Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

Evaluating Hydrological Model Performance using Information Theory-based Metrics

This title is very general. Your innovation is not that you used information-based metrics in hydrological modelling. That has been done since the 1980s. Also after reading the article you do not aim to build on performance metrics rooted in info theory to enhance calibration but rather to select model structure based on results from info-based metrics.

Yakov A. Pachepsky[1], Gonzalo Martinez[2*], Feng Pan[3,4], Thorsten Wagener[5], Thomas Nicholson[6]

[1]USDA-ARS Environmental Microbial and Food Safety Laboratory, Beltsville, MD 20705, USA
[2]Department of Agronomy, University of Cordoba, 14071, Cordoba, Spain
[3]Department of Civil & Environmental Engineering, the University of Utah, Salt Lake City, UT 84112, USA
[4]Energy & Geoscience Institute, the University of Utah, Salt Lake City, UT 84108, USA
[5]Department of Civil Engineering, University of Bristol, Bristol, UK
[6]Office of Regulatory Research, US Nuclear Regulatory Commission, Rockville, MD 20852, USA

*Correspondence to: G. Martinez (z42magag@uco.es)

I personally find very confusing to classify "standard" metrics as accuracy-based. Note that relative entropy is also an accuracy metric, i.e. distance between pdfs

what does qualitative correspondence mean here?

Abstract. The accuracy-based model performance metrics not necessarily reflect the qualitative correspondence between simulated and measured streamflow time series. The objective of this work was to use the information theory-based metrics to see whether they can be used as complementary tool for hydrologic model evaluation and selection. We simulated 10-year streamflow time series in five watersheds located in Texas, North Carolina, Mississippi, and West Virginia. Eight model of different complexity were applied. The information theory based metrics were obtained after representing the time series as strings of symbols where different symbols corresponded to different quantiles of the probability distribution of streamflow. The symbol alphabet was used. Three metrics were computed for those strings – mean information gain that measures the randomness of the signal, effective measure complexity that characterizes predictability and fluctuation complexity that characterizes the presence of a pattern in the signal. The observed streamflow time series has smaller information content and larger complexity metrics than the precipitation time series. Watersheds served as information filters and and streamflow time series were less random and more complex than the ones of precipitation. This is reflected by the fact that the watershed acts as the information filter in the hydrologic conversion process from precipitation to streamflow. The Nash-Sutcliffe efficiency metric increased as the complexity of models increased, but in many cases several models had this efficiency values not statistically significant from each other. In such cases, ranking models by the closeness of the information theory based parameters in simulated and measured streamflow time series can provide an additional criterion for the evaluation of hydrologic model performance.

Delete because this is implied from the earlier sentence

It is mentioned above

what does that mean? Are you referring to the processes?

you mean watersheds?

Please rephrase since it is not clear!

* What does this mean? Are you referring to the physical processes that alter the precipitation signal?

1 Introduction

Hydrologic modeling plays the critical role in understanding the hydrologic response of a watershed, and hence is very important for the applications such as e.g. water resources management activities, flood control, and water quality evaluation (Singh and Woolhiser, 2002; Pechlivanidis et al., 2011,

** Solely by reading this abstract I[1] cannot point towards the new insights from this research. Information-based metrics have been used in characterising time series (1-D approach) and also as performance measure (2D approach). For the latter, info theory metrics have been used as single and multi-objective. Given all these, what is your contribution?

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

* Why are you referring to lumped / distributed models? This is not relevant to your investigation. In addition, there are hundreds of models, so there is no reason to only mention 6. Please remove such ~~widely known~~ ~~in fact~~ unnecessary details.

Wagener et al., 2010). Over the last few decades, lumped and physics-based distributed hydrologic models have been developed and widely applied to simulate the hydrologic processes for understanding of watershed behaviors. Lumped models are represented, for example, by Stanford Watershed Model (SWM) (Crawford and Linsley, 1966), the Tank Model (Sugawara et al., 1976), and Xinanjiang Model (Zhao et al., 1980) etc. With the rapid development of computational power, applications of distributed models have become feasible. The family of such models include Systeme Hydrologique Europeen (SHE) (Abbott et al., 1986a, b), Physically Based Runoff Production Model (TOPMODEL) (Beven and Kirkby, 1979), Soil Water Assessment Tool (SWAT) (Arnold et al., 1998), Hydrologic Model System (Yu et al., 1999), and Variable Infiltration Capacity (VIC) model (Liang et al., 1994). The evaluation of model performance is indispensable to examine both accuracy and reliability of models.

The common model evaluation metrics in hydrology include the Nash-Sutcliffe efficiency *NSE* (Nash and Sutcliffe, 1970; Krause et al., 2005; Bai et al., 2009), the root-mean-squared error, the coefficient of determination, the Akaike information criterion *AIC* (Akaike, 1973), the Bayesian information criterion *BIC* (Schwarz, 1978), and the Kashyap information criterion *KIC* (Kashyap, 1982). Recently, new approaches have been proposed to evaluate the performance of hydrologic models, such as maximum likelihood Bayesian model averaging *MLBMA* (Ye et al., 2004), a wavelet-based multiscale performance metric (Rathinasamy et al., 2014), a data-reduction method based on self-organizing maps (Reusser et al., 2009), an interval-deviation approach (Chen et al., 2014), and a top-down methodology (Bai et al., 2009) among others. Although these metrics/approaches can evaluate the correspondence between the simulation results and observed data, they cannot capture all the features reproduced by the hydrologic models such as information content of data and model complexity under uncertainty (Gupta et al., 1998; Reusser et al., 2009; Pachepsky et al., 2006; Weijs et al., 2010). Information theory has been recently applied to develop additional metrics to characterize the patterns of observed and simulated data sets to provide the insights and complementary knowledge on the evaluation of model performances (Pachepsky et al., 2006; Pan et al., 2011, 2012; Li et al., 2012; Gong et al., 2013; Pechlivanidis et al., 2014; Beven and Smith, 2015). The predictive performance of hydrologic models was evaluated by fully exploiting the available information in the data set using the information-based indices (Gong et al., 2013). Li et al. (2012) proposed an entropy-based criterion named maximum information minimum redundancy (MIMR) to evaluate and optimize the design of the hydrometric

* I find this definition of 'common' metrics very subjective. To me, it is not about the metric itself but rather its nature, i.e. metrics based on the residuals between the modelled and observed data. I would suggest you to rephrase accordingly.

