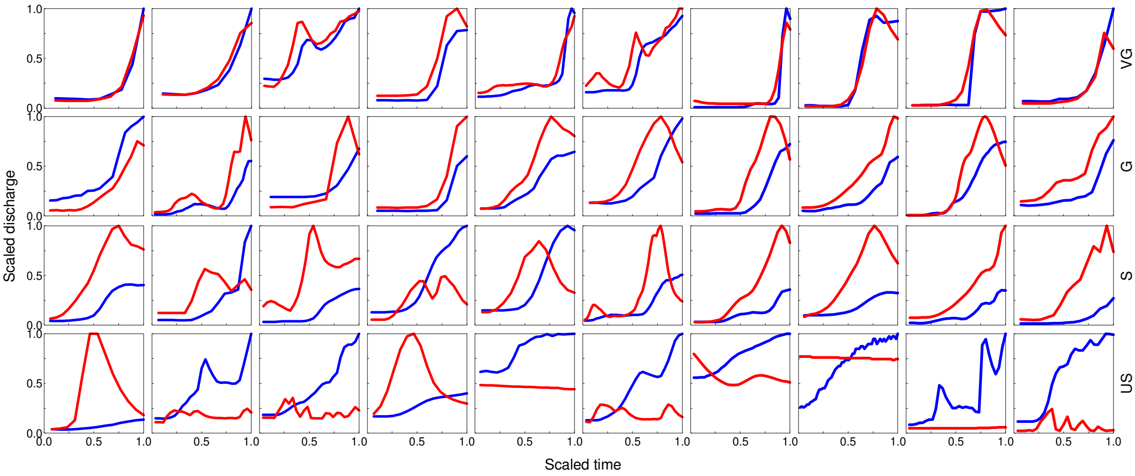

Very Good (VG)
Good (G)
Satisfactory (S)
UnSatisfactory (US)
Probability
CC
NSE
Mod-PBIAS
PF (%)
TP (hr)
US
S
G
VG
[Rising limb]
[Recession limb]
[Total hydrograph]

