AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016805

The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(ABC)}$

Mojtaba Sadegh[1], Jasper A. Vrugt[1,2], Chonggang Xu[3], and Elena Volpi[4]

[1]Department of Civil and Environmental Engineering, University of California, Irvine, California, USA, [2]Department of Earth System Science, University of California, Irvine, California, USA, [3]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA, [4]Department of Engineering, University of Roma Tre, Rome, Italy

Key Points:
- Stationarity paradigm is revisited using diagnostic model evaluation with DREAM$_{(ABC)}$
- Nonstationarity is not readily apparent from statistical analysis
- Nonstationarity is evident from analysis of summary metrics

Correspondence to:
J. A. Vrugt,
jasper@uci.edu

Citation:
Sadegh, M., J. A. Vrugt, C. Xu, and E. Volpi (2015), The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(ABC)}$, *Water Resour. Res.*, *51*, 9207–9231, doi:10.1002/2014WR016805.

Received 19 DEC 2014
Accepted 9 SEP 2015
Accepted article online 21 SEP 2015
Published online 28 NOV 2015

Abstract Many watershed models used within the hydrologic research community assume (by default) stationary conditions, that is, the key watershed properties that control water flow are considered to be time invariant. This assumption is rather convenient and pragmatic and opens up the wide arsenal of (multivariate) statistical and nonlinear optimization methods for inference of the (temporally fixed) model parameters. Several contributions to the hydrologic literature have brought into question the continued usefulness of this stationary paradigm for hydrologic modeling. This paper builds on the likelihood-free diagnostics approach of Vrugt and Sadegh (2013) and uses a diverse set of hydrologic summary metrics to test the stationary hypothesis and detect changes in the watersheds response to hydroclimatic forcing. Models with fixed parameter values cannot simulate adequately temporal variations in the summary statistics of the observed catchment data, and consequently, the DREAM$_{(ABC)}$ algorithm cannot find solutions that sufficiently honor the observed metrics. We demonstrate that the presented methodology is able to differentiate successfully between watersheds that are classified as stationary and those that have undergone significant changes in land use, urbanization, and/or hydroclimatic conditions, and thus are deemed nonstationary.

1. Introduction

The flow of water through watersheds is an incredibly complex process controlled by physical characteristics of the basin and a myriad of highly interrelated, spatially distributed, water, energy, and vegetation processes. As available measurements lack the resolution and information content required to warrant a detailed characterization of watershed structure, properties, and processes, relatively simple models are used to describe (among others) soil moisture flow, groundwater recharge, surface runoff, preferential flow, root water uptake, and river discharge at different spatial and temporal scales. This includes prediction in space (interpolation/extrapolation) and prediction in time (forecasting). These models describe spatially distributed vegetation and subsurface properties with much simpler homogeneous units using transfer functions that describe the flow of water within and between different storage compartments.

Many watershed models used within the hydrologic research community assume (by default) stationary conditions, that is, the key watershed properties that control water flow are considered to be time invariant. As a consequence, the watershed behavior as measured in hydrologic states and fluxes (jointly called variables) is assumed to vary around some constant mean value with fixed variance and serial correlation structure [*Clarke*, 2007]. This assumption is rather convenient and opens up the wide arsenal of (multivariate) statistical and nonlinear optimization methods for inference of the (temporally fixed) model parameters. Notwithstanding the progress made, several contributions to the hydrologic literature have brought into question the continued usefulness of this stationary paradigm for hydrologic modeling [*Westmacott and Burn*, 1997; *Karl and Knight*, 1998; *Strupczewski et al.*, 2001; *McCabe and Wolock*, 2002; *Groisman et al.*, 2004; *Fu et al.*, 2004; *Lins and Slack*, 2005; *Svensson et al.*, 2005; *Alexander et al.*, 2006; *Hodgkins and Dudley*, 2006; *Xu et al.*, 2006; *Leclerc and Ouarda*, 2007; *Milly et al.*, 2008; *Villarini et al.*, 2009; *Kundzewicz*, 2011; *Stedinger and Griffis*, 2011; *Vogel et al.*, 2011; *Waage and Kaatz*, 2011; *Ishak et al.*, 2013; *Salas and Obeysekera*, 2014]. For example, *Strupczewski et al.* [2001] showed evidence of nonstationarity in the annual maximum flows of 39 Polish rivers during the period of 1921–1990. *Villarini et al.* [2009] examined annual peak discharges from

50 different stations in the continental United States, and demonstrated that almost half of the records exhibit statistically significant changes in the mean and variance of annual flood peaks. Indeed, *Vogel et al*. [2011] report flood magnification factors in excess of 2–5 for many regions of the United States, particularly those regions with higher population densities. Another study by *Ishak et al*. [2013] for Australian catchments showed a significant downward trend in the annual maximum streamflow values. Moreover, *Leclerc and Ouarda* [2007] demonstrated significant improvements in the flood quantile estimates of ungaged sites in the southeastern part of Canada and northeastern part of the United States when using nonstationary regional flood frequency analysis.

Suppose that we have available a multiyear record of data, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$, of some hydroclimatologic variable, η, we would like to test this n year record for stationarity. Under H_0, the null hypothesis, it is typically assumed that the n year observations of η are homogeneous and thus have the same mean. The alternative hypothesis, H_1, assumes η to exhibit inhomogeneity, but is often rather vague because of a lack of knowledge about the expected trend in the $\tilde{y}$'s. Many different statistical tests can be devised to test the null hypothesis, homogeneity of $\tilde{\mathbf{Y}}$. Examples include the von Neumann ratio, the Kendall, Mann-Kendall, Spearman, Pearson, and Pettit tests [*Buishand*, 1982; *Lins and Slack*, 1999; *Douglas et al*., 2000; *Yue et al*., 2002; *Zhang et al*., 2009; *Bassiouni and Oki*, 2013; *Rougé et al*., 2013; *Westra et al*., 2013]. Other approaches include spectral [*Ramachandra Rao and Yu*, 1986; *Joshi and Pandey*, 2011; *Ni et al*., 2011], moving average [*Ramachandra Rao and Yu*, 1986; *Anghileri et al*., 2014], wavelet [*Karthikeyan and Nagesh Kumar*, 2013], (flood) frequency and risk analysis [*Cunnane*, 1988; *Jain and Lall*, 2001; *Cunderlik and Burn*, 2003; *Cunderlik and Ouarda*, 2006; *Leclerc and Ouarda*, 2007; *Dettinger*, 2011; *Stedinger and Griffis*, 2011; *Chebana et al*., 2013], and the use of generalized extreme value (GEV) and max-stable models [*Clarke*, 2002; *El-Adlouni et al*., 2007; *Westra and Sisson*, 2011; *Nasri et al*., 2013; *Westra et al*., 2013].

The work of *Lins* [1985] is the first study on long-term analysis of hydrologic trends. This analysis showed that broad, national patterns in streamflow are fairly closely related to annual (continental U.S. average) precipitation anomalies, especially during the post 1950 period for which the most extensive stream gage network exists. *Lettenmaier et al*. [1994] analyzed spatial patterns in trends of hydroclimatological variables and concluded that the observed trends in streamflow are not entirely consistent with the changes in temperature and precipitation and may be due to a combination of climatic and water management effects. Many later studies have focused on detection and interpretation of temporal changes in river flow regimes [*Zhang et al*., 2001; *Burn and Hag Elnur*, 2002; *Birsan et al*., 2005; *Novotny and Stefan*, 2007; *Petrow and Merz*, 2009; *Ishak et al*., 2013], among many others. These studies employ a range of parametric and nonparametric approaches to detect temporal changes in hydrologic time series. *Khaliq et al*. [2009] presented a comprehensive review of the methodologies adopted for the identification of hydrologic trends, while *Kundzewicz and Robson* [2004] provided general guidelines for the detection of trends in hydrologic data.

In a separate line of research, *Renard et al*. [2006] proposed a Bayesian framework to explicitly treat the uncertainty associated with the stationarity hypothesis. Several probabilistic models (stationary, step change, and linear trend) and extreme values distributions were tested for a 93 year record of annual peak discharges of the Drôme River in France. A similar stochastic modeling framework was advocated by *North* [1980] and used recently by *Westerberg et al*. [2011] to assess the uncertainty of the stage-discharge relationship arising from a nonstationary rating curve. *Lima and Lall* [2010] analyzed trends of annual and monthly peak streamflows using hierarchial Bayesian models and spatial scaling. Moreover, *Ouarda and El-Adlouni* [2011] introduced a maximum likelihood method to estimate the time-varying parameters of the generalized extreme value (GEV) distribution for hydrologic frequency analysis, and *Nasri et al*. [2013] introduced a GEV model with B-spline functions to estimate the quantiles of synthetic and observed rainfall data.

The term nonstationarity is a convenient way to describe characteristics of hydrologic variables that do not meet the "constancy" criteria referred to above. Yet this term fails to distinguish between long-term fluctuations (due to climate) and more gradual/abrupt changes due to human intervention (such as land use alterations and flow regulation through reservoirs and dams). For instance, the downward trend in the observed annual flood peaks observed by *Ishak et al*. [2013] is explained in large part by climate variability and likely not due to changes in physical characteristics of the watershed. Sudden multiyear changes in catchment behavior (single or multipoint) are often of anthropogenic origin, whereas decadal variations in streamflow response can be explained by climate trends and/or variability [*Graf*, 1977; *Pitman*, 1978; *Potter*, 1991; *Smith et al*., 2002; *Zhang and Schilling*, 2006; *Milly et al*., 2008; *Vaze et al*., 2010; *Villarini and Smith*, 2010]. Indeed, the extent to which changes in the

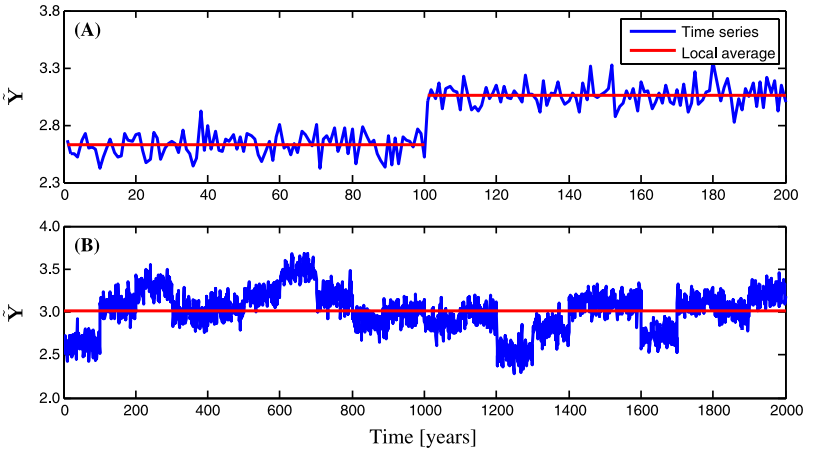

Figure 1. Clarification of the notion of stationary and nonstationary behavior using a rather simple univariate stochastic process lasting 2000 years (adapted from *Koutsoyiannis* [2011]). (top) The first 200 years of data on record, and (bottom) the full 2000 year record is displayed. The synthetic data are represented with a blue line, whereas the red line signifies the mean 100 year system response. Visual analysis demonstrates a sudden shift in system response after exactly 100 years of observation. This abrupt change of the system response would seem a textbook example of nonstationary, yet when the behavior of the system is viewed over a much larger, 2000 year data period, the mean and standard deviation of $\tilde{\mathbf{Y}}$ appear rather constant, and thus stationarity is observed. One should therefore be particularly careful proclaiming nonstationarity on the basis of a rather short data record. What is more, visual analysis and or statistical tests can provide misleading results—in part because these methods have a poor correspondence with underlying hydrologic processes.

watersheds response to rainfall are caused by hydroclimatic variations or by anthropogenic changes to the catchment characteristics (urbanization, agriculture, irrigation, dams, deforestation, and afforestation) is difficult to assess, and poses great challenges for hydrologists.

Although trends can be readily apparent in time series of hydrologic variables, one should be particularly careful not to erroneously classify this behavior as nonstationarity. For instance, short decadal variations in climate can induce trends in hydrologic variables [*Clarke*, 2002; *Parey et al*., 2007; *Ishak et al*., 2013; *Anghileri et al*., 2014], but this apparent nonstationarity might be natural variability when the behavior of the same variables is investigated over much larger time periods [*Kundzewicz et al*., 2005; *Koutsoyiannis*, 2006, 2011; *Lins and Cohn*, 2011; *Koutsoyiannis*, 2013], also referred to as long-term persistency [*Hurst*, 1951]. These long-term decadal fluctuations have been studied in detail by *Hurst* [1951], using discharge data from the Nile River in Egypt. Short-term trends of the streamflow data (up to about 100 years) were explained by long-term persistency of the Nile River system, rather than nonstationarity.

Without a sufficiently long data record, it is particularly difficult to determine whether temporal changes in watershed behavior are due to natural fluctuations in the weather and climate, or whether the physical characteristics of the watershed itself have experienced changes, the latter which we certainly classify as nonstationarity. To clarify the notion of stationarity and its antithesis nonstationarity, we draw inspiration from *Koutsoyiannis* [2011] and create a system response that is made up of 20 stochastic processes each lasting a 100 years and with mean value drawn from $\mathcal{N}(a, b)$, a normal distribution with mean $a = 3$ and standard deviation $b = 0.2$. The 2000 year data set is then perturbed with Gaussian noise using $b = 0.1$, and the resulting observations plotted in Figure 1. If we interpret the data in the top plot, then we observe a sudden shift in system response at 100 years from one mode with mean of about 2.70 to another mode with mean of approximately 3.15. At first glance, it is tempting to classify this rather dramatic shift in system response as evidence of nonstationarity. Yet when the behavior of this system is viewed over a much longer 2000 year record (bottom plot) then homogeneity is observed. Hence, one should be particularly careful in proclaiming nonstationarity based on a relatively short data time period.