** I think here you mix messages. You have started by discussing performance metrics and then you mix those with how to identify/average model structures.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences Open Access EGU
Discussions

networks. The information theory has also been applied in the calibration of hydrologic models to improve model performance (Pechlivanidis et al, 2014; Beven and Smith, 2015). The complexity and information content metrics have been employed by Pachepsky et al. (2006) to discriminate the different soil water flow models that gave the same accuracy of soil water flux estimates, and by Pan et al. (2011) to evaluate the ability of the model to reproduce the temporal trends of soil moisture content in variably saturated soil.

The objectives of this study are (1) to characterize the patterns of observed precipitation and streamflow time series in arid and humid watersheds; (2) to evaluate the performance of eight hydrologic models in five watersheds using complexity and information content metrics and (3) to compare the results of this performance evaluation with the results of performance evaluation based on the Nash-Sutcliffe efficiency metrics. The eight hydrologic model structures have been developed by Bai et al. (2009) including two evapotranspiration modules, four soil moisture accounting modules, and three routing modules. The details of model structure are referred to Bai et al. (2009). The five watersheds selected in this study include two dry watersheds, Guadalupe River and San Marcos River catchments in Texas, and three wet watersheds, Tygart Valley River in West Virginia, French Broad River in North Carolina, and Leaf River in Mississippi.

2. MATERIALS AND METHODS

2.1 Study Sites

The five watersheds were selected in Texas, North Carolina, Mississippi, and West Virginia to represent a range of hydro-climatic conditions. The eleven-year data (1960-1970) of daily precipitation (P), streamflow (Q) and potential evapotranspiration (PE) in the five watersheds were used in this study. The characteristics of the five watersheds are listed in Table 1.

The Guadalupe River and San Marcos River catchments located in Texas are two dry watersheds with mean annual precipitation of around 800 mm and mean annual PE of 1500 mm. Tygart Valley River in West Virginia, French Broad River in North Carolina, and Leaf River in Mississippi are three wet watersheds with mean annual precipitation of about 1300 mm and mean annual PE of around 800-1000 mm. More detailed information of the watersheds can be found in Bai et al. (2009).

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

2.2 Hydrologic Modeling

The eight hydrologic model structures have been selected to represent differences in hydrologic model complexity for the model evaluation with different metrics. The eight models, which are briefly described in Table 2, were derived from the different combination of three modules for soil moisture accounting, actual evapotranspiration, and routing (Bai et al., 2009). Models S1 and M1 estimated streamflow as a surface runoff resulting from the saturation excess, models S2 and M2 added subsurface flow to the streams appearing after soil reached filed capacity, models S3 and M3 added subsurface flow from saturated zone, and models S4 and M4 added the deep storage recharge. The difference between S models and M models consisted in the treatment of soil moisture accounting. S models used the single-layer models (Atkinson et al., 2002; Farmer et al., 2003), and M models used the multi-layer formulation (Son and Sivapalan, 2007). The ET module included two options with the estimation from the moisture storage as one zone, and from the unsaturated zone and shallow saturated zone (Bai et al., 2009). The routing modules were deployed to simulate flow release from storages (e.g., saturated zone, deep storage). The eight models were formed with the combination of the three modules with the increase in complexity (Bai et al., 2009). The streamflow in the five watersheds was simulated with each of eight models for ten years. The Nash-Sutcliffe efficiency index (NSE, Nash and Sutcliffe, 1970) was used as the model performance metric.

2.3 Information Content and Complexity Metrics

The general idea of information theory-based metrics in this work is to

(a) replace the time series by the string of symbols from some (small) alphabet; each letter denotes a particular range within the overall range of data variation

(b) define the number of points in the data window; for each data window, the replacement of numerical data with letters creates the word;

(c) research probabilities of changes in words as the data window moves over the time series;

(d) derive metrics of information content and complexity based on those probabilities

We represented the time series of hydrologic state variables (e.g., observed and modeled precipitation and streamflow in this study) as symbolic strings following Lange (1999) and Wolf (1999) methodologies. To do so, we chose a binary encoding using the median value of each state variable as a threshold; all the observations above the threshold were coded as one and

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

all the observations at the median value or below were coded as zero. The alphabet, therefore, had two letters – '0' and '1'. Both measured and simulated time series were encoded. Within the encoded strings we could analyze words of length L ($L \in \mathbb{N}$) composed of L consecutive symbols. Assuming that each word characterizes the state of the studied system, we have 2^L different words or states; the base '2' in this equation corresponds to the number of letters in the alphabet. For the binary encoding, we have the four (2^2) different words 11, 10, 01, 10. The first word shows the state in which the variable exceeds the median value at both times in the data window, the second word shows the transition from that state (11) to that in which the second observation falls below the median value (10), etc. For any particular string, we can compute various empirical probabilities to the occurrence and transition of states for words of length L such as:

$p_{L,i}$ probability for the word "i" to appear in the symbolic string

$p_{L,ij}$ probability for the sequence of words "i" and "j" to appear

$p_{L,i\to j}$ conditional probability of the occurrence of the j^{th} word after i^{th} word

After defining this set of probabilities we can compute two information-based metrics, namely as the metric entropy and mean information gain. The metric entropy (ME), is a normalized version of Shannon's entropy (H, Shannon, 1948):

$$ME = \frac{H(L)}{L}$$

where

$$H(L) = -\sum_{i=1}^{2^L} p_{L,i} \log_2 p_{L,i}\,, \qquad (2)$$

Shannon's entropy is a measure in bits of the average information content per code or unpredictability of the information contained in the time series. Its normalized version, *ME*, gives a measure independent of the word length. While it has a value of zero for constant strings it increases with the randomness of the string up to a value of 1 for uniformly random sequences.