Figure 7

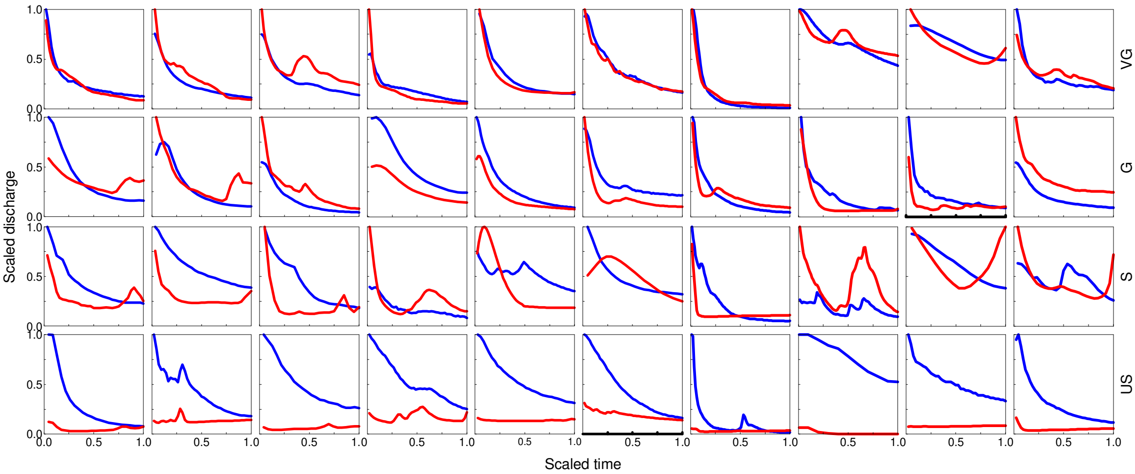


(a) Rising limb (b) Recession limb (c) Total hydrograph

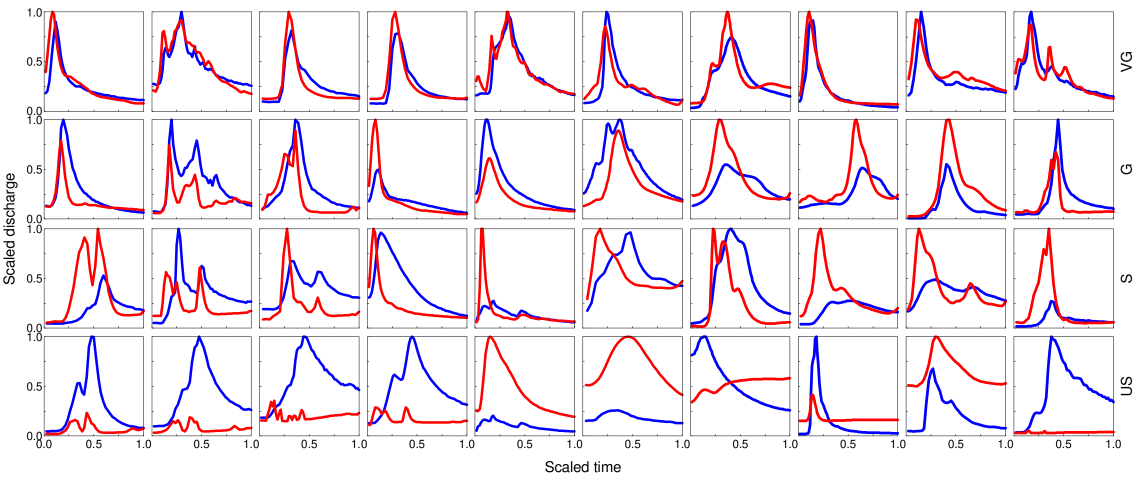