Many statistical tests (listed above) have been used in the hydrologic literature to address the stationarity null hypothesis. Without further investigation into the underlying hydroclimatologic processes, such approaches can lead to spurious conclusions [*Koutsoyiannis*, 2006, 2011]. For instance, consider Figure 2 which plots (top plot) the cumulative streamflow of the Blackberry Creek basin at Yorkville, IL for the period of 2002–2012. The dashed lines depict the slope of the cumulative streamflow for the periods of 2002–2007 (black) and 2007–2012 (grey), respectively. There appears to be a sudden increase in the watersheds streamflow response in 2007. It can be tempting to proclaim this behavior as nonstationarity (as suggested

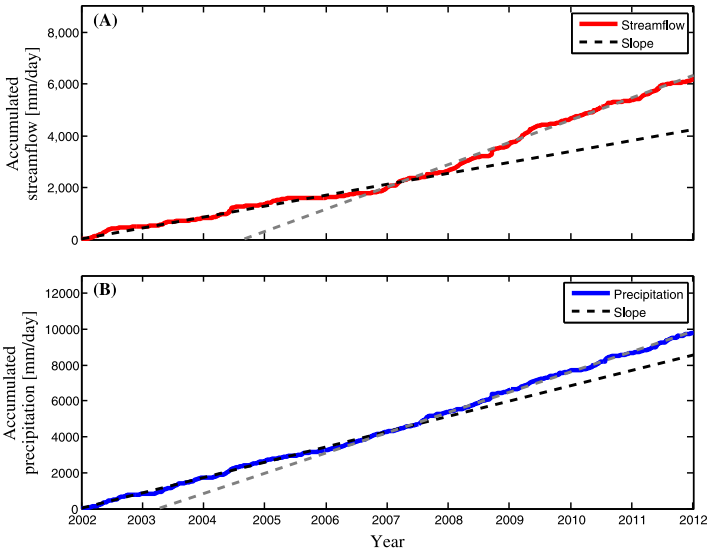

Figure 2. (a) Cumulative streamflow and (b) precipitation data record of the Blackberry Creek in Illinois, USA, for the years 2002–2012. The slope of the cumulative streamflow curve increases noticeably between the years 2007 and 2008. In the hydrologic community, this behavior is classified as nonstarionary due to growing urbanization (http://non-stationarities.irstea.fr/). Yet the bottom graph demonstrates noticeably larger precipitation amounts from 2007 onward which coincides perfectly with the increase in streamflow. With the help of diagnostic model evaluation and temporal analysis of the hydrologic signatures, we classify this watershed as stationary. A single model structure and fixed parameterization is sufficient to model the rainfall-runoff response.

by http://non-stationarities.irstea.fr/). However, the bottom plot demonstrates that the increase in streamflow from 2007 onward coincides exactly with an increase in rainfall. Thus, one should be particularly careful to proclaim nonstationarity based on the observed discharge data only. Our diagnostic analysis presented in section 4 of this paper indeed confirms that the behavior of the Blackberry watershed is considered stationary during the 2002–2012 observational period.

The results of Figure 2 reinforce the need for a detailed look into the underlying processes that drive the watershed response to hydroclimatic forcing. For instance, *Groisman et al.* [2001] found a significant relationship between the frequency of heavy precipitation in the eastern half of the United States and the presence of high streamflow events both annually and during the months of maximum streamflow. Using data from the Yellow River in China, *Cong et al.* [2009] demonstrated that streamflow trends were explained by large-scale decadal variations in climate forcing whereas land use changes were found to have a negligible impact.

In this paper, we introduce the elements of a process-based framework to address the stationarity hypothesis. Our methodology builds on recommendations of *Koutsoyiannis* [2006] and *Koutsoyiannis* [2011] and merges numerical modeling with the likelihood-free diagnostics approach of *Vrugt and Sadegh* [2013] and a set of processed-based summary metrics to pinpoint gradual/abrupt changes in watershed behavior. Our approach builds on the hypothesis (assumption) that nonstationarity catchment behavior should be evident from analysis of the temporal patterns of the watershed signatures as measured, for instance, by summary metrics of the discharge data. Such time invariant metrics cannot be properly simulated with a fixed model structure and temporally invariant parameter values. As a consequence, diagnostic inference with approximate Bayesian computation (ABC) cannot provide behavioral simulations that honor the observed summary metrics. On the contrary, for a stationary watershed, one would expect the summary metrics not to vary much temporally, and hence, a single parameterization would deem sufficient to simulate adequately the observed, time invariant, rainfall-runoff transformation. Results presented in this paper confirm our assumption—the presented methodology is able to differentiate successfully between watersheds that are classified as stationary in the literature and those that have undergone significant changes in land use, urbanization, and/or hydroclimatic conditions (forcing), and thus are deemed nonstationary.

The remainder of this paper is organized as follows. Section 2 clarifies the main terminology used in this paper. In section 3, we shortly describe the watershed data used herein. These basins have been selected after careful literature review and include stationary and nonstationary watersheds. Then in section 4, we review the basic elements of the likelihood-free diagnostics methodology of *Vrugt and Sadegh* [2013], which together with the

DREAM$_{(ABC)}$ algorithm [*Sadegh and Vrugt*, 2014] and a set of hydrologic summary metrics of the observed discharge data is used to address the stationarity null hypothesis. Here we are particularly concerned with the choice the summary metrics and value of the cutoff threshold, ϵ, that can be used to accept/reject the stationarity null hypothesis. Section 5 presents the results of our diagnostic analysis for each of the watersheds considered herein. Finally, section 6 concludes this paper with a summary of our most important findings.

2. Overview and Terminology

Before we discuss the terminology used in this paper, we first present in Figure 3 a schematic overview of our stationarity hypothesis testing methodology. The different rectangles signify the different building blocks of our methodology. The flowchart is made up of two different parts. The left-hand side ("methodology") details the various steps used for hypothesis testing of stationarity and its antithesis nonstationarity. The right-hand side ("metrics and training") explains the steps required to determine a suitable value for ϵ. This value is model and summary metrics dependent and thus needs to be estimated by training against a suite of watersheds known to exhibit stationary/nonstationary behavior. Each step is explained in full in the cited section (abbreviated with the letter "S") next to each block.

We now proceed with a discussion of the most important words used herein in a hope to provide further clarity on our terminology and avoid possible confusion. Note, the words catchment, basin, and watershed are assumed to have an equivalent meaning and are thus used interchangeably. Moreover, our analysis is considered with temporal changes of hydrologic variables—and whether such changes can be classified as stationary or not. We are not concerned with changes of hydrologic variables in space, that is, within the spatial domain of the watersheds of interest. This would require the use of spatially distributed hydrologic models.

1. Reality/real world: the natural system, understood here as the study area of the different watersheds.
2. Watershed response: the reaction of the watershed, as evidenced by observations of key hydrologic variables, to climatic forcing (e.g., rainfall, potential evapotranspiration, PET).

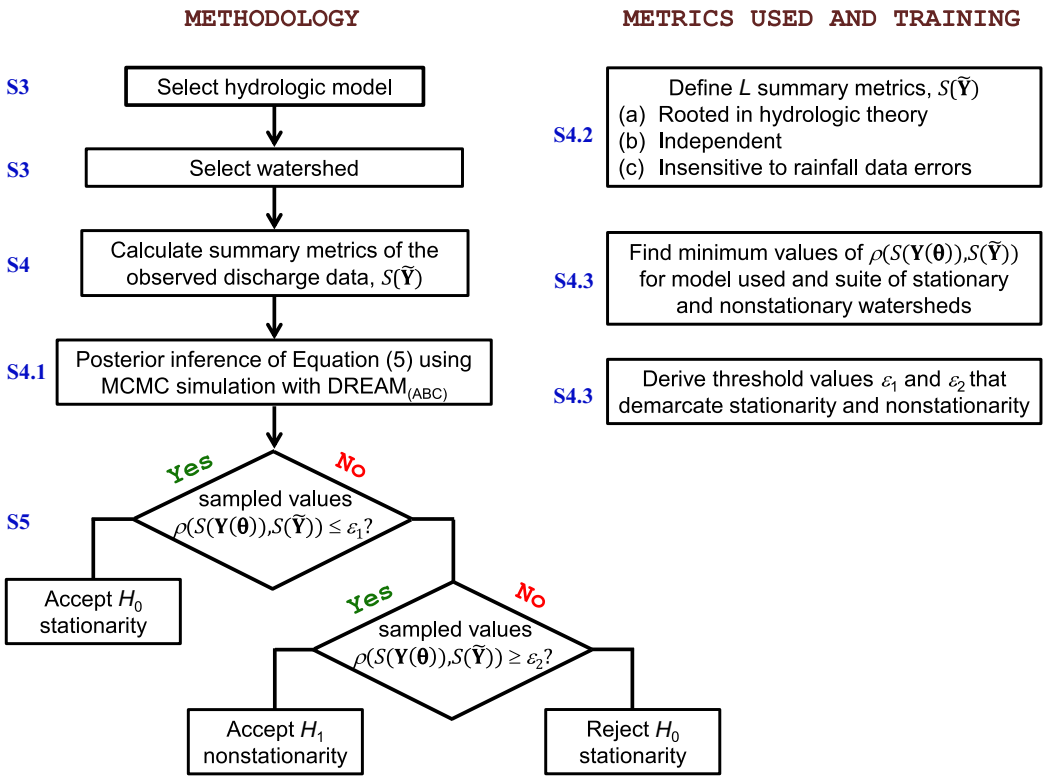

Figure 3. Schematic overview of the stationarity hypothesis testing methodology proposed herein. (left) The various steps of our stationarity hypothesis testing methodology("methodology"). (right) The steps required to determine a suitable value for ϵ ("metrics and training"). The blue labels refer to the individual sections of our paper that discuss in depth each building block.

3. Watershed behavior: the overall integrated functioning of the watershed—not necessarily visible and measurable and in response to climatic forcing.
4. Response data: a glossary of all hydrologic variables that can, in principle, be observed and simulated by a numerical model. This not only includes integrated descriptors of the watershed response to rainfall (and PET) such as the discharge emanating from the catchment outlet, but also soil moisture dynamics at a given depth and location interior to the watershed.
5. Time invariant: catchment characteristics (physical properties) or controlling (hydroclimatologic) processes that do not change over time.
6. Time variant: catchment characteristics (properties) and/or processes that change over time. Antonym of time invariant.
7. Watershed stationarity: watershed behavior that is time invariant. Main difference between the words stationarity and time invariance is that stationarity applies typically to the overall functioning (behavior) of the catchment, as measured in different response variables, whereas time invariance usually applies to a single individual process of the catchment.
8. Watershed nonstationarity: watershed behavior that is time variant. Climate change and/or alterations to the physical characteristics of the catchment (e.g., urbanization, deforestation) are main causes that can induce nonstationarity behavior. Negation of stationarity.
9. True posterior distribution: the distribution of the parameters of a model derived when the model is a perfect description of reality, the input (forcing) data are observed without error and the output (calibration) data are corrupted with random errors only (nonsystematic). The "size" of the output measurement error then determines the width of the posterior distribution.

The word time variant and its antonym time invariant are often used in connection with the parameters of a hydrologic model. Time invariant parameters are parameters whose values are assumed fixed during simulation. This is a valid assumption within the stationarity paradigm. The negation of time invariant parameters, that is, time variant parameters, are parameters whose values are not constant, but rather change, perhaps abruptly, over time. When incorporated in a numerical model, such time variance of the parameters provides a means to simulate catchment nonstationarity [see, for instance, *Westra et al*., 2014]. Note that time invariant is not synonymous to stationary. Time invariant typically is used to describe the behavior of a single process/property—whereas stationary usually applies to the behavior of the system as a whole.

The stationarity hypothesis can be difficult to accept/reject in practice. Indeed, trends in hydrologic response variables (e.g., streamflow) can be due to multiyear climatic variations. One should therefore be particularly careful to proclaim behavior as nonstationarity in the absence of a sufficiently long data record. No statistical method can compensate for insufficient data. The crux is a lack of knowledge about the long-term persistence of the watershed/climate system. This point was made in Figure 1 and *Koutsoyiannis* [2011] and *Koutsoyiannis* [2013]. Similarly, it is difficult to judge whether a process is time invariant and thus stationary or not. Nevertheless, the methodology presented herein much better recognizes the role of process physics in testing of the stationarity paradigm.

Note that our definition of stationarity/nonstationarity differs somewhat from that of *Koutsoyiannis and Montanari* [2014] who base their definitions on a theoretical treatise and state that "*stationarity and nonstationarity apply only to models, not to the real world*" since "*the laws of nature which hold now are identical to those holding for any time in the past or future*." We certainly agree with *Koutsoyiannis and Montanari* [2014] that the laws of nature hold indefinitely, yet perhaps differ in opinion about the cause of nonstationarity. The laws of nature might be invariant but if they act on a porous medium (watershed) that has time variant physical properties (due to, e.g., urbanization, deforestation) or hydroclimatic conditions (e.g., intensifying hydrologic cycle), the catchment behavior can appear nonstationary as evidenced by some trend in the streamflow response. According to this definition, processes such as erosion and sedimentation should, per definition, induce a nonstationarity catchment response, as they actively change the physical properties of the watershed in which they operate. The effect of such structural changes on the integrated catchment response might not be immediately visible unless the respective processes act at large spatial scales. What is more, in our definition, a change in land use can also induce catchment nonstationarity—even if this adaptation is described perfectly in a watershed model of the rainfall-runoff transformation!

Table 1. Description of the Watersheds Used Including Name, Region, USGS ID, Drainage Area (km^2), Mean Annual Precipitation ($\bar{\mathbf{P}}$, mm/d), Mean Annual Evapotranspiration ($\bar{\mathbf{PET}}$, mm/d), and Mean Annual Runoff Coefficient (**MAR**, -)[a]

Name	City (Region), State, Country	USGS ID	Area (km^2)	MAP (mm)	MAPET (mm)	MAR (-)	$\rho_{\min}(\cdot)$ (-)
Stationary Watersheds							
Oostanaula	Resaca, GA, USA	02387500	4150	1414	8653	0.42	0.028
Pearl	Edinburg, MS, USA	02482000	2341	1284	1052	0.31	0.043
Bogue Chitto	Bush, LA, USA	02492000	3141	1441	1068	0.34	0.035
Little Pigeon	Sevierville, TN, USA	03470000	914	1357	7981	0.41	0.036
Tangipahoa	Robert, LA, USA	07375500	1673	1459	1071	0.38	0.022
Comite	Comite, LA, USA	07378000	735	1382	1076	0.36	0.027
Calcasieu	Oberlin, LA, USA	08013500	1950	1452	1094	0.38	0.045
S. Umpqua	Tiller, OR, USA	14308000	1162	1351	813	0.63	0.022
French Broad	Asheville, NC, USA	03451500	2447	1524	819	0.49	0.016
Leaf River	Collins, MS, USA	02472000	1924	1307	1086	0.32	0.022
Nonstationary Watersheds							
Axe Creek	Longlea, Victoria, Australia	N/A	1235	559	1217	0.05	0.171
Wimmera	Glenorchy Concrete Weir Tail, Victoria, Australia	N/A	2000	571	1153	0.07	0.067
Wights	Collie, Western Australia	N/A	0.94	1135	1401	0.24	0.102
Flinders	Glendower, Queensland, Australia	N/A	1911	572	1726	0.11	0.253
Gilbert	Gilberton, South Australia	N/A	1906	827	1074	0.17	0.168
Ferson	St. Charles, IL, USA	05551200	134	1022	756	0.36	0.031
Blackberry	Yorkville, IL, USA	05551700	181	977	773	0.30	0.031
Synthetic Case I ("abrupt")				1524	819	0.52	0.122
Synthetic Case II ("gradual")				1524	819	0.52	0.144

[a]The first 10 watersheds are taken from the MOPEX data set and are deemed stationary. The other nine watersheds are classified as nonstationary in the hydrologic literature. Of these, five watersheds are located in Australia and two in the United States. The last two synthetic data sets involve simulated discharge data from the SAC-SMA model and are used to benchmark our methodology. The last column lists, for each watershed, the minimum distance, $\rho_{\min}(\cdot)$, between the observed and simulated summary metrics derived from equation (5) using the SAC-SMA model. These values are of great importance and will be revisited in later sections of this paper.

The concept of a "true" distribution is somewhat abstract as parameters are modeling constructs that might have nothing to do with the universal constants of the system. This concept is of use however to compare in a relative sense the "distance" of the inference results to their desired "truth." If the system description is erroneous in some way (as is always the case), then the parameters of the model might not map directly to their constants of the actual system. This makes regionalization efforts (among others) more difficult but is beyond the topic of the present paper.

