INCORPORATING THE SPATIO-TEMPORAL DISTRIBUTION OF RAINFALL AND BASIN GEOMORPHOLOGY INTO NONLINEAR ANALYSES OF STREAMFLOW DYNAMICS

Boyko Dodov and Efi Foufoula-Georgiou

St. Anthony Falls Laboratory
Department of Civil Engineering
University of Minnesota
Mississippi River at 3rd Ave. SE
Minneapolis, MN 55414

(Revised December, 2004)

(Accepted for publication December, 2004)

Corresponding author: Efi Foufoula-Georgiou (E-mail: efi@tc.umn.edu)

Abstract

Many recent studies have been devoted to the investigation of the nonlinear dynamics of rainfall or streamflow series based on methods of dynamical systems theory. Although finding evidence for the existence of a low-dimensional deterministic component in rainfall or streamflow is of much interest, not much attention has been given to the nonlinear dependencies of the two and especially on how the spatio-temporal distribution of rainfall affects the nonlinear dynamics of streamflow at flood time scales. In this paper, a methodology is presented which simultaneously considers streamflow series, spatio-temporal structure of precipitation and catchment geomorphology into a nonlinear analysis of streamflow dynamics. The proposed framework is based on "hydrologically-relevant" rainfall-runoff phase-space reconstruction acknowledging the fact that rainfall-runoff is a stochastic spatially extended system rather than a deterministic multivariate one. The methodology is applied to two basins in Central North America using 6-hour streamflow data and radar images for a period of five years. The proposed methodology is used to: (a) quantify the nonlinear dependencies between streamflow dynamics and the spatio-temporal dynamics of precipitation; (b) study how streamflow predictability is affected by the trade-offs between the level of detail necessary to explain the spatial variability of rainfall and the reduction of complexity due to the smoothing effect of the basin; and (c) explore the possibility of incorporating process-specific information (in terms of catchment geomorphology and an a-priori chosen uncertainty model) into nonlinear prediction. Preliminary results are encouraging and indicate the potential of using the proposed methodology to understand via nonlinear analysis of observations (i.e., not based on a particular rainfall-runoff model) streamflow predictability and limits to prediction as a function of the complexity of spatio-temporal forcing relative to basin geomorphology.

1. INTRODUCTION

Increasing implementation of remote sensing technologies for observation of atmospheric processes, and in particular, the availability of radar and satellite estimates of rainfall, provides a good base for improving the accuracy of flood prediction. Although the remotely sensed information of hydrological processes has become more detailed and reliable and the theoretical basis for developing sophisticated data mining and statistical learning algorithms has been pursued vigorously during the last decade, there is only a limited number of hydrologic studies exploring the use of spatial information of precipitation for understanding the dynamics of basin response based on a nonlinear dynamical systems perspective, as opposed to conceptual/empirical modeling.

The idea of nonparametric streamflow prediction based on past rainfall and streamflow time series was first introduced by the pioneering work of *Yakowitz* [1987] and *Yakowitz and Karlsson* [1987], who implemented the nearest neighbor approach for streamflow prediction at daily time scale. After that, many studies have been devoted to the implementation of the time series phase-space method for prediction and/or finding evidence for the existence of a deterministic component in hydrologic time series. Studies investigating hydrologic predictability based on past data and nonlinear dynamical methods can be divided into two groups: (1) studies limited to the rainfall process [e.g. *Hense,* 1987; *Rodriguez-Iturbe et al.,* 1989; *Sharifi et al.,* 1990; *Tsonis et al.,* 1993; *Jayawardena and Lai,* 1994; *Georgakakos et al.,* 1995; *Koutsoyiannis and Pachakis,* 1996; *Sivakumar et al.,* 1998], and (2) studies limited to streamflow [*Jayawardena and Lai,* 1994; *Porporato and Ridolfi,* 1996, 1997; *Liu et al.,*1998; *Wang and Gan,* 1998, *Babovic and Keijzer,* 1999, *Sivakumar et al.,* 2002, *Phoon et al.,* 2002, *Islam and Sivakumar,* 2002, among others]. Only recently [e.g. *Sivakumar et al.,* 2001, *Porporato and Ridolfi,* 2001] the phase-space approach has been applied to the coupled rainfall-runoff process. This was accomplished by considering either the phase-space of the runoff coefficient (runoff divided by precipitation) or the multivariate rainfall-runoff phase-space where rainfall was measured at several locations.