Figure 8



(a) Rising limb

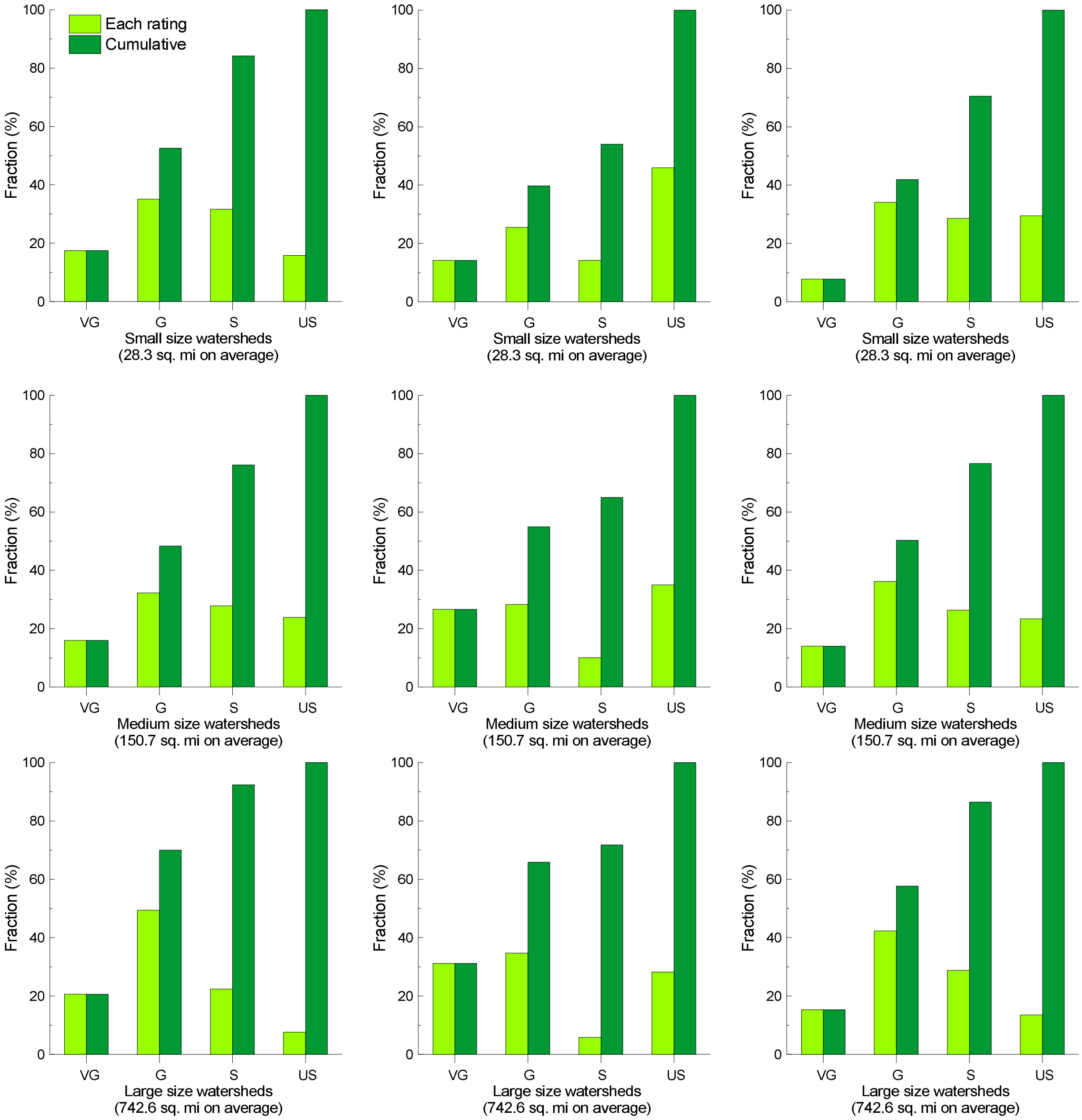


(b) Recession limb



(c) Total hydrograph

Figure 9



(a) Rising limb (b) Recession limb (c) Total hydrograph