3. Hydrologic Data and Watersheds

In this study, we analyze ten (assumed to be) stationary watersheds from the MOPEX data set [*Schaake et al*., 2006] and nine (claimed to be) nonstationary watersheds. Table 1 lists the name of each watershed, geographical location (state, country), catchment size (km^2), the mean annual precipitation (MAP, mm), potential evapotranspiration (MAPET, mm), and runoff coefficient (MAR, -). The last column will be addressed in the remainder of this paper. The watersheds are located in the United States and Australia and range in size between 0.94 and 2447 km^2.

The different watersheds listed in Table 1 have been classified as stationary or nonstationary in the hydrologic literature based on visual inspection of (multiannual) streamflow data, parametric/nonparametric statistical tests, time series analysis, and anthropogenic alterations (e.g., urbanization, deforestation). This assessment is based on a relatively short data record ($<$50 years) and one should therefore be particularly careful with hypothesis testing due to lack of knowledge of long-term climatic variations. The 10 MOPEX watersheds are classified as stationary, whereas the Wights (deforestation after 3 years from commencement of monitoring in 1974 [e.g., *Mroczkowski et al*., 1997]), Axe Creek, and Wimmera (different rainfall characteristics and much higher temperatures during period of 1997–2008) [*Thirel et al*., 2015], Flinders and Gilbert (cyclonic rainfall), and Ferson and Blackberry Creek (urbanization) watersheds are assumed to exhibit nonstationary behavior. Detailed information about these apparent nonstationary watersheds can be found at http://non-stationarities.irstea.fr/.

We base our arguments and computations of each watershed on a 10 year record of daily data. This length of record is sufficient for a robust calculation of the summary metrics, but appears rather short for

hypothesis testing. The choice of data length is determined by the shortest record on hand of the 19 watersheds. To benchmark our diagnostics approach, we also include two artificial catchments. Two 10 year records of synthetic daily streamflow data were created by driving the Sacramento soil moisture accounting (SAC-SMA) model of *Burnash et al*. [1973] using an explicit fixed-step 6 h integration time step with forcing data from the French Broad River basin in the U.S. The first catchment, coined "abrupt," experiences after 5 years, and from 1 day to the next, a sudden change in SAC-SMA parameter values. The second synthetic catchment experiences a similar adjustment of the parameter values but this adaptation takes place linearly over 365 days between year 5 and 6 of the data record. This catchment is referred to as "gradual" in the remainder of this paper.

4. Approximate Bayesian Computation

ABC was introduced by *Diggle and Gratton* [1984] to permit inference of complex models for which the likelihood is intractable, computationally too expensive to evaluate, or not explicitly available. The premise behind ABC (also called likelihood-free inference) is that the parameter vector $\boldsymbol{\theta}^*$ should be a sample of the posterior distribution if the distance between observed and simulated summary metrics, $\rho(S(\mathbf{Y}(\boldsymbol{\theta}^*)), S(\tilde{\mathbf{Y}}))$ of the data $\tilde{\mathbf{Y}}$, is less than some nominal positive threshold, ϵ [*Beaumont et al*., 2002; *Marjoram et al*., 2003; *Sisson et al*., 2007; *Del Moral et al*., 2012]. The posterior parameter distribution will converge to $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, in the limit of $\epsilon \rightarrow 0$, pending that the summary metrics of the data are sufficient [*Pritchard et al*., 1999; *Beaumont et al*., 2002; *Turner and van Zandt*, 2012]. Practical experience suggests that complex systems such as watersheds hardly admit sufficient statistics. A loss of information is hence expected when projecting the original streamflow data onto a subspace of summary statistics. This loss introduces an extra degree of approximation of the posterior distribution. This, however, is not of great concern in the present application of ABC—where our goal is not the posterior parameter distribution, but rather to analyze temporal variations in the summary statistics.

The use of summary metrics holds several great promises for diagnostic model evaluation. First, model fitting against summary metrics of the data helps ensure that the simulated response portrays accurately properties of the observed behavior deemed important to the user. Such a match cannot be guaranteed when fitting a model using a likelihood function consisting of some convoluted time series of error residuals of the observed and simulated data. Second, if we carefully design each summary metric to be sensitive to only one component (process) of the model, then after inference, it should be relatively easy to pinpoint which part of the model is malfunctioning. Finally, summary metrics can be designed so that they are relatively insensitive to forcing data errors (of which more later). This is a particularly desirable characteristic in the present context—as we want to avoid proclaiming nonstationarity based on errors of the rainfall data.

Figure 4 schematically depicts, in six different steps, the ABC framework for a simple example involving the fitting of a model to an observed hydrograph. First, the user defines one or more summary metrics, $S(\tilde{\mathbf{Y}})$ that summarize the original streamflow data (step 1: dots) in a (much) lower-dimensional space (step 2). The value of ϵ (size of purple box) defines the behavioral solution space, deemed appropriate by the user. Then, N samples, $\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*\}$, are drawn from the prior parameter distribution (step 3). In the absence of detailed prior information about the individual model parameters, this distribution is often assumed to be uniform (noninformative) to not favor a priori any parameter values. (Note that *Scharnagl et al*. [2011] present a strategy to assess informative prior distributions in vadose zone modeling.) The N proposals drawn from the prior distribution are subsequently evaluated by the hydrologic model, $\mathcal{H}(\boldsymbol{\theta}_i^*, \cdot) + \mathbf{e}$ (step 4), where $\mathbf{e}$ denotes a n-vector with draws from the (unknown) residual distribution (more of which later). The summary metrics, $S(\mathbf{Y}(\boldsymbol{\theta}_i^*))$, of each hydrograph simulation, $\mathbf{Y}(\boldsymbol{\theta}_i^*)$, are then compared to their observed values (step 5); if $\rho(S(\mathbf{Y}(\boldsymbol{\theta}_i^*)), S(\tilde{\mathbf{Y}})) \leq \epsilon$ then $\boldsymbol{\theta}_i^*$ is considered to be a behavioral (posterior) solution. The accepted samples are finally used to summarize the target distribution, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ (step 6). Note, whenever the index i is used, we mean for each $i \in \{1, \ldots, N\}$.

Obviously, the degree of approximation of the true posterior distribution depends in large part on the chosen summary statistics, along with the value of ϵ used to differentiate between behavioral and nonbehavioral solutions. The choice of summary metrics is of imminent importance and should reflect (among others) the purpose of model application. Section 4.2 of this paper will discuss, in detail, the selection of the summary metrics.

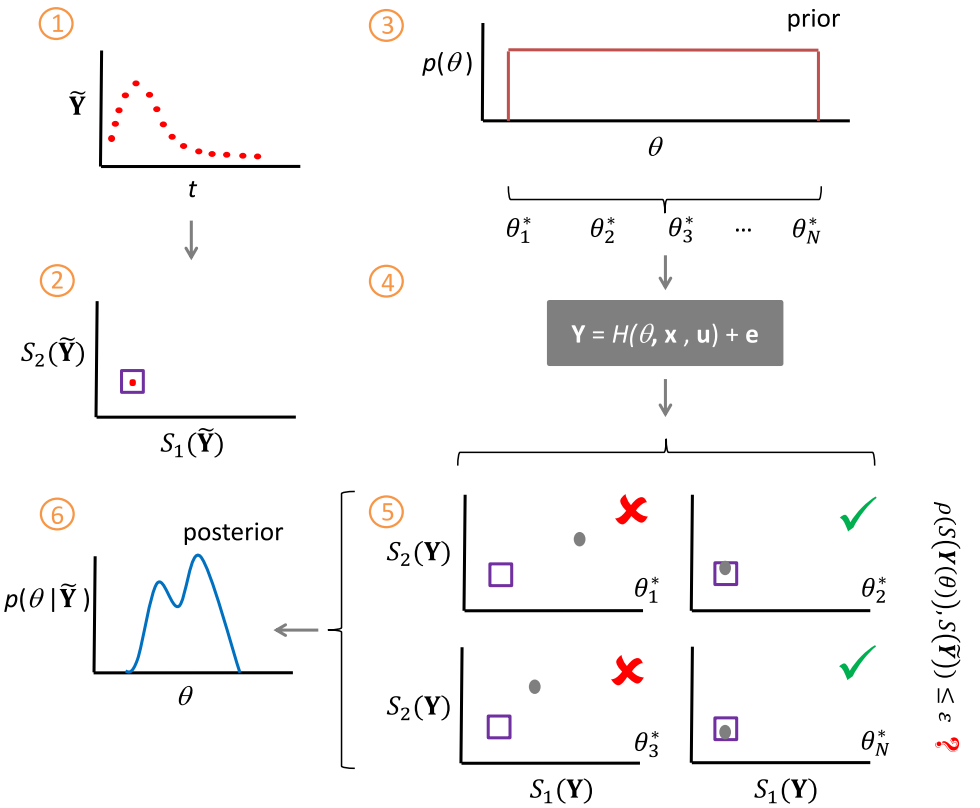

Figure 4. Schematic overview of likelihood-free inference with approximate Bayesian computation (ABC) for a model with one parameter and two different summary metrics. The original watershed data, $\tilde{\mathbf{Y}}$ is plotted in step 1 and is used to calculate S_1 and S_2, the summary metrics of the measured data (step 2). After the user has defined ϵ, step 3 proceeds by drawing N samples from the prior parameter distribution, $p(\theta)$. In step 4, each of these N samples is evaluated by the model, and used to calculate the simulated values of the summary metrics. Step 5 then proceeds with a comparison of the observed and simulated summary metrics. Model realizations whose simulated summary metrics fall within the purple box, and thus satisfy $\rho(S(\mathbf{Y}(\boldsymbol{\theta}_i^*)), S(\tilde{\mathbf{Y}})) \leq \epsilon$ are called behavioral and used in step 6 to approximate the posterior parameter distribution.

4.1. Posterior Sampling: DREAM$_{(ABC)}$

Application of likelihood-free inference with ABC requires the availability of a sampling method that can efficiently search the parameter space in pursuit of behavioral solutions. Commonly used rejection sampling methods are rather inefficient—the chance that a randomly sampled solution will fall exactly within the hypercube defined by ϵ (see Figure 4) is disturbingly small, particularly if the prior parameter distribution is large compared to the behavioral (posterior) solution space and the number of summary metrics is large [*Sadegh and Vrugt*, 2013]. We therefore take advantage of the DREAM$_{(ABC)}$ algorithm developed by *Sadegh and Vrugt* [2014].

In DREAM$_{(ABC)}$, K ($K > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chain states, $\boldsymbol{\Theta}_{t-1}$ (matrix of $K \times d$ with each chain state as row vector) using differential evolution [*Storn* and *Price*, 1997; *Price et al.*, 2005]. If A is a subset of d^* dimensions of the original parameter space, $\mathbb{R}^{d^*} \subseteq \mathbb{R}^d$, then a jump, $\mathbf{d}\boldsymbol{\Theta}^k$, in the kth chain, $k = \{1, \ldots, K\}$, at iteration $t = \{2, \ldots, T\}$ is calculated from the collection of chains $\boldsymbol{\Theta}_{t-1} = \{\boldsymbol{\theta}_{t-1}^1, \ldots, \boldsymbol{\theta}_{t-1}^K\}$ using differential evolution [*Storn and Price*, 1997; *Price et al.*, 2005]

$$\mathbf{d}\boldsymbol{\Theta}_{\mathrm{A}}^k = \boldsymbol{\zeta}_{d^*} + (\mathbf{1}_{d^*} + \boldsymbol{\lambda}_{d^*})\gamma_{(\delta,d^*)} \sum_{j=1}^{\delta} (\boldsymbol{\Theta}_{\mathrm{A}}^{\mathbf{a}_j} - \boldsymbol{\Theta}_{\mathrm{A}}^{\mathbf{b}_j})$$
$$\mathbf{d}\boldsymbol{\Theta}_{\neq \mathrm{A}}^k = 0, \tag{1}$$

where $\gamma = 2.38/\sqrt{2\delta d^*}$ is the jump rate, δ denotes the number of chain pairs used to generate the jump, and $\mathbf{a}$ and $\mathbf{b}$ are vectors consisting of δ integers drawn without replacement from $\{1, \ldots, k-1, k+1, \ldots, K\}$.

The default value of $\delta = 3$, and results, in practice, in one third of the proposals being created with $\delta = 1$, another third with $\delta = 2$, and the remaining third using $\delta = 3$. The values of $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}$ are sampled independently from $\mathcal{U}_{d^*}(-c, c)$ and $\mathcal{N}_{d^*}(0, c_*)$, respectively, the multivariate uniform and normal distribution with, typically, $c = 0.1$ and c_* small compared to the width of the target distribution, $c_* = 10^{-6}$ say. With a 20% probability, the value of the jump rate is set equal to unity, $\gamma = 1$, to enable DREAM$_{(ABC)}$ to jump between disconnected modes of the target distribution [*Vrugt et al.*, 2008, 2009; *Sadegh and Vrugt*, 2014].

The candidate point of chain k at iteration t then becomes

$$\boldsymbol{\Theta}_{\mathrm{p}}^{k} = \boldsymbol{\Theta}^{k} + \mathbf{d}\boldsymbol{\Theta}^{k}, \tag{2}$$

and a modified selection rule is used to determine whether to accept this proposal or not. This selection rule is defined as

$$P_{\mathrm{acc}}(\boldsymbol{\Theta}^{k} \rightarrow \boldsymbol{\Theta}_{\mathrm{p}}^{k}) = \begin{cases} I(f(\boldsymbol{\Theta}_{\mathrm{p}}^{k}) \geq f(\boldsymbol{\Theta}^{k})) & \text{if } f(\boldsymbol{\Theta}_{\mathrm{p}}^{k}) < 0 \\ 1 & \text{if } f(\boldsymbol{\Theta}_{\mathrm{p}}^{k}) \geq 0 \end{cases}, \tag{3}$$

If $P_{\mathrm{acc}}(\boldsymbol{\Theta}^{k} \rightarrow \boldsymbol{\Theta}_{\mathrm{p}}^{k}) = 1$, the candidate point is accepted and the kth chain moves to the new position, that is, $\boldsymbol{\theta}_{t}^{k} = \boldsymbol{\Theta}_{\mathrm{p}}^{k}$, otherwise $\boldsymbol{\theta}_{t}^{k} = \boldsymbol{\theta}_{t-1}^{k}$.

The function, $f(\cdot)$ is calculated as follows

$$f(\boldsymbol{\theta}) = \epsilon - \rho(S(\mathbf{Y}(\boldsymbol{\theta})), S(\tilde{\mathbf{Y}})), \tag{4}$$

and the distance function is defined as

$$\rho(S(\mathbf{Y}(\boldsymbol{\theta}), S(\tilde{\mathbf{Y}}))) = \max_{l=1:L} \left(|S_l(\mathbf{Y}(\boldsymbol{\theta})) - S_l(\tilde{\mathbf{Y}})| \right), \tag{5}$$

where L signifies the number of summary statistics used. Thus, we accept, $P_{\mathrm{acc}}(\boldsymbol{\Theta}^{k} \rightarrow \boldsymbol{\Theta}_{\mathrm{p}}^{k}) = 1$, if the fitness of $\boldsymbol{\Theta}_{\mathrm{p}}^{k}$ is larger than that of the current state of the kth chain, $\boldsymbol{\Theta}^{k}$ or if the summary metrics of the proposal are within ϵ of their observed counterparts, $f(\boldsymbol{\Theta}_{\mathrm{p}}^{k}) \geq 0$, otherwise the candidate point is rejected. After a burn-in period and acquisition of $f(\cdot) > 0$, the convergence of DREAM$_{(ABC)}$ can be monitored with the $\hat{R}$ diagnostic of *Gelman and Rubin* [1992].