The mean information gain (MIG), measures the average amount of new information obtained by knowing the next symbol. Given that the *MIG* includes the transition probability and the occurrence of the sequence of words, knowing the symbol that follows a word increases the local information. Therefore, the larger the MIG is the less predictable and more random is the time series.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

$$\text{[illegible]} = \sum_{i,j=1}^{n} \text{[illegible]}_{ij} \log_2 \text{[illegible]}_{ij}, \quad (3)$$

The complexity in the time series under study was assessed with the fluctuation complexity (FC) measure and the effective measure of complexity (EMC, Eq. 5). These two metrics allowed us to quantify the internal structure and the presence of patterns in the encoded symbolic strings.

$$\text{[illegible]} = \sum_{i,j=1}^{n} \text{[illegible]}_{ij} \left(\log_2 \frac{\text{[illegible]}_{ij}}{\text{[illegible]}_{ij}}\right)^2, \quad (4)$$

$$\text{[illegible]} = \sum_{i,j=1}^{n} \text{[illegible]}_{ij} \log_2 \frac{\text{[illegible]}_{ij}}{\text{[illegible]}_{ij}}, \quad (5)$$

The fluctuation complexity considers vaguely the ordering of, and relationship between, words in a sequence. It is obtained as the mean square deviation of the differences between information gained associated with the transition from the state "i" to the state "j" and the information lost associated with that transition. Strings that show a high degree of fluctuation in their symbols give larger fluctuation complexity values (Bates and Shepard, 1993). Grassberger (1986) defined the effective measure complexity (EMC) as "the minimal information that would have to be stored for optimal predictions if it could be used with 100% efficiency". Time series of random data or periodic sequences present are simple and show low values of FC and EMC. On the contrary, time series that present more structure and less randomness require a larger number of parameters to describe their behavior and show high values of FC and EMC (Pachepsky et al., 2006; Wolf, 1999).

~~One way of thinking about~~ Information theory-based metrics ~~is to consider them as metrics~~ characterizing the presence of patterns in time series. The comparison of these metrics for two time series informs about the similarity in shapes found in graphs representing the time series.

We computed the ME, MIG, FC and EMC with the SYMDYN software (Wolf, 1999). The length of words L was set as maximal word length, which guarantees the precision for the information content and complexity metrics at the worst random case. The fluctuation complexity metric usually required the largest number of time series for the same word length (Pachepsky et al., 2006). The word length was set to two in this work as in the work of Pachepsky et al. (2006).

To evaluate model performance by both information content and complexity, distances between measured and observed streamflow time series were calculated in the two-dimensional spaces of information metrics coordinates:

$$d_{i,j} = \left[(MIG_i - MIG_j)^2 + (EMC_i - EMC_j)^2\right]^{1/2} \quad (6)$$

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences Open Access
Discussions

$$d_{"mod"} != !(MIG_{"mod"} - MIG_{"obs"})^! !+ !(FC_{"mod"} - FC_{"obs"})^! / 4 \quad (7)$$

Here subscripts "mod" and "obs" denote information metrics computed from simulated and observed streamflow, respectively. The differences of FC values are normalized by division by two.

Significance of differences between Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) values was estimated based on the approximate NSE distributions developed by McCuen et al. (2006).

3. RESULTS and DISCUSSION

3.1 Watersheds Data Overview.

Figure 1 plots observed daily time series of precipitation and streamflow from Oct. 2 1961 to Oct. 1 1971. The studied watersheds vary with average elevation from 98 m to 594 m, average annual precipitation from 765 mm to 1383 mm, average annual streamflow from 116 mm to 800 mm, and average annual potential evaporation from 711 mm to 1528 mm (Table 2). Since the watersheds ranging from dry to wet represent quite different hydro-climatic conditions, the patterns of streamflow vary significantly among the watersheds. The daily precipitation and streamflow in the three wet watersheds (Tygart Valley River, French Broad River, and Leaf River) are larger than the ones in the two dry watersheds (Guadalupe and San Marcos). Prolonged and frequent periods with streamflow below the detection limit can be found in the dry watersheds as a consequence of prolonged dry periods.

3. 2 Information Content and Complexity Metrics of Precipitation and Streamflow

Information content and complexity metrics for the five watersheds studied are presented in Fig. 2 and in the Table Supp1 in Supplementary material. Since there is no definite recommendation on the word length that has appeared to be an *ad hoc* value in previous publications (e.g., Lange, 1999; Pachepsky et al., 2006; Engelhardt et al., 2009; Pan et al., 2011, 2012) the research of the effect of the word length on the efficiency of information theory based metric needs a separate research and presents an interesting avenue to explore.

The mean information gain and metric entropy of daily precipitation data are larger than 0.78 for all five watersheds (Table S1), indicating the high randomness of the daily precipitation time series and a relatively uniform distribution of the

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

So 4 out of 5 basins? I doubt this means "differentiation" of basins, but rather you should try to understand why the normal conditions basin did not show high MIG values.

system states. Similar metric entropy values were found among the wetter (0.91-0.96, Tygart, French broad and Leaf river) and among the drier watersheds (0.83, Guadalupe and San Marcos) showing the ability of the information theory-based metrics to differentiate and group precipitation time series in terms of the frequency and depth of rainfall.

Streamflow MIG values are about 0.5 less than precipitation MIGs, and the difference is approximately the same for wet and dry watersheds. High values of MIG in precipitation reflect high randomness in time series. The randomness is slightly less for precipitation in the dry watersheds than in the wet ones. The much lower values of streamflow MIG reflect the fact that the physical processes watersheds work as information filters that remove substantial random noise from the precipitation signal while converting it in the streamflow signal. Streamflow time series are not only less noisy, but also more complex. In particular, streamflow EMC values are substantially higher than precipitation EMC values (Fig. 2). This indicates that, as water is routed delivered to streams, not only noise is removed but also additional structure is in introduced in the signal, which improves chances of predictions (higher EMC) and makes fluctuations less random (higher FC). Physical processes of canopy interception, evapotranspiration, infiltration, soil water flow, etc. control the information filtering and these controls impose structure and dampen randomness in the streamflow generation (Pan et al., 2012; Roberts, 2015). Similar behavior has been described for soil water flow with the soil acting as an information filter between rainfall and the resulting soil water content (Pachepsky et al., 2006; Pan et al., 2011; Mishra et al., 2015).

Complexity metrics of precipitation appear to be inversely related to their information content (Fig. 2a, 2b). The larger is the information content and the apparent randomness of precipitation the smaller is the complexity of the time series, and less structure is found in the this time series. Wet watersheds are affected with rainfall with the visibly higher randomness (Fig. 1), and this is reflected in the higher MIG values. Values of the precipitation MIG are somewhat lower in the dry watersheds than in the wet ones. Apparently, dry watersheds receive precipitation that exhibits higher complexity that the wet ones. This indicates the presence of structure and better-expressed patterns in precipitation received in the dry watersheds.