In most of the papers, the analysis is based on the method of optimal phase-space reconstruction as it was originally developed for univariate time-series [see *Takens,*1981, *Farmer and Sidorowich,* 1987, *Casdagli et al.,* 1991, and *Sauer et al.,* 1993] and extended lately to the multivariate case [e.g. *Cao et al.,* 1998 among others]. However, the applicability of this approach in the context of rainfall-runoff nonlinear prediction (i.e. considering the rainfall-runoff as a nonlinear deterministic uni/multivariate process) has been questioned in the literature recently [see *Pasternack, G. B.,* 1999 and *Schertzer et al.,* 2002]. The reason for the criticism is the fact, that, considered simultaneously, the rainfall and runoff processes truly represent a spatially extended dynamical system and any attempt to consider such a system as a uni/multivariate one by means of spatial subsampling (i.e. rainfall and/or runoff at fixed

locations) has to be met with skepticism. This is because, an extended system has infinitely many degrees of freedom and is thus making the question of "optimal" phase-space reconstruction and, in general, the search for low dimensional attractor meaningless. Rigorous theoretical treatments of spatially extended dynamical systems are in their infancy. They typically relate to representing such systems as *coupled maps on a lattice* [e.g. see *Bohr et al.*, 1998 and *Parlitz and Merkwirth*, 2000], but this approach in hydrology would be only applicable for the spatially distributed atmospheric forcing, i.e. rainfall. When the interaction between the atmospheric forcing and the land surface is to be considered in relevance to the produced streamflow, the analysis becomes much more complicated [e.g. *Dodov and Foufoula-Georgiou*, 2001]. For example, factors such as catchment geomorphology make the "importance" of inputs at different space-time locations different. Moreover, *unobserved* processes such as land-atmosphere interactions, hydraulic dispersion, etc. can be considered to act as stochastic filters on the observations and make them uncertain, with uncertainty dependent on the space-time location of the inputs.

In the above context, any methodology for multivariate rainfall-runoff nonlinear prediction at flood time scale based on phase-space reconstruction [e.g. see recent studies *Sivakumar et al.*, 2001, and *Porporato and Ridolfi*, 2001] can be considered as an approximate technique without any formal warrantee of optimality in the reconstruction of phase-space and in the prediction. This does not mean that such techniques are not valuable tools in understanding the complex rainfall-runoff dynamics and any modifications resulting in improvements should be considered valuable.

In this paper, we propose a methodology for rainfall-runoff nonlinear analysis and prediction based on "hydrologically-relevant" phase-space reconstruction acknowledging the fact that rainfall-runoff is a stochastic and extended rather than deterministic and multivariate system. The main innovations of the proposed approach are: (a) in contrast to subsampling at a few locations or a priori aggregating the spatially variable rainfall without considering its interaction with land surface, we reduce the system's dimensionality using aggregation determined by the catchment geomorphology; (b) we explicitly incorporate the catchment's hydrologic response as an additional space-time varying dimensionality reduction factor (operating on already aggregated rainfall data) which, in fact, is used to define the "hydrologically optimal" state of the system, and respectively phase-space; (c) in order to treat the uncertainty due to the stochastic processes in the catchment (e.g. hydraulic dispersion) we project this uncertainty on the rainfall input, through an uncertainty model, thus accepting a scheme for nonlinear prediction in which the processes in the catchment are considered as deterministic and all the uncertainty is captured by the forcing.

The developed methodology is used to: (1) analyze the nonlinear dependencies between the dynamics of catchment-scale runoff and the spatio-temporal dynamics of precipitation considered at different spatial scales; (2) study the way by which runoff predictability is affected by the trade-offs between the level of detail necessary to explain the dynamics of forcing and the

reduction of complexity due to the smoothing effect of the basin; (3) analyze different ways of incorporating uncertainty in the catchment response and prior knowledge about basin geomorphology in the phase-space approach; (4) analyze the effect of geologic controls on streamflow predictability.

In Section 2 we introduce the phase-space approach and how it is used as a tool for nonlinear prediction. Section 3 discusses the specifics related to the reconstruction of the rainfall-runoff phase-space and introduces distance measures (specific for comparison of rainfall images as well as catchment-specific distance measures based on catchment geomorphology and an a-priori chosen uncertainty model) capable of evaluating the "hydrologic similarity" between rainfall-runoff states. In section 4 we analyze real data by applying the above approaches to two basins in Central North America using 6-hr rainfall accumulation radar rainfall images and 6-hr streamflow data for a period of approximately 6 years. Summary and conclusions are given in section 5.

2. THE PHASE-SPACE APPROACH

2.1. General definition

The state space (or phase-space) is a finite dimensional vector space $\boldsymbol{R}^d$, such that specifying a point in this space uniquely specifies the state of a dynamical system or stochastic process and vice versa. A large class of processes, in which a prominent member is the stochastic Markov processes, can be described by a set of states (possibly infinite) and some kind of transition rules specifying how the system proceeds from one state to another. An essential feature of Markov processes is that the future evolution at any given time t depends on the state of the system at that time only and not on states of the past. For stochastic processes, the evolution in time is given in the form of a set of transition probabilities, according to which the future states of the system are chosen probabilistically. In this general framework, a purely deterministic system can be considered as a limiting case of a Markov process on the continuum of states, where evolution to a new state occurs with probability 1 along some particular direction and any other transition has probability 0. Such a general framework is useful when it is considered in the context of predictions based on phase-space reconstructed from time series: the evolution of a deterministic system (for short enough time intervals) can be approximated by a simple linear, or low order polynomial relationship. In contrast, the transition of a stochastic system to a new state can be considered only in terms of its expectation, i.e. as the average over all possible future states, given the current one. However, from a clearly approximation point of view the two cases are very similar and consist of (i) proper phase-space reconstruction, (ii) choice of a similarity metric between the system's states, and (iii) definition of a local model for prediction.