For all watersheds, we assume default settings of the algorithmic variables of DREAM$_{(ABC)}$ and use $K = 3$ different chains and 100,000 model evaluations. To minimize the burn in, we use a related variant of DREAM$_{(ABC)}$ which uses past states in the jumping distribution of equation (1) [*Laloy and Vrugt*, 2012; *Vrugt*, 2015].

4.2. Choice of Summary Metrics

The choice which metrics to use depends on the goal of the diagnostics application. If the main subject of interest is posterior approximation, then the summary statistics should be sufficient and "*convey all relevant information*" of the data (quote of *Edwards* [1992]). Self-sufficiency is not necessarily a requirement for diagnostic model evaluation in which the goal is not to approximate the true posterior distribution but rather to detect epistemic errors arising from incomplete and/or inadequate process knowledge. The advantage of summary statistics is that they exhibit a much better diagnostic power than some (purely statistical) likelihood function of a convoluted time series of error residuals, and if properly designed and rooted in the relevant environmental theory, can help to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be refined.

We now have to decide which summary statistics to use. To facilitate diagnostic inference of watershed models, it would be highly desirable if the chosen summary statistics are, (a) properly rooted in hydrologic theory, (b) independent (uncorrelated), (c) invariant to precipitation data errors, and (d) sensitive only to one specific component (process/equation) of the model. Criteria (a), (b), and (c) are a necessary requirement to test the stationarity null hypothesis. Criterion (d) is of secondary importance in the present work and detection of epistemic errors will be pursued in other work. If the summary metrics are designed so that they extract the relevant signatures of catchment behavior, then analysis of their temporal dynamics can help elucidate subtle changes in the basin response to rainfall. At least, this is the underlying premise of our hypothesis testing methodology. If the stationarity assumption is valid then one

Table 2. Name and Description of Each of the Summary Statistics Used in Our Rainfall Error Analysis[a]

Name	Description	Reference
Base flow index	Ratio between the total base flow and streamflow volumes in the period of study	*Eckhardt* [2005]
Runoff coefficient	Ratio between the total streamflow and precipitation volumes in the period of study	*Savenije* [1996]
FDC	Relationship between the exceedence probability of streamflow and its magnitude	*Searcy* [1959]
Slope of the FDC	Slope of the midpart of the FDC (between 33% and 66% exceedance probability rates)	*Yadav et al.* [2007]
Rising Limb Density (RLD)	Ratio between the number of peaks and cumulative time of rising limbs	*Morin et al.* [2002]
Declining Limb Density (DLD)	Ratio between the number of peaks and cumulative time of declining limbs	*Shamir et al.* [2005]
Mean Daily Flow (DF)	Average of daily streamflow values	*Clausen and Biggs* [2000]
Median DF	Median of daily streamflow values	*Clausen and Biggs* [2000]
Coefficient of variation of DFs	Ratio between the standard deviation and mean of daily streamflow values	*Clausen and Biggs* [2000]
Skewness of DFs	Skewness of daily streamflow values	*Clausen and Biggs* [2000]
Ranges in DFs	Ratio of 10th/90th, 20th/80th,, and 25th/75th, percentiles of daily streamflow values	*Olden and Poff* [2003]
Mean maximum monthly flow	Average of the maximum monthly streamflow values	*Olden and Poff* [2003]
Mean minimum monthly flow	Average of the minimum monthly streamflow values	*Olden and Poff* [2003]
Median maximum monthly flow	Median of the maximum monthly streamflow values	*Olden and Poff* [2003]
Median Minimum Monthly Flow	Median of the minimum monthly streamflow values	*Olden and Poff* [2003]
Low flow pulse count	Number of events with streamflow values below 25th percentile	*Olden and Poff* [2003]
High flow pulse count	Number of events with streamflow values above 75th percentile	*Olden and Poff* [2003]
Low flow pulse duration	Mean duration of low streamflows	*Olden and Poff* [2003]
High flow pulse duration	Mean duration of high streamflows	*Olden and Poff* [2003]
Flood duration	Mean number of days that streamflow magnitude remains above flood threshold (75*th* percentile)	*Olden and Poff* [2003]
Flood frequency	Mean number of high flow (flood) events per year	*Olden and Poff* [2003]
Rise rate	Average rate of positive flow changes from 1 day to the next	*Olden and Poff* [2003]
Fall rate	Average rate of negative flow changes from 1 day to the next	*Olden and Poff* [2003]
Number of zero flow days	Number of days with zero streamflow	*Olden and Poff* [2003]
Peak distribution	Average slope between 10th and 50th percentiles of the FDC from peak flows only	*Euser et al.* [2013]

[a]The last column provides a reference for each metric, and can be used as guide for their numerical calculation.

expects the summary metrics to vary around some constant mean, and temporally invariant parameter values to be sufficient to mimic adequately the observed catchment response. The alternative hypothesis would involve temporal variant summary statistics, symptomatic for catchments subject to physical alterations (urbanization, deforestation, and/or changes in land use) or changes in hydroclimate (changing rainfall characteristics).

It is of paramount importance that the summary statistics are independent and rooted in hydrologic theory. Indeed, we can only verify the stationarity hypothesis if we use metrics that measure directly catchment behavior. These metrics should be independent to ensure that they each measure different and complementary parts of catchment functioning. Moreover, it is highly desirable that the summary metrics should be insensitive to precipitation data errors. We certainly would not want to reject the null hypothesis and proclaim nonstationarity of catchment response based on the wrong reasons.

The hydrologic literature has brought forward a plethora of summary statistics of a discharge data record that could potentially be used as signatures of catchment behavior. Table 2 summarizes our initial selection of the summary metrics potentially deemed adequate for stationarity hypothesis testing. We list the name of each metric along with a short description and reference which discusses their mathematical calculation. In summary, we consider the following 25 summary statistics of the observed discharge data: base flow index, runoff coefficient, slope [*Yadav et al*., 2007], and two other fitting parameters of the flow duration curve (FDC) [*Vrugt and Sadegh*, 2013], rising and declining limb density [*Shamir et al*., 2005], high and low flow pulse count and duration, rise and fall rate, mean, median, coefficient of variation, skewness, and the maximum range of the daily discharge observations, flood duration and frequency, number of zero flow days [*Olden and Poff*, 2003], peak distribution [*Euser et al*., 2013], and the mean and median values of the minimum and maximum monthly flows, respectively.

Table 3. Description of the SAC-SMA Model Parameters and Their (Uniform) Prior Uncertainty Ranges

Parameter	Symbol	Minimum	Maximum	Units
Upper zone tension water maximum storage	UZTWM	1.0	150.0	mm
Upper zone free water maximum storage	UZFWM	1.0	150.0	mm
Lower zone tension water maximum storage	LZTWM	1.0	500.0	mm
Lower zone free water primary maximum storage	LZFPM	1.0	1000.0	mm
Lower zone free water supplemental maximum storage	LZFSM	1.0	1000.0	mm
Additional impervious area	ADIMP	0.0	0.4	–
Upper zone free water lateral depletion rate	UZK	0.1	0.5	day^{-1}
Lower zone primary free water depletion rate	LZPK	0.0001	0.025	day^{-1}
Lower zone supplemental free water depletion rate	LZSK	0.01	0.25	day^{-1}
Maximum percolation rate	ZPERC	1.0	250.0	–
Exponent of the percolation equation	REXP	1.0	5.0	–
Impervious fraction of the watershed area	PCTIM	0.0	0.1	–
Fraction percolating from upper to lower zone free water storage	PFREE	0.0	0.6	–

We now need to determine which of these metrics satisfy the three criteria (a)–(c) listed previously. To this end, we first determine their individual sensitivity to precipitation data errors. We devised a simple numerical experiment with the SAC-SMA model [*Burnash et al.*, 1973]. Synthetic daily streamflow observations were created with 10 different parameter combinations, drawn randomly from their uniform prior ranges specified in Table 3, using forcing data from the French Broad River basin at Asheville, NC. The summary metrics of Table 2 were calculated for each of ten simulated discharge records and these vectors of metrics were then considered to be our observed values. Then, a total of thousand different precipitation time series were created by perturbing the original hyetograph of the French Broad River basin with a 20% (heteroscedastic) measurement error. On top of this, (systematic) errors in rainfall timing were introduced by moving randomly, with one day, storm events. The probability of such move was set to 20%. This leaves us with an ensemble of one thousand different rainfall records. The SAC-SMA model was subsequently executed with each of the ten different parameterizations and thousand different precipitation data records. Next, the summary metrics of Table 2 were calculated for the ensemble of 10,000 discharge simulations and compared to their respective observed values. To investigate the effect of data length on the simulated summary statistics, we consider daily discharge simulations of 1, 2, 5, and 10 years length. Note that our analysis does not consider temporal correlation of the rainfall data errors. If such errors were present then detailed prior information is required about their presence—no method can otherwise provide compelling results.

Figure 5 displays the results of our analysis, and plots histograms of the distance (residual) between the observed and simulated values of the summary statistics. Figures 5a–5e depict the five metrics that appear least sensitive to rainfall data errors, whereas the Figures 5f–5j summarize the results of the five most sensitive criteria. The results pertain to a 10 year data period and summarize the outcome of 10 different SAC-SMA model parameterizations. The base flow index, runoff coefficient, two fitting coefficients of the FDC, and annual rising limb density (RLD) appear very well defined and almost unaffected by rainfall data errors. To maximize information retrieval from the observed discharge data, the RLD is defined for each year separately equating to a total of 14 different metrics for a 10 year data. Further analysis demonstrates these $L = 14$ metrics to be rather uncorrelated—and thus to satisfy the independence criterion (b) as well. The other metrics of Table 2 show a considerable variation in response to rainfall data errors, and thus are not deemed particularly useful for stationarity hypothesis testing. A similar selection of summary metrics was made for shorter simulation periods (not shown herein).

In the present analysis, we discard errors in the streamflow data itself. A simple numerical experiment with artificial discharge data will show that the selected summary metrics appear rather insensitive to errors in the discharge data—unless these errors are nonrandom (systematic). In that case, no statistical method would provide compelling results within the present context. As the watershed response is generally dominated by rainfall events, we ignore PET errors in the current paper, and leave this for future work. It would not be difficult to devise a similar numerical test as done for rainfall data errors to explore the sensitivity of each individual summary metric to measurement errors of the potential evapotranspiration.

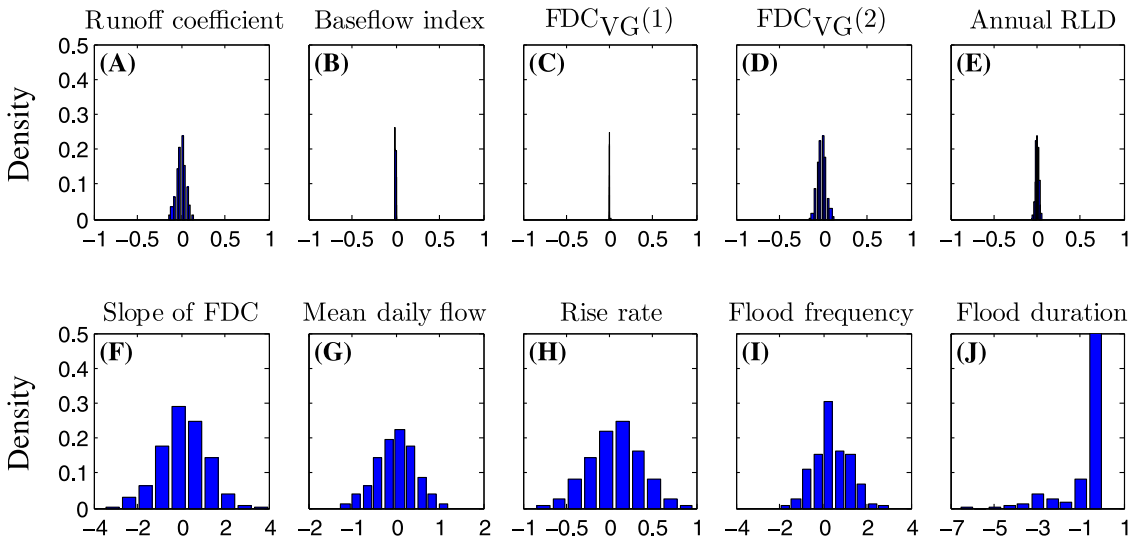

Figure 5. Histograms of the residuals between the observed and simulated summary metrics. The observed values of the summary metrics are derived from a synthetic discharge record simulated with the SAC-SMA model using the rainfall hyetograph of the French Broad River basin. The simulated metrics are derived from the SAC-SMA model but using systematic and random errors to the rainfall data record. The aggregated results of 10 different SAC-SMA parameterizations with 1000 corrupted rainfall records each are plotted. (a–e) The summary metrics whose values are least affected by rainfall data errors, and (f–j) their most sensitive counterparts.

4.3. Choice of Cutoff Threshold, ϵ

Now our summary statistics have been defined, and we are left with the values of **e** and ϵ used to differentiate between behavioral and nonbehavioral solutions (simulations). If the main goal of ABC application is to approximate the true posterior parameter distribution then the value of ϵ should be taken small (e.g., $\epsilon \leq 0.025$) and **e** should reflect accurately the probabilistic properties of the remaining error between the model operator, $\mathcal{H}(\cdot)$ and the actual data generating process. This includes possible errors in hydroclimatic forcing as well. For diagnostic model evaluation, however, the assumption **e**=0 is sufficient to help address the stationarity null hypothesis. This leaves us with specification of ϵ only, and as will be shown later, this scalar delineates stationary from nonstationary watershed behavior.

No guidelines exist what value (or vector of values) of ϵ to select for testing of the stationarity hypothesis in a diagnostics framework. It seems logical to let our choice of ϵ depend directly on the temporal variability of each individual summary metric. One would then expect the "width" around the mean value of each summary metric, or second moment, to relate directly to the "amount" of nonstationarity of $\tilde{\mathbf{Y}}$. Of course, a statistically significant change in the mean can only be detected if a sufficiently long data record is used (see Figure 1).

The last column of Table 1 lists, for each of the 19 watersheds considered in our analysis, the minimum value of $\rho(S(\mathbf{Y}(\boldsymbol{\theta}), S(\tilde{\mathbf{Y}})))$ derived from minimization of equation (5) using the SAC-SMA model and differential evolution algorithm. Equation (5) is composed of the fourteen different summary metrics selected in section 4.2. The tabulated values highlight three important findings.