Measured streamflow time series also demonstrate dependencies between information content and complexity measures (Fig. 2c, 2d). The character of these dependencies is different for two complexity measures that reflect different aspects of streamflow patterns. The EMC values reflect the presence of patterns in time series allowing predictability. Streamflow EMC values for the wet watersheds are also lower than for the dry ones. It is not clear if this happens because

* I do not believe that this analysis and conclusions are robust enough. It is known that the processes (usually with longer memory as precipitation) will result into flow dynamics that differ from P dynamics. If you want a deeper understanding you should repeat this using data of soil moisture and other state variables depending on the model structure.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
Discussions

precipitation EMC is lower in wet watersheds, or because the watershed has fewer mechanisms to impose the structure on precipitation signal. The latter suggestion may be supported by results on the dependence of FC on streamflow.

3. 3 Model Performance Evaluation Using Nash- Sutcliffe efficiency and Information Theory-based Metrics

Values of the Nash-Sutcliffe efficiency for eight modes applied at five watersheds are presented in Table 3. Models S1 and M1 perform in unsatisfactory manner. Their values of NSE are close to zero in dry watersheds, and negative in wet watersheds. The latter means that model predictions are worse than prediction using simply average. These results indicate that one cannot assume that the role of subsurface flows is insignificant and knowing runoff is sufficient to predict streamflow dynamics.

According to the classification of Moriasi et al. (2007), performance of models is very good, good, satisfactory, and unsatisfactory if the NSE statistic is larger than 0.75, between 0.65 and 0.75, between 0.5 and 0.65 and less than 0.5, respectively. Based in this classification, performance of all models appears to be unsatisfactory for the Guadalupe watershed. Only S4 and M4 perform satisfactorily in San Marcos watershed, Only S3, S4, M3 and M4 perform satisfactorily in the Tygard Valley watershed. The French Broad and Leaf watersheds have good or very good performance of S3, S3, M3 and M4. Overall, performance of models is better in wet watersheds. The significant improvement occurred for watersheds French Broad, Guadalupe and San Marcos after recharge was added as a mechanism affecting streamflow, i.e. when one changes models S3 and M3 to S4 and M4 respectively (Table 3).

NSE values increase as the conceptual complexity of models increases (see Table 2). It can be seen that the NSE values of S2 models are very close to NSE values of M2 models, NSE values of S3 models are close to NSE values of M3 models, and NSE values of S4 models are very close to the NSE values of M4 models for all watersheds except the San Marcos watershed where M2, M3, and M4 Models have larger NSE than S2, S3, and S4 models respectively.

Inspection of significance of differences between NSE of different models (Table 3) shows that no significant differences are found between average values of NSE of S4 and M4 and among S3, S2, M3, and M2 for the French Broad, among S3, S4, M3 and M4 for the Tygard Valley and Leaf River, between S4 and M4 and between S3 and M3 for the Guadalupe. The absence of significant differences indicates the opportunity of using other indicators of model performance for model selections.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

* This is already known see discussions in e.g. Pechlivanidis et all 2014, 2015, Weijs et al. 2010, 2011

You should also present modelled streamflow time series using all those metrics. Only then you can visualize the characteristics of the flow signal that these info-measures can capture.

Performance of models in terms of information content and complexity of simulated streamflow is compared with the information content and complexity of measured streamflow in Fig. 3 and 4. The corresponding distances between measured and simulated streamfows in coordinates of information-based metrics are shown in the Table Supp2 in the Supplemental materials. Inspection of graphs in Fig. 3 and 4 shows that, although there is some similarity between ranking models by NSE and by information-based metrics, the latter can provide additional insight in the model performance. In particular, the information content and complexity of the French Broad watershed are best simulated by models S2, M2 and M3 (Fig. 3 and 4) although NSE of those models is lower than the one of M4 and S4. The M4 and S4 models seem to generate simulated streamflows that are more complex than the measured ones. Ranking of models by the two complexity metrics – EMC and FC – can be quite different since these metrics reflect different aspects of the complexity in time series. The French Broad watershed provides a good example of that with regard to the model M1. It is almost perfect based on the fluctuation complexity but a very poor result based on effective complexity measure (Fig. 3 and 4).

In the Tygard Valley watershed there is no disagreement between NSE-based and information theory based top-ranked model, both methods point to the model M4. We note that whereas the NSE-based ranking does not discriminate between S4, and M4, the information theory based metrics clearly indicate that the multi-layer soil modeling (M4) better reflect the information content and complexity of this watershed's streamflow than the "single layer soil model" S4 does. A similar situation is observed for the Leaf River watershed where the values NSE for S4 and M4 are indistinguishable, and yet M4 provides much more similarity in information content and complexity between simulated and measured streamflows than S4 does. Models S3 and S4 generate streamflows with substantially smaller information content than M3 and M4. This may indicate that what looks as a noise is actually the result of soil layering.

there is definitely in the streamflow series if you present this.

better

modelled

The Guadalupe watershed gives an example of model not actually working well. Models S4 and M4 give the performance borderline with satisfactory. The information based metrics indicate that M4 is much more preferable, since the single layer models S2, S3, and S4 do not create enough variation to get the information content right. More complexity is needed and this is provided by multi-layer soil models M2, M3, and M4. The example of the Guadalupe River shows also that using two complexity metrics – EMC and FC – can be more efficient than using only one. Model M2, for example, provides values of FC that are very similar to measured ones, i.e. it generates a hidden structure in streamflow time series

inadequate performance

the

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

that is close to that in measured ones. However, this model fails to generate a correct metric EMC, which reflects the predictability of changes in the time series. The same is also true for the San Marcos watershed. The situation here is somewhat similar to the case of the French Broad watershed; the NSE values point to the preferability of S4 and M4 models, but the information content and complexity metrics show that S4 and M4 indeed perform reasonably well, but the best performance is shown by the M3 model which has the third rank in its NSE at this watershed. This indicates that although NSE values are helpful in model discrimination, they are far from capable of integrate qualitative aspects of correspondence between measured and simulated time series (Schaefli and Gupta, 2007).