2.2. Optimal phase-space reconstruction from time series

While for stochastic systems there is no specific rule for phase-space reconstruction except some physical and/or statistical considerations, the optimal phase-space reconstruction of a deterministic uni/multivariate nonlinear system is obtained by "embedding" the dynamics of the process utilizing the so-called delay time method [see *Takens*, 1981; *Casdagli et al.*, 1991, and *Sauer et al.*, 1993 for the univariate case]. This approach was originally applied for univariate time-series, representing the "state" of the system by the vector

$$\boldsymbol{X_t} = \{X_t\ , X_{t\text{-}t}\ , ..., X_{t\text{-}(m-1)t}\ \}, \tag{1}$$

where t is called the delay time, and m the embedding dimension representing the dimensionality of the phase-space. The important issues for an optimal choice of t and m, and thus, optimal phase-space reconstruction from univariate time series are: (1) two successive elements of a state-vector $X_{t-jt}, X_{t-(j+1)t}, j < m-2$, have to be "optimally" (but not completely) uncorrelated in order that every new element of the vector brings enough "new" information about the state of the system; and (2) the embedding dimension m has to be small enough to ensure maximally low dimensionality of the phase-space and at the same time large enough to uniquely "unfold" the state of the system. In other words, an optimal embedding dimension is the smallest number m, for which a neighborhood of k states in an m dimensional phase-space remains a neighborhood (i.e. the states are still close to each other) in a $(m+1)$ dimensional phase-space.

In the case of multivariate time series, say of dimensionality M, the state vector is given as

$$\begin{aligned}\boldsymbol{X_t} = \{&X_t^{(1)}, X_{t\text{-}t_1}^{(1)}, \ldots, X_{t\text{-}(m_1-1)t_1}^{(1)},\\ &X_t^{(2)}, X_{t\text{-}t_2}^{(2)}, \ldots, X_{t\text{-}(m_2-1)t_2}^{(2)},\\ &\ldots \quad \ldots \quad \ldots \quad \ldots\\ &X_t^{(M)}, X_{t\text{-}t_M}^{(M)}, \ldots, X_{t\text{-}(m_M-1)t_M}^{(M)}\}\end{aligned} \tag{2}$$

where M is the number of observables, and the dimensionality of the phase-space is $d = \sum_{j=1}^{M} m_j$. Examples of such multivariate time series could be streamflow at M gauges, or a vector composed of precipitation intensities at M pixels of a radar image, or a combination of those. The problem of choice of optimal parameters m_j and t_j, $j = 1,\ldots, M$ for this case is much more delicate and less explored. Some methodologies are suggested in *Cao et al.*, [1998] and *Phoon et al.*, [2002], but the analysis of all possible combinations still remains valuable [e.g. see *Porporato and Ridolfi*, 2001].

The phase-space reconstructed by means of the delay time method is considered to be a linear vector space where the distance between two states can be measured as the Euclidean

distance between two *d*-dimensional points. Given the reconstructed *d*-dimensional phase-space and a distance metric in this space, the overall nonlinear behavior of a dynamical system (both deterministic or stochastic) can be approximated locally by a simple nonparametric model. Below, a general form of selection of such a local model is presented, applicable for both deterministic and stochastic time series.

2.3. Choice of a local regression model for prediction

The core idea of nonlinear phase-space prediction is: given some "new" reference state $\boldsymbol{X}_0$, and the set of all observed state vectors $\mathcal{X}_N = \{\boldsymbol{X}_t, t = 1,.., N\}$ predict a future value of some of the components of the system (say the *j*-th component). In other words, construct the predictor

$$x^j(t_0 + T) = F\{\boldsymbol{X}_0, \mathcal{X}_N\} \tag{3}$$

where *T* is some prediction horizon and $\boldsymbol{F}$ is a nonlinear function.

Let us assume that we have constructed a neighborhood function (sometimes called the kernel function) $K(\boldsymbol{X}_0, d, w)$ around a reference state $\boldsymbol{X}_0$, based on some distance metric $d(\boldsymbol{X}_i, \boldsymbol{X}_j)$ measuring the similarity between two state vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, and a window parameter *w* determining the size of the neighborhood and thus the "locality" of the prediction. Let $W(w)$ denote the neighborhood defined by *w* around a state $\boldsymbol{X}_0$ and $\mathcal{X}_{W(w)}$ the subset of $\mathcal{X}_N$ defined by the kernel function $K(\boldsymbol{X}_0, d, w)$. Then, an estimate of $x^j(t_0 + T)$, $\hat{x}^j(t_0 + T)$, is computed by the local prediction model $\hat{F}\{\boldsymbol{X}_0, \mathcal{X}_{W(w)}\}$, i.e.,

$$\hat{x}^j(t_0 + T) = \hat{F}\{\boldsymbol{X}_0, \mathcal{X}_{W(w)}\} \tag{4}$$

where the nonlinear function $\hat{F}$ is chosen among a set of polynomial functions $P(Q)$ (usually, first or zero-th order) with parameters Q such that a quadratic error between the predicted and observed states is minimized (in the usual least squares optimization framework).