In the first place, the minimized value of $\rho(\cdot)$ has important diagnostic power to address the stationarity null hypothesis. Indeed, a value of $\rho(\cdot) \leq 0.05$ demarcates stationary watershed behavior, whereas values of $\rho(\cdot) > 0.10$ support the presence of nonstationary. Second, the Ferson and Blackberry Creek basins in the U.S. satisfy the stationary test of $\rho(\cdot) \leq 0.05$, but have been classified as nonstationary in the hydrologic literature due to significant urbanization. We believe that this literature classification is erroneous, and this finding is supported by the evidence presented in Figure 2 for the Blackberry basin. Third, the Wimmera River basin is classified in Table 1 as nonstationary, but its value of $\rho(\cdot)$=0.067 does not satisfy formally $\rho(\cdot) \geq 0.10$. For values of $\rho(\cdot) \leq 0.05$, we can declare with a high level of confidence the presence of stationarity. For values of $\rho(\cdot) \geq 0.10$, nonstationarity is implied. For values of $\rho(\cdot) \in [0.05, 0.10]$, the null hypothesis (stationarity) is rejected but the evidence is not strong enough to support with high confidence the nonstationary hypothesis. In principle, this classification can be adapted to support a probabilistic interpretation of the degree of stationarity, P_s, and its antithesis, nonstationarity, $P_s = 1 - P_n$, where P_n signifies the probability of nonstationarity. We leave this for future work. Of course, in theory, every watershed is nonstationary, to some extent, due to global changes of the hydrologic cycle, geomorphologic and land

use changes. The presented approach should help ascertain whether change is present in the data, and if so, how significant this change is (probability of nonstationarity).

Thus, the value of $\rho(\cdot)$ appears to be a useful proxy for the degree of watershed nonstationarity. In practice, if the SAC-SMA model cannot find a value of equation (5) smaller than 0.05, the respective watershed under consideration fails the stationarity null hypothesis. For values of $\rho(\cdot) > 0.10$, we can assume the watershed to have experienced changes to its physical characteristics and/or hydroclimatic conditions. The resulting behavior is then classified as nonstationary. This binary classification scheme serves to accept/reject the stationary null hypothesis, nevertheless a probabilistic interpretation of $\rho(\cdot)$ is preferred statistically. This would provide, for each watershed, an estimate of its degree of nonstationarity. Based on this analysis, we assume a value of ϵ=0.05 in all our numerical experiments.

A few remarks are in order. The criteria of $\rho(\cdot) \leq 0.05$ used to demarcate stationarity apply to the $L = 14$ summary metrics and SAC-SMA model used herein and/or models with similar structural complexity. For more parsimonious models with fewer calibration parameters, one would expect the reported values of $\rho(\cdot)$ to increase due to a reduced ability of the model to describe closely the observed signatures. For instance, consider a simple linear model that is used to approximate the rainfall-runoff transformation. It will not be a surprise that this linear model is unable to closely fit the observed values of the summary metrics for each of the nineteen watersheds considered herein. Indeed, the values of $\rho(\cdot)$ for this linear model will be substantially larger than their counterparts of the SAC-SMA model. The largest values of $\rho(\cdot)$ will still be observed for the nonstationary watersheds, and hence for this linear model the threshold values of $\rho(\cdot)$ used to test for stationarity/nonstationarity simply increases beyond 0.05 and 0.10, respectively. The diagnostic power of $\rho(\cdot)$ for hypothesis testing thus remains unaffected. A similar change to these criteria is expected with the use of another set of summary metrics—pending the assumption that they each satisfy the three criteria listed previously (hydrologic relevance, independent, and insensitive to rainfall data errors).

Now we have defined ϵ=0.05 to be appropriate for the SAC-SMA model, we can now use our methodology to confront the stationarity null hypothesis. In the next section, we illustrate the results of our methodology for three watersheds deemed stationary in the hydrologic literature, and three basins that are assumed to exhibit nonstationarity behavior.

5. Results and Discussion

We now illustrate the results of our numerical simulations with the SAC-SMA model using the DREAM$_{(ABC)}$ algorithm with the $L = 14$ summary statistics of section 4.2 derived from 10 years of daily hydrologic data from the 19 different watersheds listed in Table 1. In our discussion, we focus attention on three stationary watersheds (Oostanaula, French Broad, and Leaf River) and three nonstationary watersheds (Wights catchment, and two synthetic data sets). The findings for these six basins are representative for the entire collection of watersheds studied herein, and demonstrate the ability of the proposed diagnostics methodology to differentiate between stationary and nonstationary watershed behavior.

Figure 6 presents histograms of the marginal posterior distribution of a representative set of five SAC-SMA parameters including (a) PCTIM, (b) ADIMP, (c) LZTWM, (d) LZPK, and (e) LZSK using the observed values of the $L = 14$ different summary metrics of the Oostanaula (top), French Broad (middle), and Leaf (bottom) River basins and ϵ=0.05. These three basins are classified as stationary in the hydrologic literature. The x axis of each plot matches exactly the ranges of each parameter used in the (uniform) prior distribution (Table 3). The histograms are created from the last 20% of the joint samples of the Markov chains. The results in this figure highlight several important findings. First, most of the SAC-SMA parameters are not particularly well resolved by calibration against the observed summary metrics. The marginal distributions encompass a large part of the prior distribution. The exception is LZPK which appears reasonably well constrained for each of the stationary watersheds. Second, most of the histograms are rather irregular with multiple modes and poorly described with a traditional probability density function. These results suggest that the posterior surface is nonsmooth and exhibits many local minima. Similar conclusions were drawn in our earlier diagnostics work presented in *Vrugt and Sadegh* [2013]—but then using a model with an implicit numerical solver. Thus, the rather erratic posterior parameter distributions are likely the consequence of the metrics used and less likely caused by inferior model numerics. Finally, although the marginal distributions

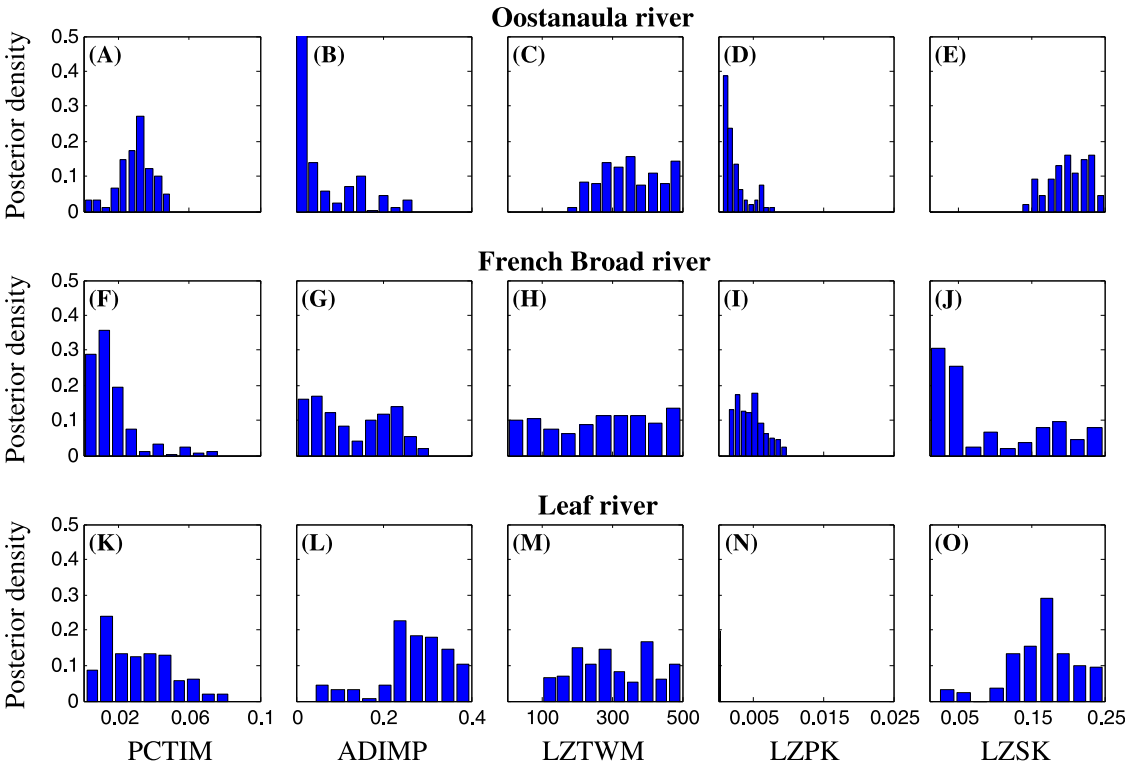

Figure 6. Histograms of the marginal posterior distribution of a representative set of five SAC-SMA model parameters (PCTIM, ADIMP, LZTWM, LZPK, and LZSK) derived from DREAM$_{(ABC)}$ using historical data from the (top) Oostanaula, (middle) French Broad, and (bottom) Leaf River watershed, respectively. Some of the SAC-SMA parameters are well resolved by calibration against the observed summary metrics, whereas others exhibit considerable uncertainty.

exhibit considerable scatter—multivariate plots of the behavioral solutions demonstrate that the posterior samples occupy only a small portion of the uniform prior hypercube (not shown).

To provide better insights into the sampled values of equation (5), Figure 7 plots trace plots of $\rho(\cdot)$ in each of the $K = 3$ Markov chains simulated with DREAM$_{(ABC)}$. We illustrate the results for the (a) Oostanaula, (b) French Broad, and (c) Leaf River watersheds. These three watersheds are deemed stationary in the hydrologic literature. Each of the three Markov chains is coded with a different color. The dashed red line signifies the threshold of ϵ=0.05 used to demarcate stationarity. The DREAM$_{(ABC)}$ algorithm rapidly converges to a limiting distribution. About 5000 function evaluations are required for each watershed to locate solutions that satisfy the stationarity assumption. The number of function evaluations with DREAM$_{(ABC)}$ is more than sufficient to generate a large sample of the posterior distribution. Note that the different chains mix relatively well—indeed the acceptance rate varies between 3.24 and 5.54%.

We now move on to the sampled summary statistics. Figure 8 plots histograms of the marginal distribution of the sampled summary metrics of the SAC-SMA model for the Oostanaula (top), French Broad (middle), and Leaf River (bottom) basins. The observed values of each summary metric are indicated separately in each plot with a red cross. The histograms are created from the last 20% of the joint samples of the Markov chains. As it is particularly difficult to summarize the results of all the $L = 14$ summary statistics, we plot only the distributions of S_1 (Figures 8a, 8f, and 8k) the runoff coefficient, S_2 (Figures 8b, 8g, and 8l) the base flow index, S_3 (Figures 8c, 8h, and 8m) and S_4 (Figures 8d, 8i, and 8n) the two fitting coefficients of the van Genuchten formulation of the FDC, and S_{11} (Figures 8e, 8j, and 8o) the RLD of year 7 of the simulation period. The sampled distributions center nicely around the observed values of the summary metrics and appear rather uniform. The different plots confirm stationarity, that is, the sampled summary metrics reside within a distance of ϵ=0.05 of their observed values. A fixed parameterization of the SAC-SMA model appears sufficient to mimic adequately close the observed summary statistics.

Figure 9 illustrates how the SAC-SMA posterior parameter uncertainty translates into streamflow simulation uncertainty. A representative 365 day period of the 10 year data record is used to illustrate our findings. The

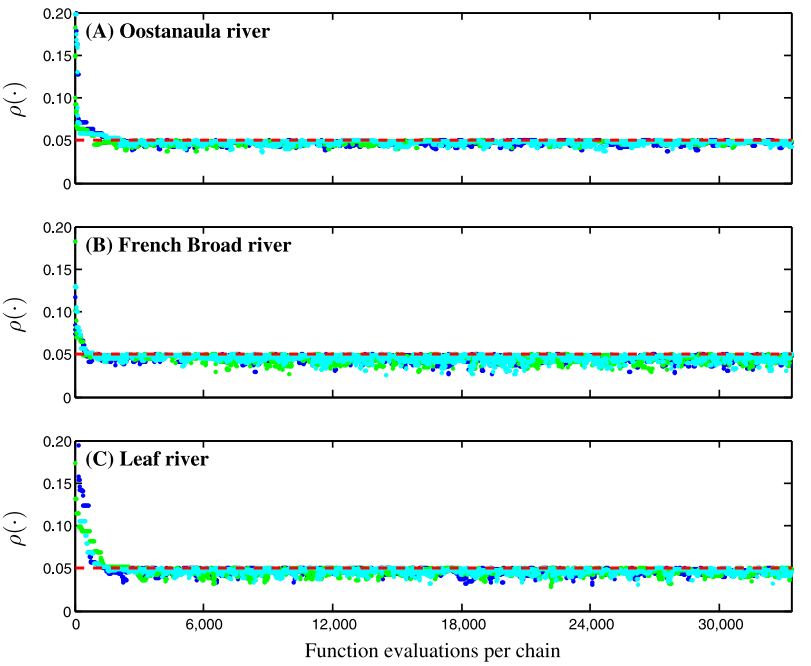

Figure 7. Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the three Markov chains for the (a) Oostanaula, (b) French Broad, and (c) Leaf River watersheds. The dashed red line depicts the threshold of ϵ=0.05 used to demarcate stationarity.

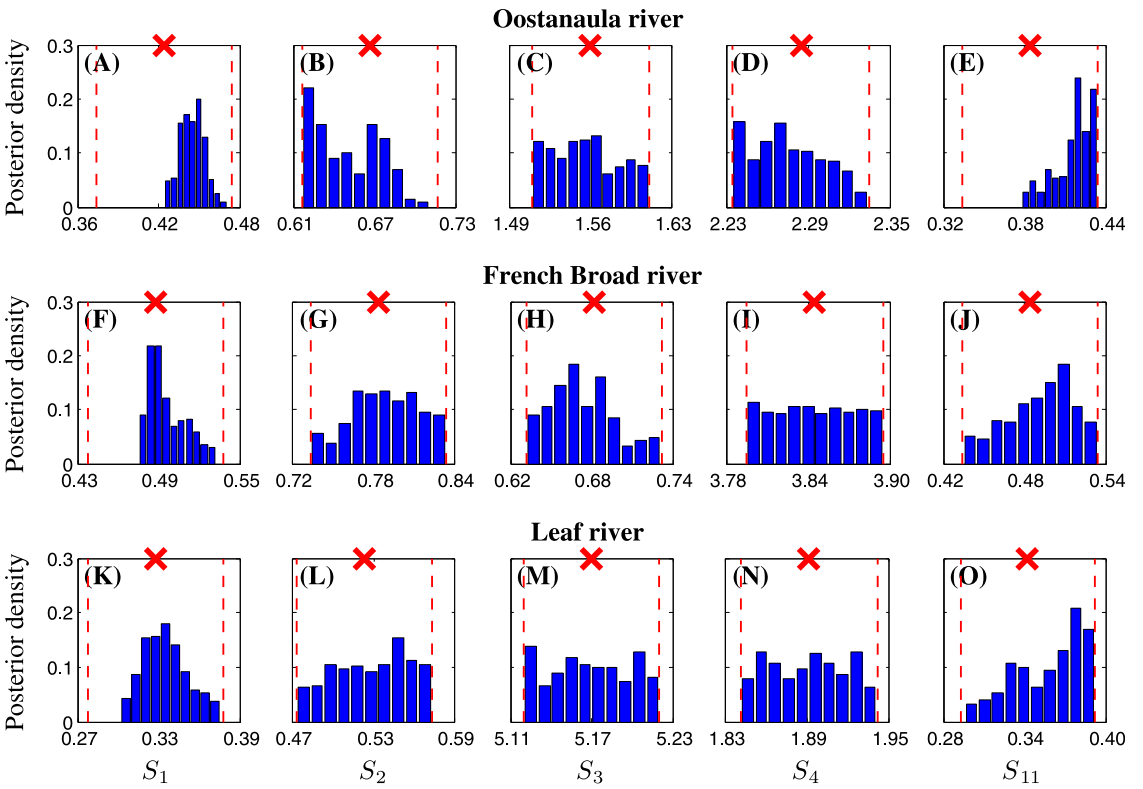

Figure 8. Histograms of the marginal posterior distribution of S_1: runoff coefficient, S_2: base flow index, S_3, S_4: fitting parameters of the flow duration curve, and S_5: the rising limb density of year 7 of the calibration data period. We separately display the results for the (top) Oostanaula, (middle) French Broad, and (bottom) Leaf River watersheds. The red cross in each plot signifies the measured value of each summary metric. The vertical dashed lines delineate the behavioral solution space. A necessary condition for stationarity is that all $L = 14$ summary metrics fall within the dashed interval around the observed value of each summary metric.