The simple notion of squared error (Eq. 5) is the first attempt to define the distance between time series in the coordinates of complexity and information content metrics. Weights may be needed to account for the different roles that information content metrics and complexity metrics may play in the evaluation of models. It is possible that these weights can be found from the comparative evaluation of predictive capability of the models. We note that other recently suggested information theory-based methods, such as the so-called Hodrick-Prescott filter (Arias-Hidalgo, 2012), Jensen–Shannon divergence and phase space reconstruction called complexity–entropy causality plane (Serinaldi et al., 2013), can be used to find series patterns and identify recurrent changes in hydrographs. Also, methods of this work may be applied with different word lengths dependent on the length of available time series (Wolf, 1999). Further search for information theory-based metrics to complement accuracy-based metrics presents an interesting research avenue to explore.

5. CONCLUSIONS

The information theory-based metrics were applied in this study to characterize the patterns of observed precipitation and streamflow time series in arid and humid watersheds and to evaluate the performance of eight hydrologic model structures in five watersheds using both traditional Nash-Sutcliffe efficiency (NSE) statistic and usability of information theory-based metrics as complementary to NSE means for comparison and selection models.

We found that:

- patterns of precipitation and streamflow in humid watersheds were more random and less complex than the ones in arid watersheds;

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

- watersheds served as information filters and the streamflow time series were much less random and much more complex than the precipitation time series,
- information content and complexity were substantially different in watersheds with wet and dry climate;
- in pairs of models that differed only by the use of the single-layer or multilayered soil model, the multi-layer model simulated information content and complexity better than the single-layer model in majority of cases;
- values of NSE appeared to be not significantly different for two or more models for each watersheds; in all these cases the information-theory based metrics provided a clear distinction between models and the best models could be selected.

ACKNOWLEDGEMENTS

The Interagency Agreement IAA-NRC-05-005 of USDA-ARS with the US Nuclear Regulatory commission supported YP and FP; GM was supported by the Spanish Ministry of Economy and Competitiveness through the grant FPDI-2013-16742.

REFERENCES

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., Rasmussen, J. 1986a. An introduction to European Hydrologic System-Systeme Hydrologique Europeen, SHE, 1: History and philosophy of a physically-based, distributed modeling system. *J. Hydrol.*, 87, 45-59.

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., Rasmussen, J. 1986b. An introduction to European Hydrologic System-Syteme Hydrologique Europeen, SHE, 2: Structure of a physically-based, distributed modeling system. *J. Hydrol.*, 87, 61-77.

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaksi, F. (Eds.), 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 267-281.

Arias Hidalgo, M. E. 2012. A Decision Framework for Integrated Wetland-River Basin Management In A Tropical And Data Scarce Environment. UNESCO-IHE, Institute for Water Education.

Arnold, J. G., Srinivasan, R., Muttiah, R. S. and Williams, J. R. 1998. Large area hydrologic modeling and assessment part I: Model development, JAWRA J. Am. Water Resour. Assoc., 34(1), 73–89.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Atkinson S., Woods, R.A., Sivapalan, M., 2002. Climate and landscape controls on water balance model complexity over changing time scales. *Water Reours. Res.* 38(12), 1314, doi:10.1029/2002WR001487.

Bai, Y., Wagener, T., Reed, P., 2009. A top-down framework for watershed model evaluation and selection uncertainty. *Environ. Modell. Softw.* 24, 901-916.

Bates, J.E., Shepard, H.K., 1993. Measuring complexity using information fluctuation. *Phys. Lett. A* 172(6), 416-425.

Beven, K.J., Kirkby, M.J., 1979. A physically-based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.*, 24(1), 43-69.

Beven, K.J., Smith, P., 2015.Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *J. Hydrol. Eng.* 20(1), A4014010.

Chen, L., Shen, Z., Yang, X., Liao, Q., Yu, S.L., 2014. An interval-deviation approach for hydrology and water quality model evaluation within an uncertainty framework. *J. Hydrol.* 509, 207-214.

Crawford, N.H., Linsley, R.K., 1966. Digital simulation in hydrology: Stanford Watershed MODEL IV. Technical Report No. 39, Stanford University, Palo Alto, California.

Engelhardt, S., Matyssek, R. and Huwe, B.: Complexity and information propagation in hydrological time series of mountain forest catchments, Eur. J. For. Res., 128(6), 621–631, doi:10.1007/s10342-009-0306-2, 2009.

Farmer, D., Sivapalan, M., Jothiyangkoon, C., 2003. Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis. *Water Resour. Res.* 39(2), 1035, doi:10.1029/2001WR000328.

Gong, W., Gupta, H.V., Yang, D., Sricharan, K., Hero III, A.O., 2013. Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resour. Res.* 49, 2253-2273, doi: 10.1002/wrcr.20161.

Grassberger, P., 1986. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* 25, 907-938.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.*, 34(4), 751-763.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

Kashyap, R.L., 1982. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE T. Pattern Anal.* 4(2), 99-104.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* 5, 89-97.

Lange, H., 1999. Time series analysis of Ecosystem variables Uwe Ehret with complexity measures. *InterJournal for Complex Systems* Mauscript #250. New England Complex Systems Institute, Cambridge, MA.

Li, C. Singh, V.P., Mishra, A.K., 2012. Entropy theory-based criterion for hydrometric network evaluation and design: Maximum information minimum redundancy. *Water Resour. Res.* 48, W05521, doi: 10.1029/2011WR011251.

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.* 99(D7), 14415-14428.

McCuen, R. H., Knight, Z., & Cutter, A. G. 2006. Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering*. 11(6): 597-602.

Mishra, V., Ellenburg, W., Al-Hamdan, O., Bruce, J., Cruise, J., 2015. Modeling Soil Moisture Profiles in Irrigated Fields by the Principle of Maximum Entropy. Entropy 17, 4454–4484. doi:10.3390/e17064454

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. Asabe, 50*(3), 885-900

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, Part I – A discussion of principles. *J. Hydrol.* 10, 282-290.