Figure 10

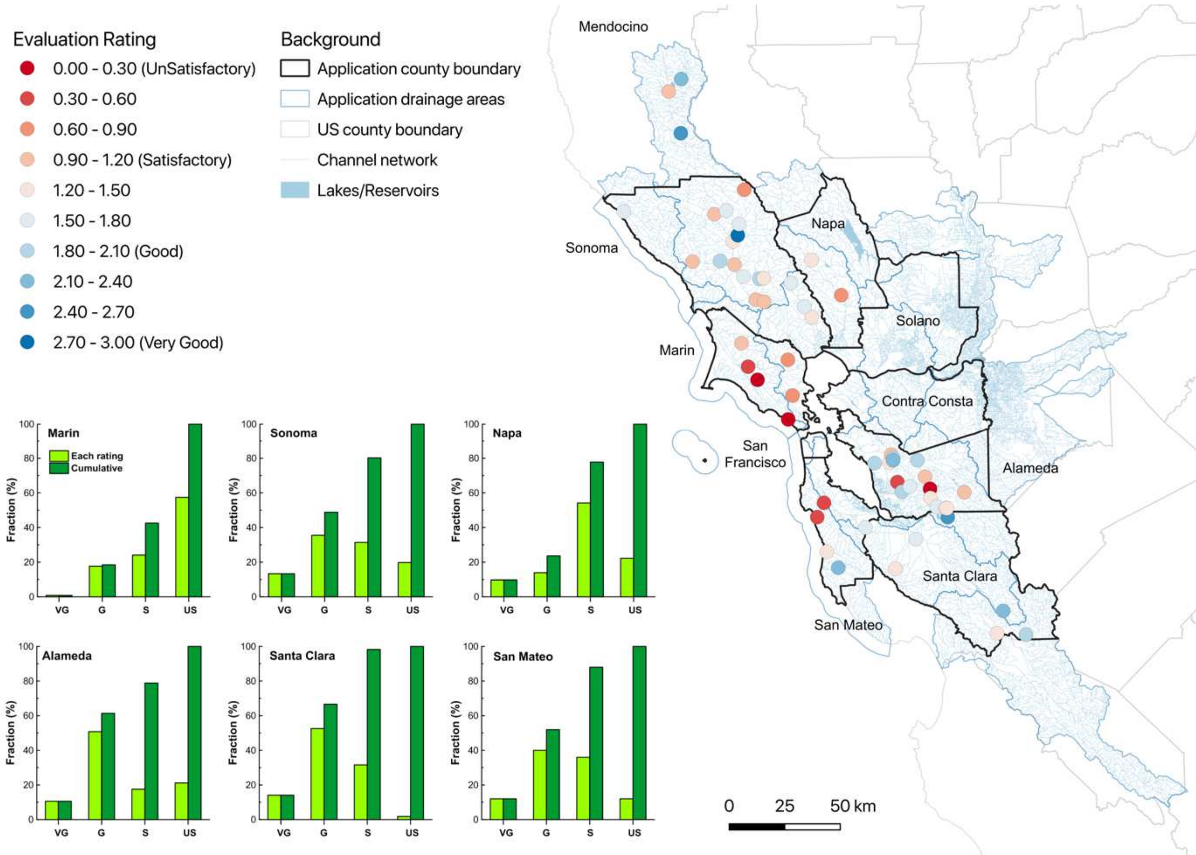
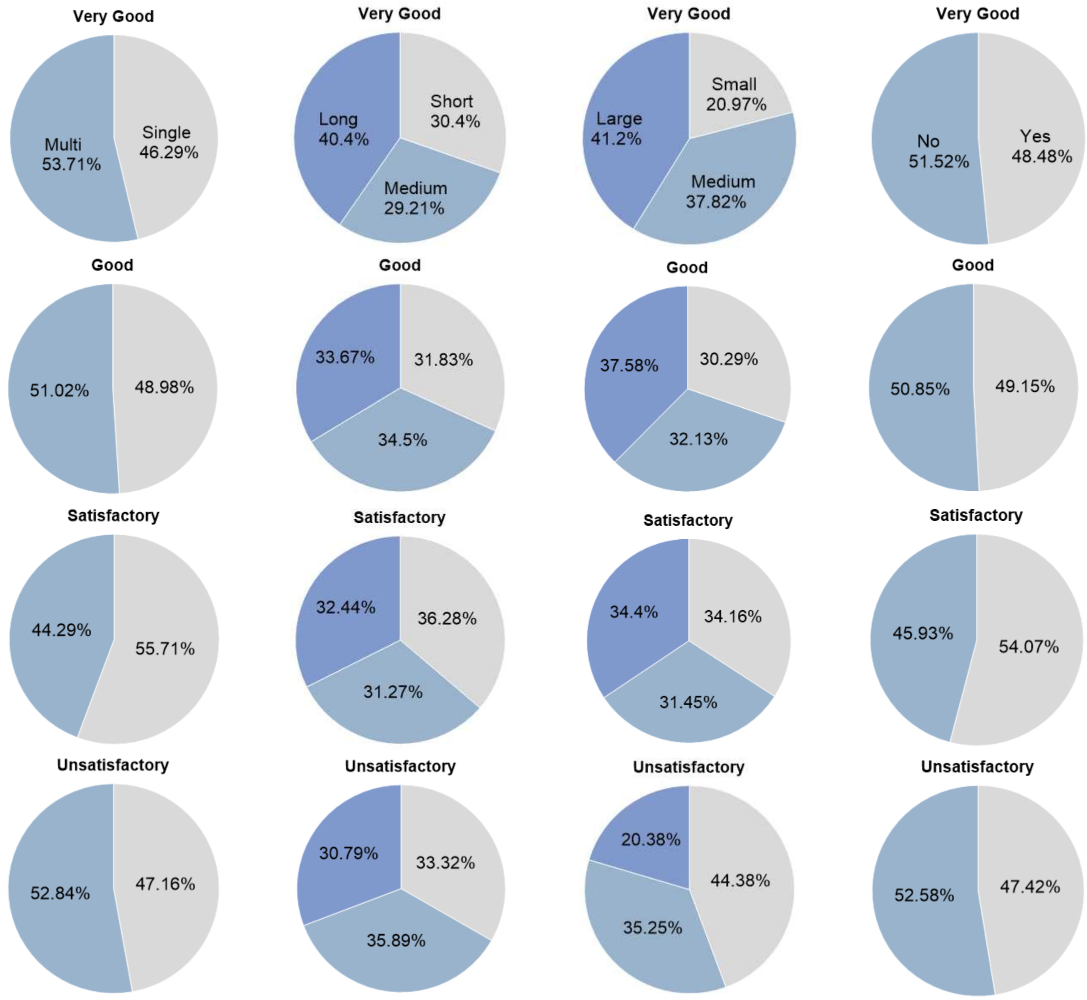
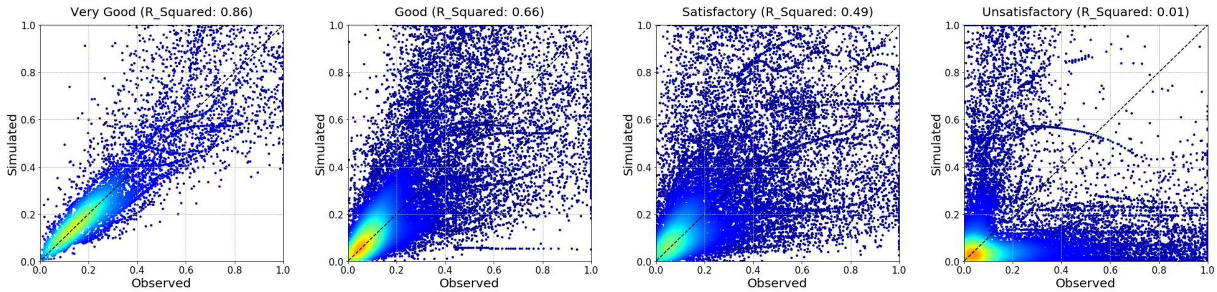