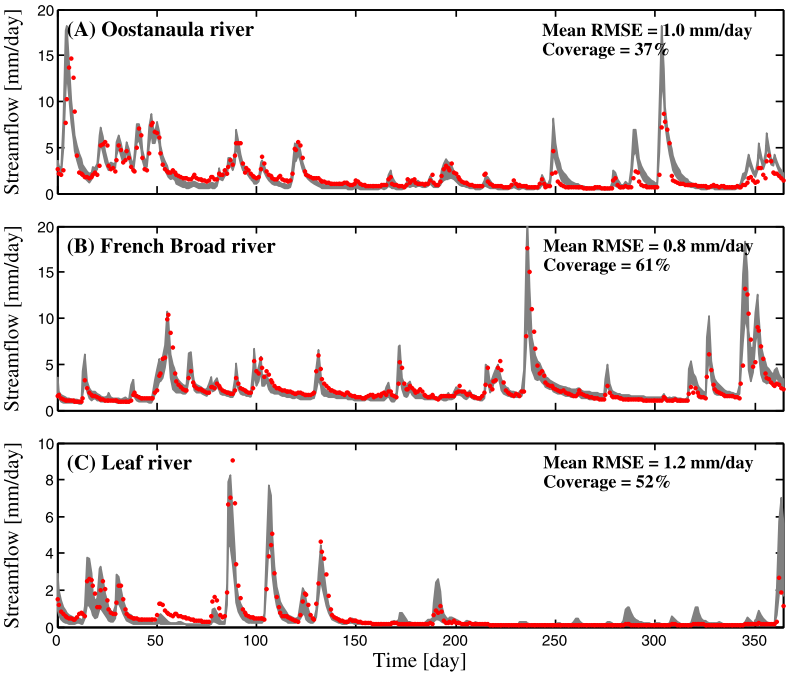

Figure 9. The 95% posterior simulation uncertainty ranges of the SAC-SMA model for the (a) Oostanaula, (b) French Broad, and (c) Leaf River watersheds for a representative 365 days portion of the calibration data period. The simulation uncertainty ranges (grey region) envelop between 37 and 61% of the streamflow observations (red dots).

top, middle, and bottom plots display the results of the Oostanaula, French Broad, and Leaf River basins, respectively. The observed discharge data are indicated with red dots. The 95% streamflow uncertainty ranges appear rather narrow but nicely track the observed discharge data. About 40–60% of the discharge observations are contained in the 95% simulation intervals. The root-mean-square error (RMSE) of the ensemble mean simulation ranges between 0.8 and 1.2 mm/d. This value of the RMSE would be substantially lower if the SAC-SMA model was fitted directly against the observed discharge data using a *n*-variate Gaussian likelihood function [*Schoups and Vrugt*, 2010]. Yet this classical approach to model fitting leads to simulated summary metrics that deviate considerably from their observed counterparts in an effort of the model to compensate (optimally) for model structural and hydroclimatic forcing data errors (among others) [see also *Vrugt and Sadegh*, 2013]. Our diagnostics approach is aimed at correctly representing the signatures of the watershed observed in the discharge data—and hence the parameter values are carefully chosen to represent this behavior with the model.

We now illustrate our findings for three watersheds that are deemed to exhibit nonstationary behavior. Figure 10 presents trace plots of the sampled values of $\rho(\cdot)$ (computed from equation (5)) in each of the $K = 3$ different Markov chains simulated with DREAM$_{(ABC)}$ for the (a) Wights catchment, (b) "abrupt," and (c) "gradual" catchment. Each Markov chain is coded with a different color. As a reminder, the watersheds "abrupt" and "gradual" constitute synthetic streamflow data. "Abrupt" experiences a sudden change in the SAC-SMA model parameterization half way through its 10 year simulated record of daily discharge values. The second artificial catchment, called "gradual," experiences a daily linear adjustment of the SAC-SMA parameter values during year 5–6 of the 10 year data period. The dashed red line depicts the threshold value of ϵ=0.05 derived from Table 1 and used to demarcate stationarity.

The sampled distance values in each of the three Markov chains are substantially larger than the value of ϵ=0.05 required to accept the stationarity null hypothesis. The SAC-SMA model is unable to simulate sufficiently close, with time invariant parameter values, the observed values of each of the $L = 14$ summary metrics. In fact, the sampled distance values of $\rho(\cdot)$ are larger than 0.10 and thus satisfy the nonstationarity condition. The DREAM$_{(ABC)}$-derived samples are therefore considered nonbehavioral and discarded for posterior inference (the posterior distribution is unbounded). If time variant parameter values were used then, in principle, much lower values of $\rho(\cdot)$ could be attained, pending the assumption that the temporal trend

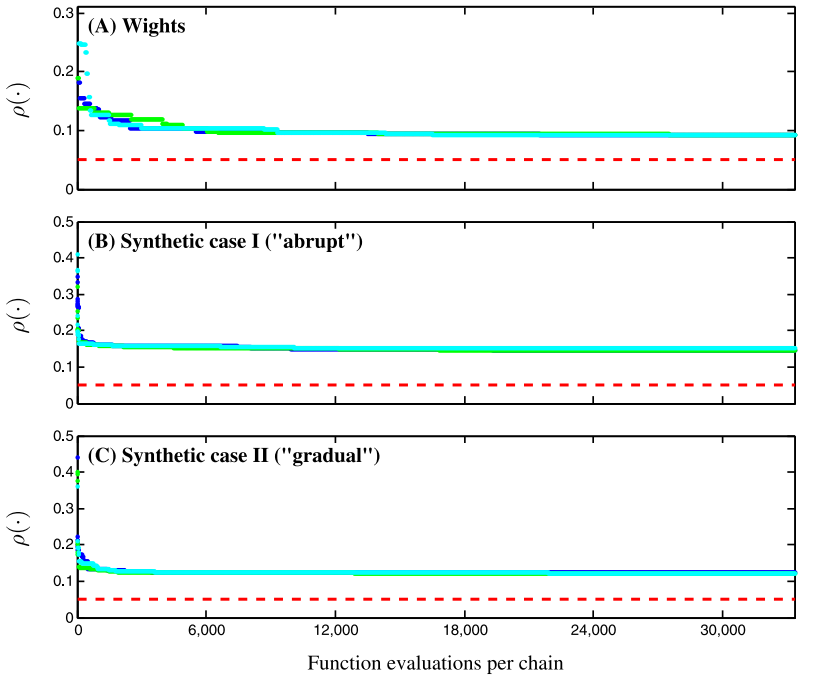

Figure 10. Trace plots of the sampled distance values, $\rho(\cdot)$, in each of the three Markov chains simulated with DREAM$_{(ABC)}$ for the (a) Wights, (b) "abrupt," and (c) "gradual" watersheds. The dashed red line depicts the threshold of ϵ=0.05 used to demarcate the stationarity null hypothesis. For each of the three watersheds, the SAC-SMA model is unable to satisfy the stationary condition, $\rho(\cdot) \leq 0.05$—in fact, the simulated values of $\rho(\cdot) \geq 0.10$ illuminate the presence of nonstationarity in the observed discharge data records.

of each parameter is described adequately. Altogether, our methodology confirms prior knowledge that the three watersheds have undergone changes.

Finally, Figure 11 plots histograms of a representative set of summary statistics (S_1, S_2, S_3, S_4, and S_{11}) simulated with the SAC-SMA model for each of the three nonstationary basins of Figure 10. The plotted distributions are derived from the last 20% of the samples in each of the three Markov chains simulated with DREAM$_{(ABC)}$. The observed summary metrics are indicated separately in each plot with a red cross, whereas the vertical dotted red lines demarcate the stationarity null hypothesis.

The histograms (in blue) of S_1, S_2, S_3, S_4, and S_{11} are made up of solutions that do not satisfy the stationary hypothesis, and hence cannot be called posterior marginal (behavioral) distributions. Instead, these distributions simply plot a frequency distribution of the last 20% of the samples of the $K = 3$ chain trajectories plotted in Figure 10 and in pursuit of $\epsilon \leq 0.05$. A particularly poor fit is observed for summary metric R_{11}, the RLD of year seven of the discharge data record. For all three watersheds, this metric cannot be simulated adequately with the SAC-SMA model assuming a single parameterization. In fact, a significant discrepancy is observed between the observed and fitted RLD (as measured in the remaining nine statistics) in several other years of the discharge data record (not shown). Thus, we conclude that the RLD is an important summary metric that can help diagnose watershed nonstationarity in response to urbanization, deforestation, and changes in hydroclimatic forcing (amongst others).

A value of $\rho(\cdot) > 0.10$ satisfies the nonstationarity condition but cannot convey the cause of the inhomogeneity observed in the discharge data record. Land use changes, urbanization, deforestation, and hydroclimatic variations (among others) can render the catchment response to precipitation (and other forcing variables) nonstationary. Only one of these driving variables needs to exhibit nonstationarity (can be a deterministic change) for the watersheds response to rainfall to appear nonstationary. Without further information, it is virtually impossible to determine from the present analysis whether the catchment characteristics have altered during the period of observation or whether the climate is time variant. Such separation requires separate analysis of the forcing data in a pursuit to answer, for each watershed individually, whether climate change is the culprit of the observed nonstationarity.

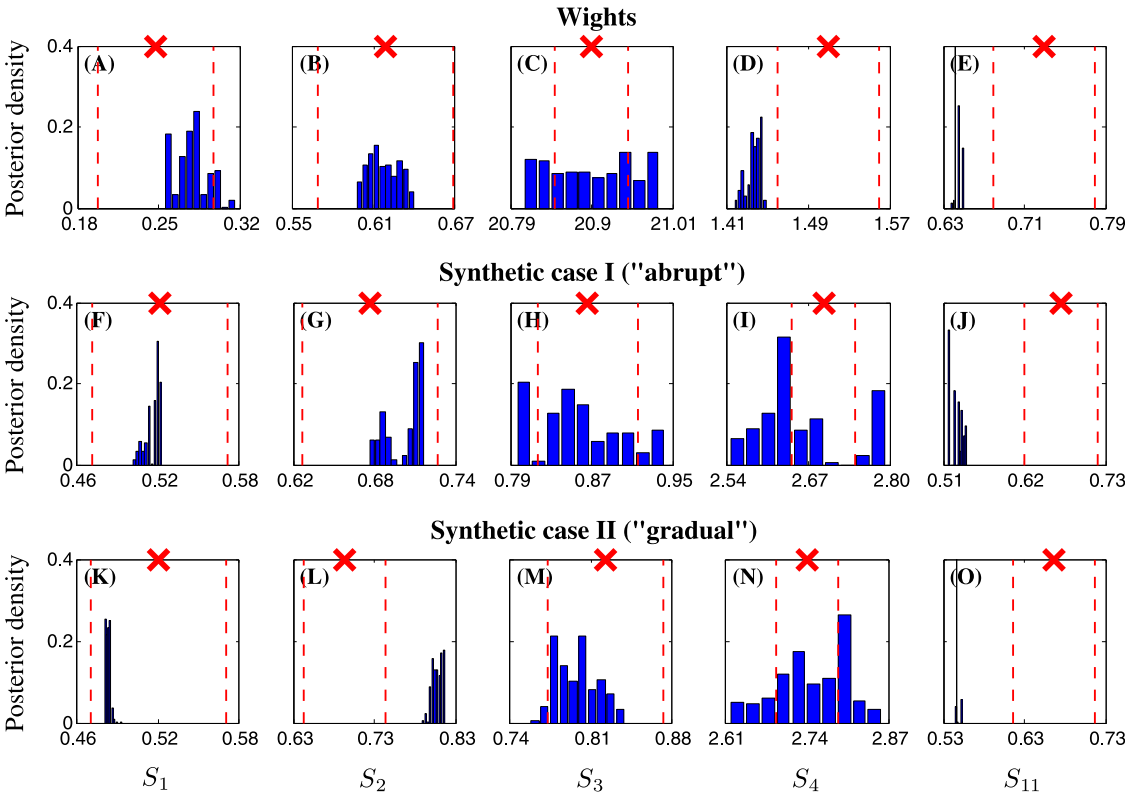

Figure 11. Histograms of the marginal posterior distribution of S_1: the runoff coefficient, S_2: base flow index, S_3, S_4: fitting coefficients of the flow duration curve, and S_5: the rising limb density of the seventh year of the calibration data record. We separately display the results for the (a) Wights, (b) "gradual," and (c) "abrupt" watersheds. The red cross "x" symbol marks the measured values of the summary metrics. A time invariant parameterization of the SAC-SMA model is unable to satisfy the stationarity hypothesis (demarcated with the vertically dashed red lines) for each of the three basins. This suggests that the watershed has experienced changes to its physical characteristics and/or hydroclimatic forcing during the period of observation.