Pachepsky, Y., Guber, A., Jacques, D., Simunek, J., Van Genuchten, M.T., Nicholson, T., Cady, R., 2006. Information content and complexity of simulated soil water fluxes. Geoderma 134, 253–266. doi:10.1016/j.geoderma.2006.03.003

Pan, F., Pachepsky, Y. a., Guber, A.K., Hill, R.L., 2011. Information and complexity measures applied to observed and simulated soil moisture time series. Hydrol. Sci. J. 56, 1027–1039. doi:10.1080/02626667.2011.595374

Pan, F., Pachepsky, Y. a., Guber, A.K., McPherson, B.J., Hill, R.L., 2012. Scale effects on information theory-based measures applied to streamflow patterns in two rural watersheds. J. Hydrol. 414-415, 99–107. doi:10.1016/j.jhydrol.2011.10.018

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences Open Access
Discussions

Pechlivanidis, I.G., Jackson, B., McMillan, H., Gupta, H., 2014. Use of an entropy-based metric in multiobjective calibration to improve model performance. *Water Resour. Res.* 50, 8066-8083, doi: 10.1002/2013WR014537.

Pechlivanidis, I.G., Jackson, B.M., Mcintyre, N.R., Wheater, H.S., 2011. Catchment scale hydrological modeling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. *Global Nest J.* 13(3), 193-214.

Rathinasamy, M., Khosa, R., Adamowski, J., Ch, S., Partheepan, G., Anand, J., Narsimlu, B., 2014. Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models. *Water Resour. Res.* 50, 9721-9737, doi: 10.1002/2013WR014650.

Reusser, D.E., Blume, T., Schaefli, B., Zehe, E., 2009. Analysiing the temporal dynamics of model performance for hydrological models. *Hydro. Earth Syst. Sc.* 13, 999-1018.

Roberts, A.D., 2015. The effects of current landscape configuration on streamflow within selected small watersheds of the Atlanta metropolitan region. J. Hydrol. Reg. Stud. doi:10.1016/j.ejrh.2015.11.002

Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Stat.* 6(2), 461-464.

Serinaldi, F., Zunino, L., Rosso, O. a., 2013. Complexity–entropy analysis of daily stream flow time series in the continental United States. Stoch. Environ. Res. Risk Assess. 28, 1685–1708. doi:10.1007/s00477-013-0825-8

Shannon, C.E., 1948. A mathematical theory of communication. *AT&T Tech. J.* 27, 379-423, 623-656.

Singh, V.P., Woolhiser, D.A., 2002. Mathematical modeling of watershed hydrology. J. *Hydrol. Eng.* 7(4), 270-292.

Son, K., Sivapalan, M., 2007. Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resour. Res.* 43, W01415, doi: 10.1029/2006WR005032.

Sugawara, M., Ozaki, E., Wantanabe, I., & Katsuyama, Y. (1976). Tank Model and its Application to Bird Creek, Wollombi Brook, Bihin River, Sanaga River, and Nam Mune. National Center for Disaster Prevention, Tokyo, Research Note, 11, Kyoto, Japan, pp. 1-64.

Wagener, T., Sivapalan, M., Troch, P.A., McGlynn, B.L., Harman, C.J., Gupta, H.V., Kumar, P., Rao, P.S.C., Basu, N.B., Wilson, J.S., 2010. The future of hydrology: An evolving science for a changing world. *Water Resour. Res.* 46, W05301, doi: 10.1029/2009WR008906.

Weijs, S.V., Schoups, G., van de Giesen, N., 2010. Why hydrological predictions should be evaluated using information theory. *Hydrol. Earth Syst. Sc.* 14, 2545-2558.

Wolf, F., 1999. Berechnung von Information und Komplexität von Zeitreihen – Analyse des Wasserhaushaltes von bewaldeten Einzugsgebieten. Bayreuth. Forum Okol. 65, 164 S.

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resour. Res.* 40, W05113, doi:10.1029/2003WR002557.

Yu, Z., Lakhtakia, M.N., Yarnal, B., White, R.A., Miller, D.A., Frakes, B., Barron, E.J., Duffy, C., Schwartz, F.W., 1999. Simulating the river-basin response to atmospheric forcing by linking a mesoscale meteorological model and hydrologic model system. *J. Hydrol.* 218, 72-91.

Zhao, R., Zhuang, Y., Fang, L., Liu, X., Zhang, Q., 1980. The Xinanjiang model. Proceedings of Oxford Symposium on Hydrological Forcasting, IAHS Publication No. 129, International Association of Hydrological Sciences, Wallingford, U.K., 351-356.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

Table 1. Selected properties of watersheds in this study.

Basin name and sampling location	**Area [km^2]**	**Mean elevation [m]**	**Mean annual P [mm]**	**Mean annual Q [mm]**	**Mean annual PE [mm]**
French Broad River near Asheville, NC	2448	594	1383	800	819
Tygart Valley River near Pipestem, WV	2372	390	1166	736	711
Leaf River near Collins, MS	1950	111	1346	415	1052
Guadalupe River near Spring Branch, TX	3406	289	765	116	1528
San Marcos River near Luling, TX	2170	98	827	179	1449

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

I am still confused on the model structures. You say that they differ on soil moisture accounting, evaporation and routing. So why not define the ~~table~~ structure based on this? You could have seperate columns for SMA, ET and ROUT.

Table 2. General description of the models used (after Bai et al., 2009).

ID	General description
S1	Single-layer model with single store. Runoff generation controlled by maximum soil water storage
S2	Single-layer model with single store. Runoff generation by saturation excess and subsurface flow controlled by threshold storage
S3	Single-layer model with two stores (unsaturated and saturated zones). Evaporation and transpiration from both stores. Runoff generation by saturation excess and subsurface flow from the saturated zone
S4	Single-layer model with three stores (unsaturated and saturated zones and deep store). Evaporation and transpiration from saturated and saturated zones. Base flow losses from deep store. Runoff generation by saturation excess and subsurface flow from the saturated zone
M1	Multi-layer (10 layers to represent a soil moisture profile that fits the Xinanjiang model distribution) model with single store. Runoff generation controlled by maximum soil water storage
M2	Multi-layer model with single store. Runoff generation by saturation excess and subsurface flow controlled by threshold storage
M3	Multi-layer model with two stores (unsaturated and saturated zones). Evaporation and transpiration from both stores. Runoff generation by saturation excess and subsurface flow from the saturated zone
M4	Multi-layer model with three stores (unsaturated and saturated zones and deep store). Evaporation and transpiration from saturated and saturated zones. Recharge of the deep store. Runoff generation by saturation excess and subsurface flow from the saturated zone

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

Table 3. The Nash-Sutcliffe efficiency values for eight models in five watersheds.