The neighborhood function K_f can be hard-thresholded, in which case $\mathcal{X}_{W(w)} \subset \mathcal{X}_N$ is the subset of states contained in the (hyper)sphere of radius *w* around $\boldsymbol{X}_0$, or soft-thresholded, for example a Gaussian kernel of the type $K_f = \exp(-\| \boldsymbol{X}_i - \boldsymbol{X}_j \|^2 / w^2)$, where $i, j \in \{1,..., N\}$, in which case the neighborhood is extended to all the state vectors with weights that become smaller the further away we move from the reference state vector $\boldsymbol{X}_0$. It is noted that if the neighborhood function K_f is hard-thresholded, there are two (technically speaking) different approaches to nonlinear prediction: (i) neighborhood of fixed radius *w* and changing mass (number of states) from one reference state to another, and (ii) neighborhood of fixed mass (number of states) and changing radius *w*. In the rest of the paper the second approach is adopted for nonlinear prediction.

If the local prediction model $P(Q)$ is of zero order the predicted value is the local (or weighted in the case of soft threshold) average over the future values $x^j(t_0+T)$ of the states $\mathcal{X}_{W(w)}$ in the neighborhood $W(w)$. If $P(Q)$ is of first order the optimization is reduced to local (or weighted) linear fit on $\mathcal{X}_{W(w)}$. For deterministic dynamical systems with low noise level and densely populated phase-space, i.e. long enough time series, local linear fit is the most appropriate approach. When the noise level is high and/or the determinism is weak or not present, zero-th order prediction usually gives better results [*Kantz and Schreiber*, 1997].

In the next section we consider the choice of optimal phase-space parameters and distance measures as they apply specifically to the rainfall runoff process.

3. PHASE-SPACE TECHNIQUES FOR RAINFALL-RUNOFF ANALYSIS AND PREDICTION: INCORPORATING KNOWLEDGE ABOUT CATCHMENT GEOMORPHOLOGY

As discussed in Section 1, it is not correct to apply methods designed for multivariate time series analyses to data representing a spatially extended system by reducing the infinite dimensional system to a finite dimensional one via subsampling at a few discrete locations. This is especially so when the process under consideration has high fine-scale spatial variability, such as rainfall. However, during the process of runoff generation and runoff routing, a natural smoothing and a dampening of the effect of rainfall's spatial variability occurs, by the storage effects of the most shallow aquifers and by the integration (smoothing) process taking place because of the natural hierarchy of the stream network topology. The most appropriate way then to incorporate the fine-scale spatial variability of rainfall and the smoothing effect of the basin simultaneously is not to increase the number of rainfall point observations to infinitely many or subsample a few of them arbitrarily, but rather to aggregate the spatially distributed rainfall input according to the catchment topology.

Due to the effect of geomorphology, the streamflow at the outlet of the catchment at any moment depends in a different way on the past rainfall occurrences at different space-time locations. For example, rainfall occurring far back in time and far upstream may still contribute to the discharge at the outlet after, say, three hours, while in contrast, the one occurring relatively recently close to the outlet may have no contribution at all to the streamflow at this particular time horizon. Following this line of thinking, one could easily understand the idea that the information of catchment geomorphology could be used for *additional reduction of dimensionality*, by weighting or even discarding rainfall inputs according to their contributions to the discharge at the outlet at particular instants in time.

The aggregation of spatially distributed rainfall according to catchment topology will reduce the dimensionality of the phase–space but will also reduce the uncertainty (due to smoothing) in the observed spatially distributed rainfall. However, there is uncertainty introduced by

unobserved processes such as hydraulic dispersion, acting as stochastic filters on the observations a posteriori. This type of uncertainty depends on the space-time locations of inputs and has to be addressed in the framework of nonlinear prediction.

In order to reflect the above considerations, we propose a methodology for rainfall-runoff nonlinear prediction, consisting of two techniques: (1) a scheme similar to the classical multivariate time-series analysis with "optimal" phase-space reconstruction based on the introduction of a "normalized distance space", with distance metrics that are better able to reflect comparison of spatially distributed fields (such as rainfall images) rather than comparison of multidimensional vectors, and (2) a scheme without phase-space reconstruction in the above sense, but based on the definition of a neighborhood in a "hydrologically optimal" phase-space through catchment-specific distance measures and an uncertainty model. The details of the proposed schemes are considered below.