Evaluation Rating
0.00 - 0.30 (UnSatisfactory)
0.30 - 0.60
0.60 - 0.90
0.90 - 1.20 (Satisfactory)
1.20 - 1.50
1.50 - 1.80
1.80 - 2.10 (Good)
2.10 - 2.40
2.40 - 2.70
2.70 - 3.00 (Very Good)
Background
Application county boundary
Application drainage areas
US county boundary
Channel network
Lakes/Reservoirs
Mendocino
Sonoma
Napa
Solano
Marin
Contra Consta
San Francisco
Alameda
Santa Clara
San Mateo
0 25 50 km
Marin
Each rating
Cumulative
Sonoma
Napa
Alameda
Santa Clara
San Mateo
Fraction (%)
VG G S US

Figure 11

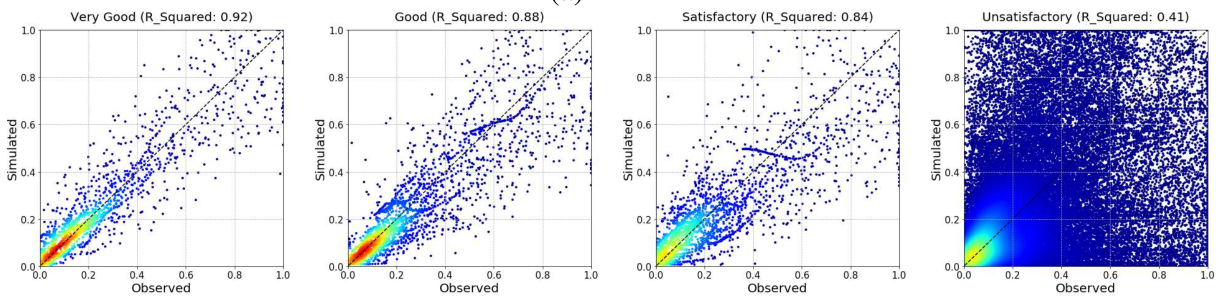


(a) Hydrograph (b) Runoff duration (c) Drainage size (d) Regulation

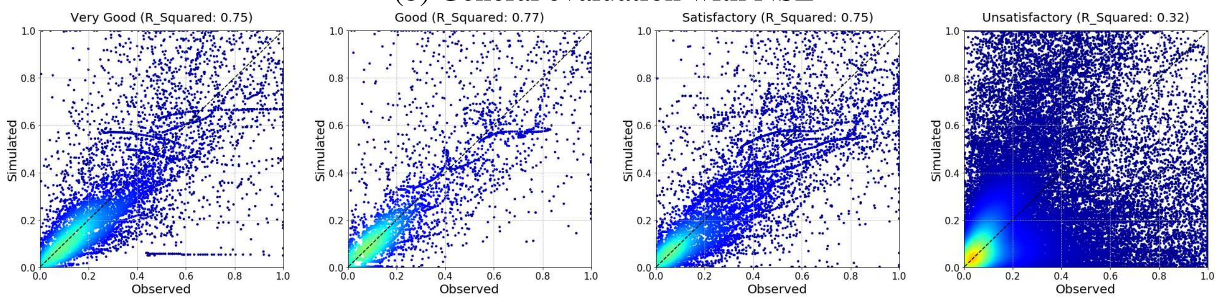
Figure 12



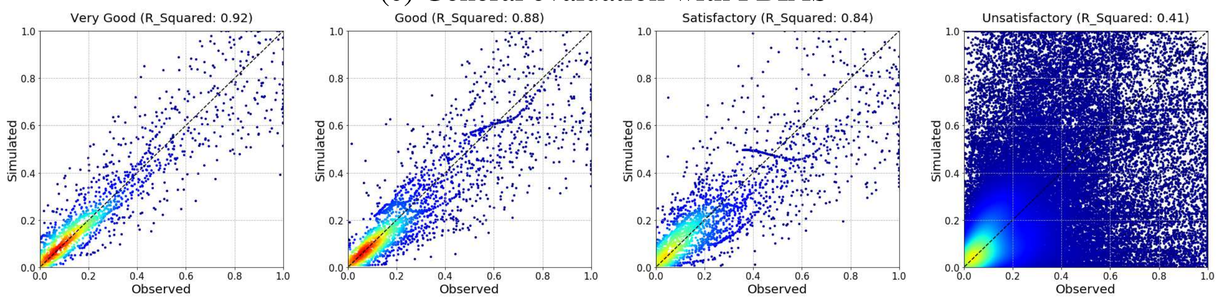
(a) HAT



(b) General evaluation with NSE



(c) General evaluation with PBIAS



(d) General evaluation with RSR

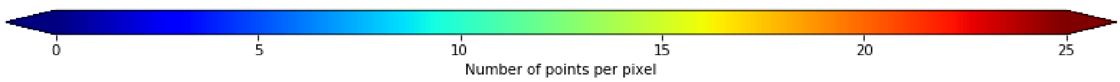


Figure 13

List of tables

Table 1

Error Indices	Acronym (Range)	Equation
Correlation coefficient	CC [-1, 1]	$\frac{\sum(Q_{sim}-\overline{Q_{sim}})(Q_{obs}-\overline{Q_{obs}})}{\sqrt{\sum(Q_{sim}-\overline{Q_{sim}})^2}\sqrt{\sum(Q_{obs}-\overline{Q_{obs}})^2}}$
Nash-Sutcliffe efficiency	NSE (-inf, 1]	$1-\frac{\sum(Q_{sim}-Q_{obs})^2}{\sum(Q_{obs}-\overline{Q_{obs}})^2}$