What are the subscripts for?

Model	French Broad	Tygard Village	Leaf River	Guadalupe	San Marcos
S1	-1.499	-0.231	-0.227	0.205	0.076
S2	0.590[b]	0.477[b]	0.643[b]	0.407[c]	0.378[e]
S3	0.608[b]	0.541[a]	0.682[a]	0.450[b]	0.389[e]
S4	0.764[a]	0.567[a]	0.700[a]	0.508[a]	0.548[b]
M1	-1.236	-0.198	-0.130	0.211	0.114
M2	0.589[b]	0.476[b]	0.640[b]	0.418[c]	0.448[d]
M3	0.609[b]	0.545[a]	0.704[a]	0.460[ab]	0.497[c]
M4	0.754[a]	0.559[a]	0.699[a]	0.478[a]	0.584[a]

The same superscript indicates that NSE values are not significantly different at the 0.05 significance level.

The performance for S1 and M1 is very poor for all basins; Any analysis using these two structures can not be

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

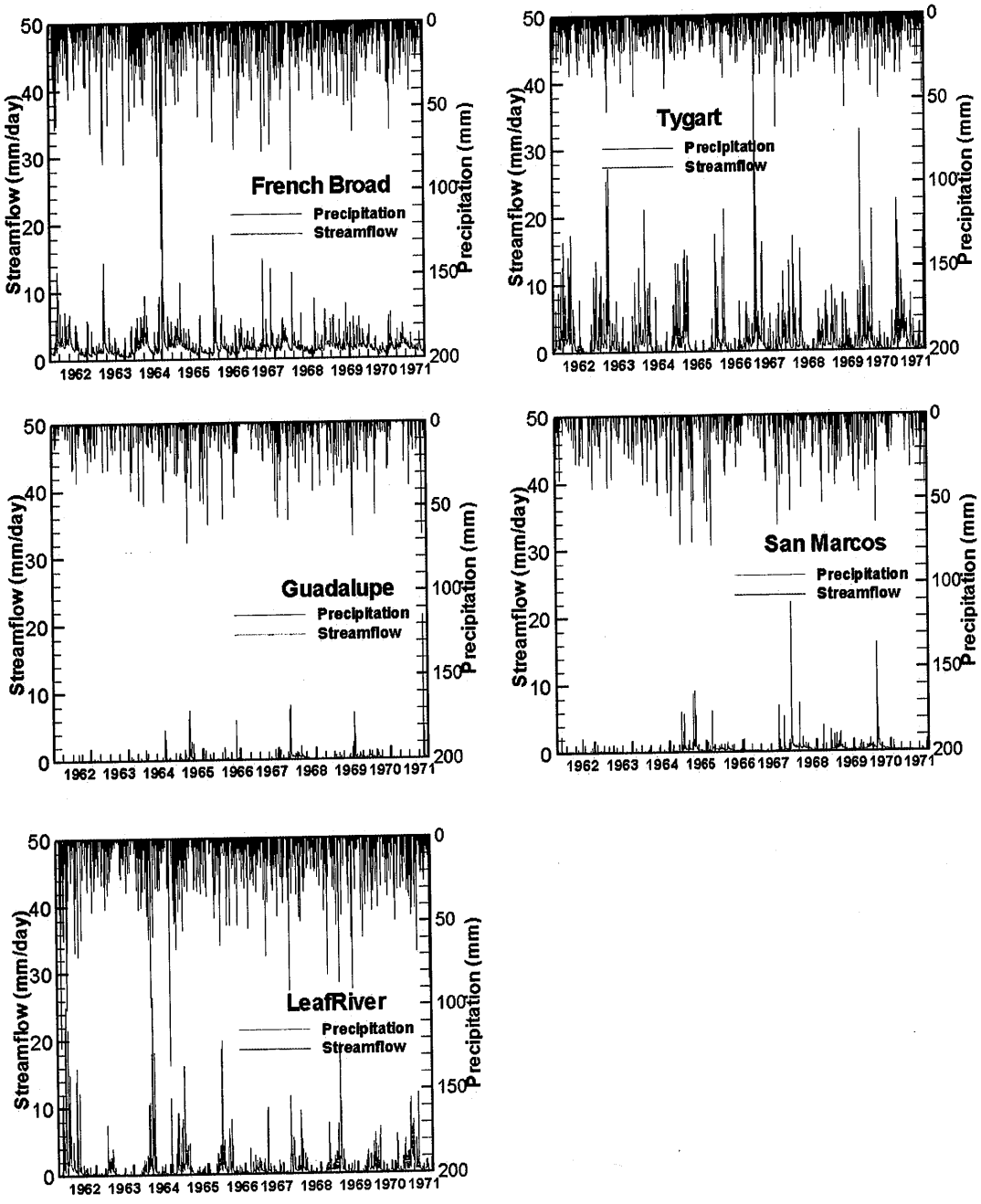
Hydrology and Earth System Sciences
Open Access
Discussions

List of figures.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
Discussions

Figure 1. Daily observed precipitation and streamflow time series from Oct. 2 1961 to Oct. 1 1971 at five different watersheds across US.