3.1. Technique with "optimal" phase-space reconstruction (T1)

Considering the high spatial variability of rainfall, even comparison of an image to its slight translation in terms of their Euclidean distance could result in a huge separation in the phase-space. However, considering the aforementioned smoothing effects in the catchment, small perturbations in local rainfall intensities and/or small spatial translations in rain cells are not expected to affect the discharge at the outlet significantly. In order to account for this effect, we introduce two more distance measures for image comparison, namely, Multi-Scale Root Mean Square error – MSRMS, and Hausdorff distance based norm – HD (see Appendix I for details). Since both MSRMS and the HD are not Euclidean measures, we cannot consider the rainfall and streamflow data together in a single phase-space. Rather, we calculate the distances between the reference and all other "rainfall states" and the distances between the corresponding "streamflow states" (measured as Euclidean distances in the *optimally reconstructed* univariate phase-space of streamflow) and project those distances (normalized by their medians) on what we call "distance space". This space can be considered as the nonnegative quadrangle of a two dimensional Cartesian coordinate system in which the x axis represents "streamflow distances", i.e., the distances between the reference and all other streamflow states, and the y axis represents "rainfall distances" i.e., distances between the reference and all other rainfall states. It is in this space that we define the neighborhood of the collective reference state (rainfall and streamflow) as shown in the schematic of Fig.1. A similar approach is applied to the spatially aggregated rainfall time-series, with Euclidean distances between the "rainfall states" measured in the appropriate multivariate phase-space of aggregated rainfall.

To assess the optimal reconstruction of the phase-space we consider the prediction accuracy for all possible combinations of embedding dimensions, number of neighbors and prediction horizons within reasonable intervals of these parameters. The delay time is chosen to be 1 time

step, since we assume that the catchment acts as a continuous integrator of the input and all the information about the rainfall is important for the discharge at the outlet.

3.2. Technique with phase-space reconstruction based on catchment-specific distance measure and an uncertainty model (T2)

This type of analysis is applied only with respect to spatially aggregated (over subcatchments) rainfall time series. In order to incorporate catchment geomorphology into the nonlinear prediction, we employ a geomorphologically derived response function (implementing the spatially distributed velocity field approach of *Maidment et al.*, [1996] based on geomorphologic analysis of the catchment's Digital Elevation Model - DEM). This response function can be used to assign weights to the elements of the state vector according to how much they affect the streamflow at the basin outlet. The response function is nothing but a geomorphologically derived unit hydrograph (UH) and in particular, its discrete normalized version, such that $\sum_{\ell=0}^{\infty} U_\ell = 1$ (see Fig.2a). More specifically, the basin response function is derived as follows:

- First, using a standard DEM analysis, a slope and an upstream contributing area are assigned to every pixel of the DEM.
- Second, the at-pixel travel time is computed assuming a power-law dependence of velocity on the area draining to the pixel and the local slope in the form $L/T = V = aA^b S^{0.5}$, where a and b are parameters, A is the contributing area and S is the local slope. This assumption is based on previous observations showing: (i) a power law relationship between mean velocity V and discharge Q measured at different locations along a river network for some characteristic discharge of constant frequency (known as downstream hydraulic geometry - HG, see *Leopold and Maddock*, 1953 for details); (ii) power law dependence between characteristic discharges and contributing area (e.g. see *Stall and Fok*, 1968; *Rodriguez-Iturbe and Rinaldo*, 1997 and *Dodov and Foufoula-Georgiou*, 2004 among others), and (iii) power law dependence of local channel slope on contributing area (e.g. see *Rodriguez-Iturbe and Rinaldo*, 1997 and references therein). It is noted that although based on the above observations the velocity at any pixel can be expressed solely in terms of contributing area, such a relationship would ignore the information about the local slopes available from the analysis of the DEM. Here we have retained the effect of the variability of the local slopes by explicitly using them in the velocity computations.

  The parameters a and b can be estimated from measurements of V corresponding to some characteristic discharge at many locations (gauging stations) along a river network if the local channel slopes at these locations are known. Specifically, these parameters are computed from a least squares fit of the relationship $\ln(V_i S_i^{-1/2}) = \ln a + b \ln A_i$ where V_i, S_i

and A_i are respectively the mean velocity, the channel slope and the area draining to location *i*.

- Third, the travel time to the outlet is computed for every pixel as:

$$T_i = \sum_{j \in W_D} L_j / a A_j^b S_j^{0.5} \tag{5}$$

where W_D is the set of all pixels downstream of pixel *i*. To avoid infinite at-pixel travel times, zero-valued local slopes are first replaced by the local averages

$$S_j = \left[\sum_{k \in L_M}^{N_M} S_k \right] \Big/ N_M \tag{6}$$

of all pixels *k* that are upstream and downstream from the pixel *j* along the main channel with an averaging length L_M chosen to be scale dependent, i.e. $L_M \propto A_j^{0.5}$, where A_j is the area draining to pixel *j*.

- Finally, the response function of any arbitrary basin/subbasin is then extracted as the normalized travel time histogram of all pixels covered by the corresponding areas. Only travel times with frequencies 5% to 95% are considered in order to obtain reasonable length of the state vectors.