Percent bias	PBIAS (-inf, inf)	$\left(\sum(Q_{obs}-Q_{sim})\right) \div \sum Q_{obs} \times 100\ (\%)$
RMSE-observations standard deviation ratio	RSR [0, inf)	$\frac{\sqrt{\sum(Q_{obs}-Q_{sim})^2}}{\sqrt{\sum(Q_{obs}-\overline{Q_{obs}})^2}}$
Time to peak error	TP (-inf, inf)	T_{obs}- T_{sim}
Peak flow error	PF (-inf, inf)	(Max(Q_{obs}) – Max(Q_{sim})) ÷ Max(Q_{obs}) × 100 (%)

Table 2

Rating	Statistic	Error index		
		CC	NSE	Mod PBIAS (MPBIAS)
Very good	min≤ , ≤max	0.74≤CC≤1.00	-8.16≤NSE≤1.00	0.00≤MPBIAS≤18.50
	Q1[a]≤ , ≤Q3[a]	0.84≤CC≤0.92	0.25≤NSE≤0.72	11.31≤MPBIAS≤15.74
	mean (variance)	0.88 (0.004)	0.22 (1.342)	13.63 (10.148)
Good	min≤ , ≤max	0.44≤CC≤0.98	-54.34≤NSE≤0.87	5.64≤MPBIAS≤45.44
	Q1[a]≤ , ≤Q3[a]	0.68≤CC≤0.88	-1.79≤NSE≤0.34	21.81≤MPBIAS≤32.89
	mean (variance)	0.78 (0.015)	-2.23 (33.984)	27.36 (60.616)
Satisfactory	min≤ , ≤max	-0.41≤CC≤0.89	-165.40≤NSE≤0.72	8.63≤MPBIAS≤55.72
	Q1[a]≤ , ≤Q3[a]	0.24≤CC≤0.65	-5.62≤NSE≤-0.13	27.69≤MPBIAS≤42.26
	mean (variance)	0.41 (0.065)	-5.13 (161.571)	34.78 (108.115)
Unsatisfactory	min≤ , ≤max	-0.92≤CC≤0.98	-534.44≤NSE≤0.60	25.36≤MPBIAS≤99.38
	Q1[a]≤ , ≤Q3[a]	0.01≤CC≤0.76	-13.22≤NSE≤-0.44	53.92≤MPBIAS≤74.89
	mean (variance)	0.37 (0.176)	-21.74 (3713.448)	64.89 (214.739)

[a] Q1 and Q3 indicate the lower (25%) and upper (75%) quartiles.