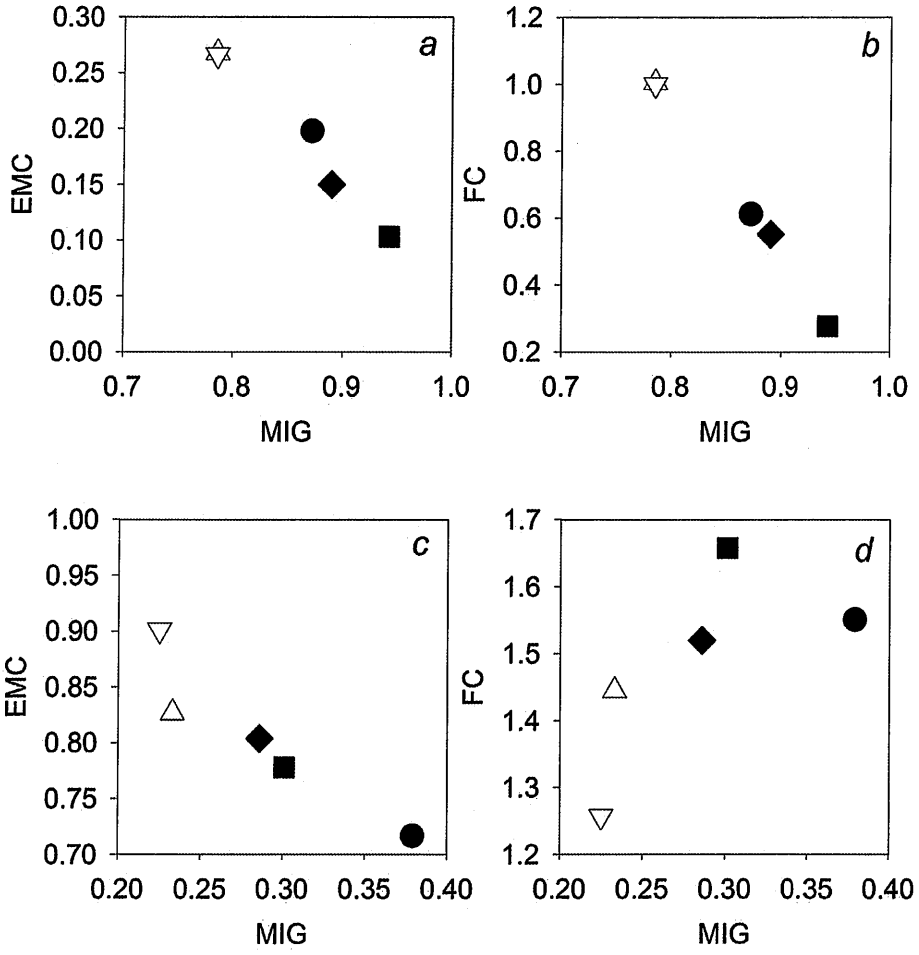
5

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

The caption does not allow understanding of the figure. What do the symbols refer to and what are the sub plots for?

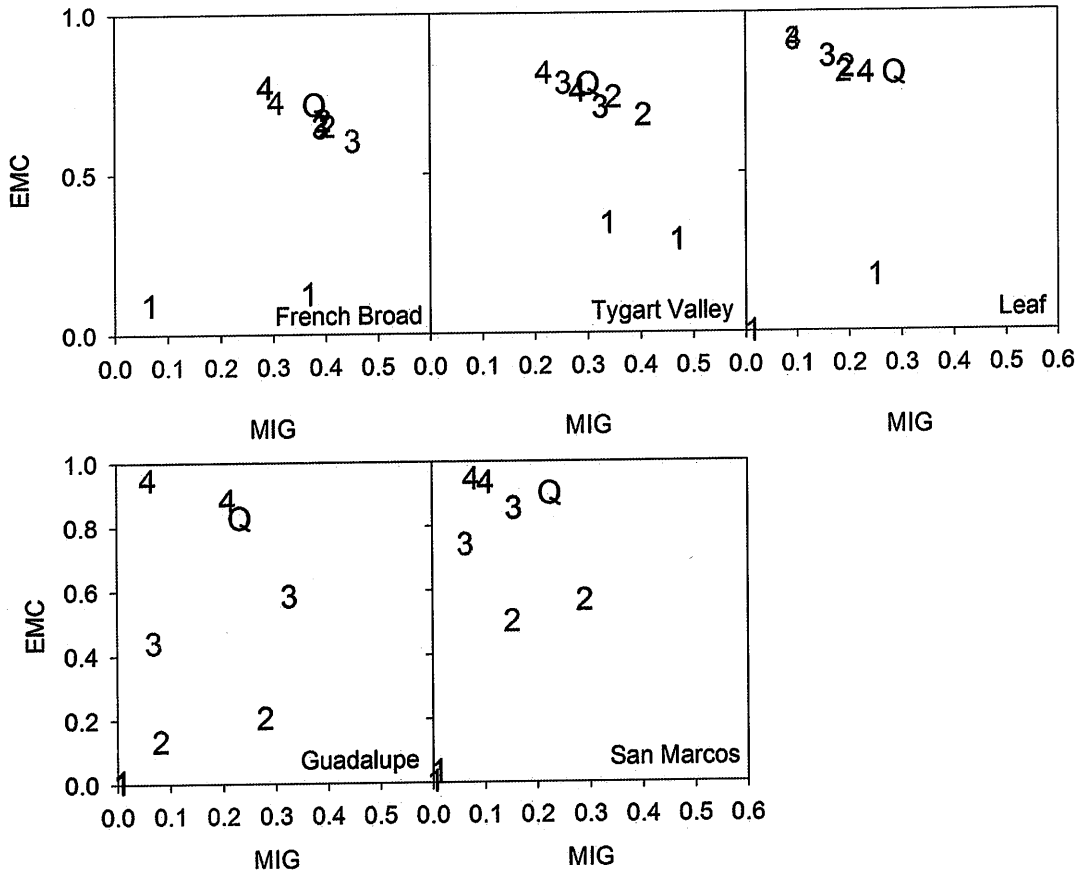
Figure 2. Relationships between the mean information content (MIG) and complexity metrics – effective complexity measure (EMC) and fluctuation complexity (FC) in precipitation time series of watersheds in this study: ! - French Broad river, " - Tygard Valley river, # - Leaf river, ! - Guadalupe river, " - San Marcos river.



Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Open Access
Discussions

Figure 3. Relationships between mean information content (MIG) and effective measure of complexity (EMC) in measured (Q) and simulated (numbers) streamflow time series. Blue symbols 1, 2, 3, 4 correspond to single-layer soil models S1, S2, s3, and S4, red symbols 1, 2, 3, 4 correspond to multi-layer soil models M1, M2, M3, M4.



Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-46, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 15 February 2016

Hydrology and Earth System Sciences
Discussions

Figure 4. Relationships between mean information content (MIG) and fluctuation complexity (EMC) in measured (Q) and simulated (numbers) streamflow time series. Blue symbols 1,2,3,4 correspond to single-layer soil models S1, S2, s3, and S4, red symbols 1,2,3,4 correspond to multi-layer soil models M1, M2, M3, and M4.