To better understand our approach, the reader should consider the fact that the response function of a basin or subbasin is actually a descriptor of the "importance" of the rainfall input over the basin area at moment $(t-\ell)$ with respect to the output from the system at moment *t*. In other words, rainfall at moment $(t-\ell)$ is seeing as contributing to the streamflow at moment *t* with weight $w_{t-\ell} \equiv U_{t-\ell}$, i.e., with weight equal to the ordinate of the normalized response function at time $(t-\ell)$. Consequently, if we measure how "different" or "similar" two rainfall events are with respect to the streamflow they produce at the basin outlet at moment *t*, the weight $w_{t-\ell}$ can be applied to the difference between the $(t-\ell)$-th components of the state vectors. For the case of the whole basin the distance between two states is defined similar to the standard L_p norm (see Appendix I) but "weighted" by the response function, i.e.

$$L_p = \left\{ \sum_{\ell=0}^{T_{\max}} w_{t-\ell} (e_{t-\ell})^p \right\}^{1/p} \tag{7}$$

where $e_{t-\ell}$ is the difference between any two state vectors at time $(t-\ell)$, and T_{max} is the maximum time of response. In the case of N_S subbasins the distance between two states is computed as:

$$L_p = \left\{ \left[\sum_{i=1}^{N_S} \sum_{\ell=0}^{T_{\max}} w_{t-\ell}^i (e_{t-\ell}^i)^p \right] \Big/ N_S \right\}^{1/p} \tag{8}$$

where T_{max} is the maximum response time for all subbasins. It is noted that the implementation of (7) and (8) defines a "hydrologically optimal" phase-space reconstruction and is proposed here as a replacement of the optimal reconstruction in the context of the "classical" nonlinear prediction.

Although the above idea is by itself new as it applies to nonlinear analysis of the rainfall-runoff process, we would like to mention its general conceptual similarity to other dimensionality reduction approaches. Note that the discrete normalized response function can be considered as a multidimensional unit vector, representing the "basis" on which our data has to be projected in order to be transformed to some "hydrological space" relevant to the streamflow at the catchment's outlet. In this context, the way we apply our weights to the difference between two state vectors is equivalent to the way the state vectors are projected onto the eigenbasis in the case of Principal Component Analysis (PCA) dimensionality reduction as it was done by *Broomhead and King*, [1986]. However, in our case the "basis" reflects the interaction between rainfall and land surface and includes information about basin geomorphology.

In order to treat the uncertainty related to hydraulic dispersion, we introduce the notion of "comparison within some tolerance". Let us look at the example rainfalls given in Fig. 2b. If we consider the two vectors formed by the history of each of rainfalls, say $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, as multidimensional points in Euclidean space, the difference between the points will be large. However, the effect of the rainfalls at times $(t-\ell_1)$ and $(t-\ell_2)$ on the streamflow at moment t will be quite similar if these rainfall pulses occur far away from the catchment outlet and ℓ_1 and ℓ_2 are large compared to the time of concentration of the basin. This is why we consider the difference between the ℓ-th components of the state vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ as:

$$e_\ell = \begin{cases} \boldsymbol{X}_{i,\ell} - \max\limits_{k\in(\ell-e,\,\ell+e)} \left\{\boldsymbol{X}_{j,k}\right\}, & \text{if } \boldsymbol{X}_{i,\ell} > \boldsymbol{X}_{j,\ell} \\ \boldsymbol{X}_{j,\ell} - \max\limits_{k\in(\ell-e,\,\ell+e)} \left\{\boldsymbol{X}_{i,k}\right\}, & \text{if } \boldsymbol{X}_{i,\ell} < \boldsymbol{X}_{j,\ell} \end{cases} \tag{9}$$

where e is a tolerance interval around the $(t-\ell)$-th time instance. In other words, when we compare the distance between the ℓ-th components of two vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, we take the largest value of $\boldsymbol{X}_{i,\ell}$ and $\boldsymbol{X}_{j,\ell}$, say $\boldsymbol{X}_{i,\ell}$, and compute the minimal difference between this value and the values of $\boldsymbol{X}_{j,k}$, $k \in \{l-e, l+e\}$. Thus, the difference between two strong events occurring in the "tail" of the system response function will be considered to be small if the time interval within which these events occur is less than $2e$. We consider e to be monotonically increasing with time back from the moment *t*, which in the discrete time case is a stepwise constant function, e.g. $e=0$ at moment $(t-1)$, $e=1$ at moment $(t-2)$, etc. In the following, we refer to the monotonically increasing e with the time back from the reference moment as the "uncertainty model". It is noted that once the "uncertainty" difference e_ℓ between the ℓ-th

components of state vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is computed from (9), any norm can be used to express the “similarity” between these two state vectors.

As was the case in the first technique, here too the distances between “rainfall states” and the distances between the corresponding “streamflow states” are measured differently (applying the above weights with respect to “rainfall states”, and for streamflow states by computing the Euclidean distance in the optimally reconstructed univariate phase-space) and cannot be considered together in a single phase-space. Therefore, similarly to T1, the two distances have to be projected on a “distance space” in order for the neighborhood of the collective reference space to be defined.