Table 3

Ratings	Hydrograph						(d) Entire[a]	
	(a) Rising[a]		(b) Recession[a]		(c) Total[a]			
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
VG	96.8	3.2	100.0	0.0	100.0	0.0	98.8	1.2
G	97.7	2.3	96.7	3.3	96.8	3.2	97.1	2.9
S	98.4	1.6	100.0	0.0	92.8	7.2	96.1	3.9
US	97.8	2.2	100.0	0.0	100.0	0.0	99.6	0.4
Mean	97.7	2.3	99.2	0.8	97.4	2.6	97.9	2.1

[a] (a)-(c) represent each limb and total hydrograph results, and (d) indicates correct and incorre

ct percentages for entire results regardless limbs.

Table 4

Error Index	Hydrograph		
	(a) Rising	(b) Recession	(c) Total
Mod PBIAS	0.55	0.56	0.52
CC	0.30	0.29	0.26
NSE	0.15	0.15	0.07
TP	-	-	0.08
PF	-	-	0.07
Sum	1.00	1.00	1.00

Table 5

Set	Incorrect (%)			Overrated (%)			Underrated (%)		
	Rising	Recession	Total	Rising	Recession	Total	Rising	Recession	Total
1	5.7	3.6	6.4	3.6	2.1	2.1	2.1	1.4	4.3
2	1.4	0.7	3.6	0.0	0.0	2.1	1.4	0.7	1.4
3	1.4	2.9	5.0	1.4	1.4	2.9	0.0	1.4	2.1
4	2.1	2.1	0.7	0.7	0.7	0.0	1.4	1.4	0.7
5	0.7	3.6	6.4	0.7	0.7	2.9	0.0	2.9	3.6
6	2.9	0.7	1.4	0.7	0.7	1.4	2.1	0.0	0.0
7	0.7	1.4	2.1	0.0	0.7	0.0	0.7	0.7	2.1
8	1.4	2.1	7.1	0.7	0.0	5.0	0.7	2.1	2.1
9	3.6	1.4	5.7	1.4	1.4	2.1	2.1	0.0	3.6
10	4.0	0.7	5.3	2.6	0.0	1.3	1.3	0.7	4.0
Mean	2.4	1.9	4.4	1.2	0.8	2.0	1.2	1.1	2.4

Table 6

Rating	Method	Statistic	Error index		
			RSR	NSE	PBIAS
Very good	HAT	min≤ , ≤max	0.00≤RSR≤3.03	-8.16≤NSE≤1.00	-38.33≤PBIAS≤28.49
		Q1≤ , ≤Q3	0.53≤RSR≤0.86	0.25≤NSE≤0.72	-18.70≤PBIAS≤10.52
		mean (variance)	0.78 (0.181)	0.22 (1.342)	-2.99 (333.921)
	General[a]	min≤ , ≤max	0.00≤RSR≤0.50	0.75<NSE≤1.00	PBIAS≤±10
Good	HAT	min≤ , ≤max	0.36≤RSR≤7.44	-54.34≤NSE≤0.87	-166.54≤PBIAS≤62.48
		Q1≤ , ≤Q3	0.81≤RSR≤1.37	-1.79≤NSE≤0.34	-56.17≤PBIAS≤23.39
		mean (variance)	1.46 (1.112)	-2.23 (36.984)	-22.52 (2726.906)
	General[a]	min≤ , ≤max	0.50<RSR≤0.60	0.65<NSE≤0.75	±10<PBIAS<±15
Satisfactory	HAT	min≤ , ≤max	0.53≤RSR≤12.90	-165.40≤NSE≤0.72	-189.11≤PBIAS≤71.17
		Q1≤ , ≤Q3	1.06≤RSR≤2.57	-5.62≤NSE≤-0.13	-55.33≤PBIAS≤32.84
		mean (variance)	2.01 (2.083)	-5.13 (161.571)	-19.25 (4010.565)
	General[a]	min≤ , ≤max	0.60<RSR≤0.70	0.50<NSE≤0.65	±15<PBIAS<±25
Unsatisfactory	HAT	min≤ , ≤max	0.63≤RSR≤23.14	-534.44≤NSE≤0.60	-1409.10≤PBIAS≤99.52
		Q1≤ , ≤Q3	1.20≤RSR≤3.77	-13.22≤NSE≤-0.44	-154.89≤PBIAS≤80.40
		mean (variance)	3.20 (12.469)	-21.74 (3713.448)	-43.37 (40543.580)
	General[a]	min≤ , ≤max	RSR>0.70	NSE≤0.50	PBIAS≥±25

[a] general evaluation approach by Moriasi et al. (2007).