The diagnostics methodology described herein provides hydrologists with a new methodology for hypothesis testing and analysis of stationarity. We posit that this methodology has several advantages over classical Frequentist and Bayesian inference methods that use purely statistical metrics to summarize (quantify) model adequacy [e.g., *Schoups and Vrugt*, 2010]. Such likelihood based metrics have little correspondence with the underlying hydrologic processes that control watershed behavior [*Gupta et al.*, 2008; *Vrugt and Sadegh*, 2013; *Sadegh and Vrugt*, 2014]. What is more these methods act on the error residuals of the observed and simulated discharge data which makes the inference very sensitive to (among others) forcing data errors. On the contrary, the summary statistics we used herein are far less sensitive to forcing data errors, and are much better grounded in hydrologic theory. What is more, summary metrics can be carefully designed to extract different and complementary parts of watershed behavior. When such metrics are used within the diagnostics framework of *Vrugt and Sadegh*

Table 4. Comparison of the Minimum Values of Equation (5) Derived From the SAC-SMA Model and HYMOD

Name	$\rho_{\min}(\cdot)$ SAC-SMA	$\rho_{\min}(\cdot)$ HYMOD
Stationary Watersheds		
Oostanaula	0.028	0.086
Pearl	0.043	0.191
Bogue Chitto	0.035	0.059
Little Pigeon	0.036	0.065
Tangipahoa	0.022	0.039
Comite	0.027	0.184
Calcasieu	0.045	0.084
S. Umpqua	0.022	0.039
French Broad	0.016	0.046
Leaf River	0.022	0.036
Nonstationary Watersheds		
Axe Creek	0.171	0.430
Wimmera	0.067	0.275
Wights	0.102	0.101
Flinders	0.253	0.249
Gilbert	0.168	0.188
Ferson	0.031	0.110
Blackberry	0.031	0.092
Synthetic Case I ("abrupt")	0.122	0.152
Synthetic Case II ("gradual")	0.144	0.128

[2013] then DREAM$_{(ABC)}$ [*Sadegh and Vrugt*, 2014] can be used for hypothesis testing of the stationary paradigm. For instance, if nonstationary is present in a discharge data record then one would expect at least one of the summary metrics to exhibit temporal variations. This behavior cannot be simulated adequately with temporally invariant model parameters. The minimum achievable distance between the observed and simulated metrics should therefore be a good proxy for the degree of catchment nonstationarity. For the SAC-SMA model, this purports to a value of $\rho(\cdot) \geq 0.10$ of equation (5) to satisfy nonstationarity.

To verify the main conclusions of this paper, we repeat the analysis but now using a more parsimonious watershed model. This model, called HYMOD, was developed by *Boyle* [2001] and has a much lower structural complexity (five parameters) than the SAC-SMA model to simulate the catchment response to rainfall. The results of our analysis are presented in Table 4 which lists the minimum values of equation (5) for each of the watersheds used herein. For completeness, we also summarize separately the SAC-SMA values of $\rho_{\min}(\cdot)$ of Table 1. The main conclusions are as follows.

1. The minimum values of $\rho(\cdot)$ of HYMOD are substantially larger than their counterparts of the SAC-SMA model. This confirms our earlier hypothesis that watershed models with lower structural complexity (fewer parameters) than SAC-SMA will generally have a larger discrepancy to the observed summary statistics of the discharge data. The exception to this are the Wights and Flinders watersheds in Australia for which both watersheds models receive similar values of $\rho(\cdot)$ of about 0.10 and 0.25, respectively. The HYMOD model appears to be of sufficient complexity for these two watersheds.
2. Watersheds classified as nonstationary exhibit the largest HYMOD values of $\rho(\cdot)$. This confirms another of our hypotheses that the value of $\rho(\cdot)$ is a useful proxy for catchment stationarity/nonstationarity. For values of HYMOD of $\rho(\cdot) \leq 0.09$, we can designate the behavior of the watershed as stationary, whereas values of $\rho(\cdot) > 0.12$ suggest the presence of nonstationarity in the observed discharge data. For values of $\rho(\cdot) \in [0.09, 0.12]$, the null hypothesis is rejected but the level of nonstationarity is insufficient to accept the alternative hypothesis.
3. The Ferson and Blackberry River basins in Illinois, USA, are classified as stationary watersheds with the SAC-SMA model, whereas HYMOD proclaims these two watersheds to exhibit some level of nonstationarity, but insufficient enough to reject formally the stationarity null hypothesis.
4. The HYMOD model designates the Pearl and Comite watersheds in the U.S. as nonstationary. This classification is erroneous, and explained by an inability of the model to represent adequately the observed discharge dynamics. Stationarity hypothesis testing thus requires the use of a model that can mimic sufficiently close the observed watershed behavior.

The results of HYMOD substantiate the main findings of this paper, and provide further support for our claims. The value of $\rho(\cdot)$ is a useful proxy for catchment stationarity. The larger the value of this diagnostic, the more likely that the watershed has experienced changes during the period of observation. Note the use of an ensemble of models can have several practical advantages for stationarity hypothesis testing. If all the different watershed models are in agreement in their assessment of stationarity (nonstationarity) then this is very strong evidence that the discharge record is indeed homogenous (inhomogeneous). Those data records for which the different watershed models of the ensemble differ in their assessment deserve careful further analysis. In all this work, it is of importance that the watershed model (or ensemble of models) is able to mimic reasonably well the observed response, otherwise epistemic errors play too large of a role and the inference becomes rather meaningless. For instance, a stationary basin may be erroneously classified as nonstationary if processes such as snowmelt are not adequately described.

A few final remarks are appropriate. In this work, we have assumed a single value of ϵ that is used for each different summary metric. This approach is defensible in the present context as the summary metrics have a somewhat similar magnitude and variation. Care should be exercised if metrics are used with disparate scales. Hence, it would seem logical to use a different value of ϵ for each individual summary metric. Temporal analysis of each summary metric can help to discern suitable values of ϵ for each individual metric. Alternatively, one could analyze the variability of each summary metric for watersheds with similar climate, soil, and topographic properties. We leave this for future work.

The summary metrics used herein have shown to exhibit the necessary diagnostic power to help detect catchment nonstationarity. The analysis of catchment nonstationarity would be much more difficult if a formal prior distribution and likelihood function were used. The model residuals would not only be diluted but

also lack obvious patterns and/or shifts that are easy to pinpoint to process nonstationarity. More work is required to provide further support for some of our claims made regarding the advantages of summary metrics for hypothesis testing and inference of epistemic errors. Whatever the verdict, formal likelihoods will always play a key role in statistical inference.

Lastly, most catchments will exhibit some level of nonstationarity. Climate change has modified the hydrological cycle globally, and in the Anthropocene, most catchments have experienced some level of change. The inference methodology we have presented not only helps to detect nonstationarity in the data, but also provides guidance on whether this nonstationarity is small or large. The $\rho(\cdot)$ value of equation (5) conveys the "size" of the nonstationarity. Of course, the analysis of change requires the use of a very long record of discharge data, certainly longer than used herein.

6. Summary and Conclusions

Many watershed models used within the hydrologic research community assume (by default) stationary conditions, that is, the key watershed properties that control water flow are considered to be time invariant. This assumption is rather convenient and pragmatic and opens up the wide arsenal of (multivariate) statistical and nonlinear optimization methods for inference of the (temporally fixed) model parameters from (for instance) a historical record of discharge data emanating from the catchment outlet. Evidence of an intensifying hydrologic cycle and the presence of trends (shift points) in long-term records of discharge data have brought into question the continued usefulness of this stationary paradigm for hydrologic modeling. The alternative hypothesis or negation of stationarity, i.e., nonstationarity, assumes the presence of trends in hydroclimatic variables. Several authors cast doubt on the validity of claims of nonstationarity, in particular, because climate change and other long-term atmospheric disturbances can explain a large part of the observed trends and variations in multidecadal streamflow observations. This long-term persistency of the climate and weather is not readily apparent in short hydrologic data sets that span only 10–20 years. It would be highly desirable to have available a methodology that can determine whether the observed catchment behavior is considered stationary or nonstationary. Ideally, such approach would also distinguish between the reasons of catchment nonstationarity, and determine whether the trends in streamflow dynamics are explained by decadal variations in hydroclimate or whether the physical characteristics of the catchment have changed as a result of (among others) anthropogenic influences.

This paper has built on the likelihood-free diagnostics approach of *Vrugt and Sadegh* [2013] and has introduced a methodology for stationarity hypothesis testing. This methodology uses a diverse set of hydrologic summary metrics to detect gradual/abrupt changes in the watersheds response to rainfall. If the stationarity assumption is valid then one expects the summary metrics to vary around some constant mean, and temporally invariant parameter values to be sufficient to mimic adequately the observed catchment response. The alternative hypothesis would involve temporal variant summary statistics, symptomatic for catchments subject to physical alterations (urbanization, deforestation, and/or changes in land use) or changes in hydroclimate (changing rainfall characteristics).

Numerical simulations with artificial discharge data of the SAC-SMA model have identified the runoff coefficient, base flow index, two fitting coefficients of the flow duration curve, and annual rising limb density to be most suitable for testing of the stationarity hypothesis. These summary statistics of the discharge data were selected from a large group of hydrologic metrics and appear independent and relatively insensitive to rainfall data errors. These constitute two important criteria to avoid, among others, proclaiming nonstationarity based on the wrong reasons.

The analysis and findings presented in this paper confirm our hypothesis that summary metrics of catchment behavior convey important information about hydrologic functioning, and that temporal variations of these fingerprints elucidates the presence of nonstationarity. Based on a suite of 19 different watersheds, 10 of which are classified in the hydrologic literature as stationary, and 9 of which are deemed nonstationary, we demonstrate that a value of $\rho(\cdot) \leq 0.05$ is a necessary condition for stationarity when the SAC-SMA model is used to simulate the rainfall-runoff transformation. On the contrary, a SAC-SMA value of $\rho(\cdot) > 0.10$ is required to proclaim nonstationarity. This simple distance criterion cannot convey the cause of the inhomogeneity observed in the discharge data record. Land use changes, urbanization, deforestation, and hydroclimatic variations (among others) can render the catchment response to precipitation (and other forcing

variables) nonstationary. For values of $\rho(\cdot) \in [0.05, 0.10]$, the null hypothesis (stationarity) is rejected but the evidence is not strong enough to support with high confidence the nonstationary hypothesis.

The suggestion made in the hydrologic community that the Ferson and Blackberry Creek basins in the U.S. exhibit nonstationary behavior is not confirmed by our diagnostic analysis with DREAM$_{(ABC)}$. The summary metrics of these watersheds appear rather constant over the duration of the 10 year data period, and the observed trends in streamflow after 2006 is due to an increase in precipitation. From all summary metrics, the annual rising limb density (RLD) appears to be most sensitive to changes to the physical characteristics of the watershed and/or hydroclimatic variations. Thus, the RLD is a good proxy for watershed nonstationarity.

In summary, the diagnostics methodology described herein provides hydrologists with a new methodology for hypothesis testing and analysis of stationarity. This framework is easy to use in practice and readily employs a wide variety of process-based metrics to capture different signatures of watershed behavior. Of course, additional testing of the proposed methodology against a much larger set of watersheds is required to further benchmark and refine our findings. Moreover, our binary classification of (non)stationarity should be replaced with a probabilistic interpretation of $\rho(\cdot)$. This would provide an estimate of the degree of nonstationarity of each watershed. Also, it would seem logical to use a different value of ϵ for each individual summary metric. In all this work, it is of importance that the watershed model (or ensemble of models) is able to mimic reasonably well the observed response, otherwise epistemic errors play too large of a role and the inference becomes rather meaningless. Finally, due to a lack of data for some watersheds, we base our arguments and computations in this paper on a rather short 10 year record of daily streamflow observations. We would strongly advise to use a longer period of data in future studies. We have made some suggestions how to tackle some of these problems and leave this for future work.

Acknowledgments
The comments of Seth Westra and three anonymous reviewers are greatly appreciated and have helped to significantly enhance the current manuscript. The first and second author appreciate the support and funding from the UC-Lab Fees Research Program award 237825. The contribution of the last author was made possible by the Italian Ministry of University and Research through project PRIN 20102AXKAJ. We would also like to thank Guillaume Thirel and George Kuczera for kindly sharing with us the data of the watersheds used herein. The data of these watersheds can be found at ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/ and http://non-stationarities.irstea.fr/.

References

Alexander, L. V., et al. (2006), Global observed changes in daily climate extremes of temperature and precipitation, *J. Geophys. Res.*, *111*, D05109, doi:10.1029/2005JD006290.
Anghileri, D., F. Pianosi, and R. Soncini-Sessa (2014), Trend detection in seasonal data: From hydrology to water resources, *J. Hydrol.*, *511*, 171–179.
Bassiouni, M., and D. S. Oki (2013), Trends and shifts in streamflow in Hawaii, 1913-2008, *Hydrol. Processes*, *27*(10), 1484–1500.
Beaumont, M. A., W. Zhang, and D. J. Balding (2002), Approximate Bayesian computation in population genetics, *Genetics*, *162*(4), 2025–2035.
Birsan, M. V., P. Molnar, P. Burlando, M. Pfaundler (2005), Streamflow trends in Switzerland, *J. Hydrol.*, *314*, 312–329.
Boyle, D. (2001), Multicriteria calibration of hydrological models, PhD dissertation, Univ. of Arizona, Tucson.
Buishand, T. A. (1982), Some methods for testing the homogeneity of rainfall records, *J. Hydrol.*, *58*, 11–27.
Burn, D. H., and M. A. Hag Elnur (2002), Detection of hydrologic trends and variability, *J. Hydrol.*, *255*, 107–122.
Burnash, R. J., R. L. Ferral, and R. A. McGuire (1973), *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers*, Joint Fed.-State River Forecast Cent., Sacramento, Calif.
Chebana, F., T. B. M. J. Ouarda, and T. C. Duong (2013), Testing for multivariate trends in hydrologic frequency analysis, *J. Hydrol.*, *486*, 519–530.
Clarke, R. T. (2002), Estimating trends in data from the Weibull and a generalized extreme value distribution, *Water Resour. Res.*, *38*(6), doi: 10.1029/2001WR000575.
Clarke, R. T. (2007), Hydrological prediction in a non-stationary world, *Hydrol. Earth Syst. Sci.*, *11*, 408–414, doi:10.5194/hess-11-408-2007.
Clausen B., and B. J. F. Biggs (2000), Flow variables for ecological studies in temperate streams: Grouping based on covariance, *J. Hydrol.*, *237*(3-4), 184–197.
Cong, Z., D. Yang, B. Gao, H. Yang, and H. Hu (2009), Hydrological trend analysis in the Yellow River basin using a distributed hydrological model, *Water Resour. Res.*, *45*, W00A13, doi:10.1029/2008WR006852.
Cunderlik, J. M., and D. H. Burn (2003), Non-stationary pooled flood frequency analysis, *J. Hydrol.*, *276*, 210–223.
Cunderlik, J. M., and T. B. M. J. Ouarda (2006), Regional flood-duration-frequency modeling in a changing environment, *J. Hydrol.*, *318*, 276–291.
Cunnane, C. (1988), Methods and merits of regional flood frequency analysis, *J. Hydrol.*, *100*(1-3), 269–290.
Del Moral, P., A. Doucet, and A. Jasra (2012), An adaptive sequential Monte Carlo method for approximate Bayesian computation, *Stat. Comp.*, *22*(5), 1009–1020.
Dettinger, M. (2011), Climate change, atmospheric rivers, and floods in California—A multimodel analysis of storm frequency and magnitude changes, *J. Am. Water Resour. Assoc.*, *47*, 514–523.
Diggle, P. J., and R. J. Gratton (1984), Monte Carlo methods of inference for implicit statistical models, *J. R. Stat. Soc., Ser. B*, *46*, 193–227.
Douglas, E. M., R. M. Vogel, and C. N. Kroll (2000), Trends in floods and low flows in the United States: Impact of spatial correlation, *J. Hydrol.*, *240*(1-2), 90–105.
Eckhardt, K. (2005), How to construct recursive digital filters for baseflow separation, *Hydrol. Processes*, *19*, 507–515.
Edwards, A. W. F. (1992), *Likelihood*, The Johns Hopkins Univ. Press, Baltimore, Md.
El-Adlouni, S., T. B. M. J. Ouarda, X. Zhang, R. Roy, and B. Bobée (2007), Generalized maximum likelihood estimators for the nonstationary generalized extreme value model, *Water Resour. Res.*, *43*, W03410, doi:10.1029/2005WR004545.
Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije (2013), A framework to assess the realism of model structures using hydrological signatures, *Hydrol. Earth Syst. Sci.*, *17*, 1893–1912, doi:10.5194/hess-17-1893-2013.