Acknowledging the fact that the dynamics of the rainfall-runoff process is not considered deterministic, and also due to the limited length of our data series and the high level of noise in the radar rainfall estimates, both techniques use the zero-th order prediction scheme described in section 2.

4. CASE STUDY

4.1. Basin characteristics and data used

In the present study, streamflow and rainfall data from the Chikaskia River basin, Oklahoma and the Gasconade River basin, Missouri, USA are used (see Fig. 3). The streamflow data are given as 6 hour average flows and the rainfall data consist of radar rainfall images representing 6 hour rainfall accumulations. Every pixel in the radar image corresponds to 4×4 km area. For every record in the database (say 05/10/1995 6:00 am) there is one streamflow value representing the mean flow between 0:00 and 6:00 am and one image of 6hr rainfall accumulations with ending time 6:00 am. The reliability of the streamflow and rainfall data is ensured by the fact that the flows are unregulated and that the basins are well within (at ranges of 40 to 120 km) the coverage of the KVNX, KICT and KSGF radar stations. This yields reliable radar rainfall estimates (within a radius of ≈ 40 km a "ground clutter" generally appears, and at distances longer then ≈ 140 km the estimates of only intensive rainfalls are reliable). As can be seen from Fig. 4, the response of both basins to the basin average rainfall is realistically reflected in the measured streamflow at the basin outlets.

In order to analyze the effect of geologic controls on the quality of prediction, the two basins were chosen to have completely different geologic environments: deep limestone aquifer underlying the Gasconade River basin and absence of dominating aquifer in the case of the Chikaskia River basin. The autocorrelation functions of the two streamflow time-series plotted in Fig. 5 clearly show the effect of geology on flow regime: while the autocorrelation function of the Gasconade River indicates high persistence due to a significant effect of a deep aquifer, in the case of the Chikaskia River the temporal “memory” of streamflow is much shorter.

Data from 37 USGS gauging stations in Kansas, Missouri, Oklahoma and Arkansas (see Fig. 3) were used for the estimation of the parameters *a* and *b* of relationship (5). The data consisted of: (i) independent measurements of maximum width, mean depth, cross-sectional area, mean velocity and discharge under different flow conditions (up to several hundred measurements per station), (ii) time-series of at least 10 years of unregulated daily discharges, and (iii) channel slope at every station. In Fig. 3. a plot of $VS^{-0.5}$ versus A is also given with the estimates of the parameters *a* and *b*.

4.2. Reduction of dimensionality (data preprocessing), choice of reconstruction parameters and derivation of a catchment-specific distance function

In order to analyze the effect of rainfall spatial aggregation on streamflow prediction accuracy, several different types of multivariate time-series were created: (1) 6-hr rainfall accumulations at the scale of the radar pixels covering the two basins (respectively 305 and 415 pixels); (2) average rainfall accumulations over 9/10 subcatchments respectively for Chicaskia/Gascodate River basins; (3) average rainfall accumulations over 3 subcatchments; and (4) average accumulations over each of the two basins. The dimensionality of the state vectors in the above four cases for the two basins respectively are 305/415; 9/10; 3 and 1.

All possible combinations of the number of neighbors, prediction horizons and embedding dimensions (only for the case of T1) were computed respectively for the intervals 1 to 50, 1 to15 and 1 to 20. In the case of T1, the similarity between every two states was evaluated by means of root mean square error (RMSE), multiscale root mean square error (MSRMS), and Hausdorff distance (HD) when radar images were considered for rainfall at pixel level (see Appendix I) and by means of Euclidean distances otherwise.

The response functions of the two basins (and the corresponding subbasins) were obtained following the procedure described in section 3.2. (see Fig. 6a,b for the basin-average response of the two catchments, and Fig. 6c for the gray-color coded spatial travel time distribution and the response functions of the 3 subcatchments of Chikaskia River basin).

To incorporate the response functions for prediction n_T time steps ahead, we incorporate only the weights $w_{t-\ell}$ for which $\ell > n_T$. Thus the distances between rainfalls, contributing maximally to the runoff at time $(t+n_T)$ will be enhanced, i.e. given more weight. For example, for prediction at time two time steps ahead (n_T = 2), the distances are multiplied by the coefficients w_3 to w_{16}, for three time steps ahead (n_T = 3) by the coefficients w_4 to w_{17}, etc. Since streamflow prediction at lead times larger than 10/15 (10/15 means 10 for Chikaskia and 15 for Gasconade River) time steps is mostly related to prediction of future rainfall (since the “old” rainfall has already drained out of the basin), we restrict ourselves to prediction horizon of n_T = 10/15 time steps.

In order to account for the uncertainty related to hydraulic dispersion, we consider the fact that the longer the hydraulic path of a particle in the catchment, the longer the time interval of its possible arrival at the outlet. This in turn means that the larger the uncertainty in the time of its possible occurrence at a specific location in the catchment for a given prediction time horizon at the outlet. This translates to an increasing uncertainty (and thus greater tolerance) for increasing time difference between the occurrence of a rainfall pulse and its appearance at the outlet (called here "time lag"). Thus tolerance can be expressed as a step function whose height increases with "time lag" (the uncertainty model in section 3.2). The graphs of the tolerance thresholds we apply are given in Fig. 6a,b.