Fu, G., S. Chen, C. Liu, and D. Shepard (2004), Hydro-climatic trends of the Yellow River basin for the last 50 years, *Clim. Change*, *65*(1-2), 149–178.
Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*, 457–472.
Graf, W. L. (1977), Network characteristics in suburbanizing streams, *Water Resour. Res.*, *13*(2), 459–463.
Groisman, P. Y., R. W. Knight, and T. R. Karl (2001), Heavy precipitation and high streamflow in the contiguous United States: Trends in the twentieth century, *Bull. Am. Meteorol. Soc.*, *82*, 219–246.
Groisman, P. Y., R. W. Knight, T. R. Karl, D. R. Easterling, B. Sun, and J. H. Lawrimore (2004), Contemporary changes of the hydrological cycle over the contiguous United States: Trends derived from in situ observations, *J. Hydrometeorol.*, *5*, 64–85.
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813.
Hodgkins, G. A., and R. W. Dudley (2006), Changes in the timing of Winter-Spring streamflows in Eastern North America, 1913-2002, *Geophys. Res. Lett.*, *33*, L06402, doi:10.1029/2005GL025593.
Hurst, H. E. (1951), Long term storage capacities of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 776–808.
Ishak, E. H., A. Rahman, S. Westra, A. Sharma, and G. Kuczera (2013), Evaluating the non-stationarity of Australian annual maximum flood, *J. Hydrol.*, *494*, 134–145.
Jain, S., and U. Lall (2001), Floods in a changing climate: Does the past represent the future?, *Water Resour. Res.*, *37*(12), 3193–3205, doi: 10.1029/2001WR000495.
Joshi, M. K., and A. C. Pandey (2011), Trend and spectral analysis of rainfall over India during 1901-2000, *J. Geophys. Res.*, *116*, D06104, doi: 10.1029/2010JD014966.
Karl, T. R., and R. W. Knight (1998), Secular trends of precipitation amount, frequency, and intensity in the United States, *Bull. Am. Meteorol. Soc.*, *79*(2), 231–241.
Karthikeyan, L., and D. Nagesh Kumar (2013), Predictability of nonstationary time series using wavelet and EMD based ARMA models, *J. Hydrol.*, *502*, 103–119.
Khaliq, M. N., T. B. M. J. Ouarda, P. Gachon, L. Sushama, and A. St-Hilaire (2009), Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers, *J. Hydrol.*, *368*, 117–130.
Koutsoyiannis, D. (2006), Nonstationarity versus scaling in hydrology, *J. Hydrol.*, *324*, 239–254.
Koutsoyiannis, D. (2011), Hurst-Kolmogorov dynamics and uncertainty, *J. Am. Water Resour. Assoc.*, *47*(3), 481–495.
Koutsoyiannis, D. (2013), Hydrology and change, *Hydrol. Sci. J.*, *58*(6), 1177–1197, doi:10.1080/02626667.2013.804626.
Koutsoyiannis, D., and A. Montanari (2014), Negligent killing of scientific concepts: The stationarity case, *Hydrol. Sci. J.*, *60*(7–8), 1174–1183, doi:10.1080/02626667.2014.959959.
Kundzewicz, Z. W. (2011), Nonstationarity in water resources—Central European perspective, *J. Am. Water Resour. Assoc.*, *47*, 550–562.
Kundzewicz, Z. W., and A. J. Robson (2004), Change detection in hydrological records: A review of the methodology, *Hydrol. Sci.*, *49*, 7–19.
Kundzewicz, Z. W., D. Graczyk, T. Maurer, I. Pińskwar, M. Radziejewski, C. Svensson, and M. Szwed (2005), Trend detection in river flow series: 1. Annual maximum flow, *Hydrol. Sci. J.*, *50*(5), 797–810, doi:10.1623/hysj.2005.50.5.797.
Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try $\text{DREAM}_{(ZS)}$ and high-performance computing, *Water Resour. Res.*, *48*, W01526, doi:10.1029/2011WR010608.
Leclerc, M., and T. B. M. J. Ouarda (2007), Non-stationary regional flood frequency analysis at ungauged sites, *J. Hydrol.*, *343*, 254–265.
Lettenmaier, D. P., E. F. Wood, and J. R. Wallis (1994), Hydroclimatological trends in the continental United States, 1948–1988, *J. Clim.*, *7*, 586–607.
Lima, C. H. R., and U. Lall (2010), Spatial scaling in a changing climate: A hierarchical Bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *J. Hydrol.*, *383*(3-4), 307–318.
Lins, H. F. (1985), Streamflow variability in the United States: 1931–78, *J. Clim. Appl. Meteorol.*, *24*, 463–471.
Lins, H. F., and T. A. Cohn (2011), Stationarity: Wanted dead or alive?, *J. Am. Water Resour. Assoc.*, *47*(3), 475–480.
Lins, H. F., and J. R. Slack (1999), Streamflow trends in the United States, *Geophys. Res. Lett.*, *26*(2), 227–230.
Lins, H. F., and J. R. Slack (2005), Seasonal and regional characteristics of U.S. streamflow trends in the United States from 1940–1999, *Phys. Geogr.*, *26*, 489–501.
Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003), Markov chain Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *100*(26), 15,324–15,328.
McCabe, G. J., and D. M. Wolock (2002), A step increase in streamflow in the coterminous United States, *Geophys. Res. Lett.*, *29*(24), 2185, doi:10.1029/2002GL015999.
Milly, P. C. D., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer (2008), Stationarity is dead: Whiter water management?, *Science*, *319*, 573–574.
Morin, E., K. P. Georgakakos, U. Shamir, R. Garti, and Y. Enzel (2002), Objective, observations-based, automatic estimation of the catchment response timescale, *Water Resour. Res.*, *38*(10), 1212, doi:10.1029/2001WR000808.
Mroczkowski, M., G. P. Raper, and G. Kuczera (1997), The quest for more powerful validation of conceptual catchment models, *Water Resour. Res.*, *33*(10), 2325–2335.
Nasri, B., S. El-Adlouni, and T. B. M. J. Ouarda (2013), Bayesian estimation for GEV-B-Spline model, *Open J. Stat.*, *3*(2), 118–128.
Ni, C. F., C. P. Lin, S. G. Li, and C. J. Liu (2011), Efficient approximate spectral method to delinate stochastic well capture zones in nonstationary groundwater flow systems, *J. Hydrol.*, *407*(1–4), 184–195.
North, M. (1980), Time-dependent stochastic model of floods, *J. Hydraul. Div.*, *106*(5), 649–665.
Novotny, E. V., and H. G. Stefan (2007), Stream flow in Minnesota: Indicator of climate change, *J. Hydrol.*, *334*, 319–333.
Olden, J. D., and N. L. Poff (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, *River Res. Appl.*, *19*, 101–121.
Ouarda, T. B. M. J., and S. El-Adlouni (2011), Bayesian nonstationary frequency analysis of hydrological variables, *J. Am. Water Resour. Assoc.*, *47*(3), 496–505.
Parey, S., F. Malek, C. Laurent, and D. Dacunha-Castelle (2007), Trends and climate evolution: Statistical approach for very high temperatures in France, *Clim. Change*, *81*(3-4), 331–352.
Petrow, T., and B. Merz (2009), Trends in flood magnitude, frequency and seasonality in Germany in the period 1951–2002, *J. Hydrol.*, *371*, 129–141.
Pitman, W. V. (1978), Trends in streamflow due to upstream land-use changes, *J. Hydrol.*, *39*, 227–237.

Potter, K. W. (1991), Hydrological impacts of changing land management practices in a moderate-sized agricultural catchment, *Water Resour. Res.*, *27*(5), 845–855.
Price, K. V., R. M. Storn, and J. A. Lampinen (2005), *Differential Evolution: A Practical Approach to Global Optimization*, Springer, Berlin.
Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. T. Feldman (1999), Population growth of human Y chromosomes: A study of Y chromosome microsatellites, *Mol. Biol. Evol.*, *16*(12), 1791–1798.
Ramachandra Rao, A., and G. H. Yu (1986), Detection of nonstationarity in hydrologic time series, *Manage. Sci.*, *32*(9), 1206–1217.
Renard, B., M. Lang, and P. Bois (2006), Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: Case study with peak-over-threshold data, *Stochastic Environ. Res. Risk Assess.*, *21*, 97–112.
Rougé, C., Y. Gea, and X. Cai (2013), Detecting gradual and abrupt changes in hydrological records, *Adv. Water Res.*, *53*, 33–44.
Sadegh, M., and J. A. Vrugt (2013), Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation, *Hydrol. Earth Syst. Sci.*, *17*, 4831–4850, doi:10.5194/hess-17-4831-2013.
Sadegh, M., and J. A. Vrugt (2014), Approximate Bayesian computation using Markov Chain Monte Carlo simulation: $\text{DREAM}_{(\text{ABC})}$, *Water Resour. Res.*, *50*, 6767–6787, doi:10.1002/2014WR015386.
Salas, J. D., and J. Obeysekera (2014), Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events, *J. Hydrol. Eng.*, *19*(3), 554–568.
Savenije, H. H. G. (1996), The runoff coefficient as the key to moisture recycling, *J. Hydrol.*, *176*, 219–225.
Schaake, J., S. Cong, and Q. Duan (2006), *U.S. MOPEX Data Set, IAHS Publ. Ser.*, Lawrence Livermore Natl. Lab., Livermore, Calif. [Available at http://goo.gl/wthGvS.]
Scharnagl, B., J. A. Vrugt, H. Vereecken, and M. Herbst (2011), Bayesian inverse modeling of soil water dynamics at the field scale: Using prior information about the soil hydraulic properties, *Hydrol. Earth Syst. Sci.*, *15*, 3043–3059, doi:10.5194/hess-15-3043-2011.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.
Searcy, J. K. (1959), Flow-duration curves, *U.S. Geol. Surv. Water Supply Pap., 1542-A*, 17–21.
Shamir, E., B. Imam, E. Morin, H. V. Gupta, and S. Sorooshian (2005), The role of hydrograph indices in parameter estimation of rainfall-runoff models, *Hydrol. Processes*, *19*, 2187–2207.
Sisson, S. A., Y. Fan, and M. M. Tanaka (2007), Sequential Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(6), 1760–1765.
Smith, J. A., M. L. Baeck, J. E. Morrison, P. Sturdevant-Rees, D. F. Turner-Gillespie, and P. D. Bates (2002), The regional hydrology of extreme floods in an urbanizing drainage, *J. Hydrometeorol.*, *3*(3), 267–282.
Stedinger, J. R., and V. W. Griffis (2011), Getting from here to where? Flood frequency analysis and climate, *J. Am. Water Resour. Assoc.*, *47*, 506–513.
Storn, R., and K. Price (1997), Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces, *J. Global Optim.*, *11*, 341–359.
Strupczewski, W. G., V. P. Singh, and H. T. Mitosek (2001), Non-stationary approach to at-site flood frequency modelling. III. Flood analysis of Polish rivers, *J. Hydrol.*, *248*(1-4), 152–167.
Svensson, C., W. Z. Kundzewicz, and T. Maurer (2005), Trend detection in river flow series: 2. Flood and low-flow index series, *Hydrol. Sci. J.*, *50*(5), 811–824.
Thirel, G., et al. (2015), Hydrology under change: An evaluation protocol to investigate how hydrological models deal with changing catchments, *Hydrol. Sci. J.*, *60*, 1184–1199, doi:10.1080/02626667.2014.967248.
Turner, B. M., and T. van Zandt (2012), A tutorial on approximate Bayesian computation, *J. Math. Psychol.*, *56*, 69–85.
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, *J. Hydrol.*, *394*, 447–457.
Villarini, G., and J. A. Smith (2010), Flood peak distributions for the eastern United States, *Water Resour. Res.*, *46*, W06504, doi:10.1029/2009WR008395.
Villarini, G., F. Serinaldi, J. A. Smith, and W. F. Krajewski (2009), On the stationarity of annual flood peaks in the continental United States during the 20th century, *Water Resour. Res.*, *45*, W08417, doi:10.1029/2008WR007645.
Vogel, R. M., C. Yaindl, and M. Walter (2011), Nonstationarity: Flood magnification and recurrence reduction factors in the United States, *J. Am. Water Resour. Assoc.*, *47*(3), 464–474.
Vrugt, J. A. (2015), Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, *Environ. Model. Software*, doi:10.13140/RG.2.1.4692.6569.
Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi:10.1002/wrcr.20354.
Vrugt, J. A., C. G. H. Diks, W. Bouten, H. V. Gupta, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.
Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.
Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, D. Higdon, B. A. Robinson, and J. M. Hyman (2009), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.
Waage, M. D., and L. Kaatz (2011), Nonstationary water planning: An overview of several promising planning methods, *J. Am. Water Resour. Assoc.*, *47*, 535–540.
Westerberg, I., J.-L. Guerrero, J. Seibert, K. J. Beven, and S. Halldin (2011), Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrol. Processes*, *25*, 603–613, doi:10.1002/hyp.7848.
Westmacott, J. R., and D. H. Burn (1997), Climate change effects on the hydrologic regime within the Churchill-Nelson River basin, *J. Hydrol.*, *202*(1-4), 263–279.
Westra, S., and S. A. Sisson (2011), Detection of non-stationarity in precipitation extremes using a max-stable process model, *J. Hydrol.*, *406*(1-2), 119–128.
Westra, S., L. V. Alexander, and F. W. Zwiers (2013), Global increasing trends in annual maximum daily precipitation, *J. Clim.*, *26*, 3904–3918.
Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014), A strategy for diagnosing and interpreting hydrologic non-stationarity, *Water Resour. Res. 50*, 5090–5113, doi:10.1002/2013WR014719.
Xu, C.-Y., L. Gong, T. Jiang, D. Chen, and V. P. Singh (2006), Analysis of spatial distribution and temporal trend of reference evapotranspiration and pan evaporation in Changjiang (Yangtze River) catchment, *J. Hydrol.*, *327*(1-2), 81–93.
Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*, 1756–1774.

Yue, S., P. Pilon, and G. Cavadias (2002), Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series, *J. Hydrol.*, *264*(1-4), 262–263.

Zhang, Q., C.-Y. Xu, S. Becker, Z. X. Zhang, Y. D. Chen, and M. Coulibaly (2009), Trends and abrupt changes of precipitation maxima in the Pearl River basin, China, *Atmos. Sci. Lett.*, *10*, 132–144, doi:10.1002/asl.221.

Zhang, X., K. D. Harvey, W. D. Hogg, and T. R. Yuzyk (2001), Trends in Canadian streamflow, *Water Resour. Res.*, *37*(4), 987–998.

Zhang, Y. K., and K. E. Schilling (2006), Increasing streamflow and baseflow in Mississippi River since the 1940s: Effect of land use change, *J. Hydrol.*, *324*, 412–422.