A summary of the state vectors and similarity distances used, and their notations are given in Table 1.

4.3. Prediction accuracy improvement by incorporating rainfall in the nonlinear prediction

In this work we are interested not only in assessing maximum predictability, but also in better understanding the effects of the spatial distribution of rainfall and catchment geomorphology on the accuracy of nonlinear prediction. More specifically, we would like to explore how the interplay between the spatial distribution of rainfall, time of concentration of a basin and prediction horizon affect streamflow predictability. For this reason, our first step is to analyze what lead time (prediction horizon) gives maximum difference between prediction based on only streamflow data and prediction based on rainfall and streamflow measurements. In Fig.7 we plot the prediction accuracies, measured as correlation between the true and predicted series, based on different types of state vectors and different types of similarity measures (see Table 1), along with the "improvement" in prediction, where the "improvement" is measured in comparison to prediction based solely on streamflow.

Clearly, for short lead times (1-4 time steps, i.e., 6 to 24 hours) the improvement is negligible and, in the case of Gasconade River, introducing rainfall information gives worse prediction than using streamflow data only (shown as negative improvement in Fig. 7). This result coincides with the result reported by *Porporato and Ridolfi* [2001], where the improvement was of the order of 1%. Maximal improvement is achieved at lead times of 6/9 steps (i.e. 30 hrs for Chikaskia and 54 hrs for Gasconade River), followed by fast decrease in the differences between the prediction accuracies based on rainfall and streamflow observations combined versus streamflow observations only. Three facts could explain such a behavior: (1) Lead times of 6/9 (Chikaskia/Gasconade) steps coincide with the corresponding mean travel times of the two catchments (see Fig. 6). Thus the information provided about the spatial distribution of rainfall will have maximum effect for these prediction horizons; (2) High correlation in the streamflow data at time lags up to 3-4/5-6 time steps and sharp decrease in the

correlation after that (see Fig.5) makes rainfall information redundant for lead times up to approximately 4/6 time steps. Thus, for prediction up to 4/6 time steps ahead, streamflow predictability is dominated by the persistence in streamflow data; (3) For prediction horizons longer than 7/10 time steps the spatial patterns of rainfall, contributing mostly to the future flows, are not yet observed. Therefore, for such lead times the uncertainty, and respectively the prediction error, grows rapidly.

Incorporating information about the spatial distribution of rainfall gives an improvement in the prediction accuracy of ~10 to 15 % for Chikaskia River and ~3 to 8 % for Gasconade River. Naturally the question arises as to whether this improvement is significant. In order to measure the statistical significance of the result, we evaluate the correlation coefficient for which the linear dependence is still significant. This critical level for the correlation coefficient r_a is given by:

$$r_a = t_a / \sqrt{t_a^2 + N - 2} \tag{10}$$

where t_a and $\boldsymbol{N}$ denote respectively the two-sided a percentile of the Student's t distribution and the sample size. For data size of $\boldsymbol{N}$ = 3615 (all points for which rainfall happens in the last 5 time steps over the Chikaskia River basin) and confidence interval a = 0.05 the critical correlation coefficient is 0.033. For Gasconade River the values are respectively $\boldsymbol{N}$ = 3939 and the critical correlation coefficient 0.031. Therefore the improvement can be considered statistically significant.

The aforementioned improvement is significantly larger (but for longer prediction horizons) than the one reported by *Porporato and Ridolfi* [2001], suggesting that: (i) The information about rainfall should be incorporated into the nonlinear prediction of streamflow, with an aggregation of the detail of the spatial distribution of rainfall, rather than by incorporation of point measurements at fixed locations of the basin, and (ii) nonlinear prediction of streamflow has to be considered in the context of its relation to such important factors as catchment geomorphology and underlying geology.

For the purpose of comparison, the results of all analyses are reported with the same number of nearest neighbors of 30. This is because for most of the cases this number provided best prediction accuracy (see Fig. 8 for the case of embedding dimension 14 and lead times 6/9).

4.4. Effect of the embedding dimension

In the case of both T1 and T2 techniques, the optimal embedding dimension for the streamflow was found to be 4 for Chikaskia and 6 for Gasconade River, and this result was shown to be independent of the change of embedding dimension of rainfall in the case of T1. Fig. 9 shows the prediction accuracy for different distance measures as a function of the embedding dimension of rainfall for prediction horizons 6/9 (Chikaskia/Gasconade) time steps,

and keeping the embedding dimension of streamflow 4/6. The prediction is better for small embedding dimensions of rainfall and gradually decreases for longer state vectors. Notice that for small embedding dimensions the prediction based on the mean rainfall is comparable to that based on different catchment subdivision schemes, while for larger lengths of the state vectors the discrepancy rapidly grows. These two facts could be expected for the following reasons. First, for lead time 6/9 steps and embedding dimension up to 6, according to Fig. 6, only the most important rainfalls are taken into account. For longer state vectors the probability to find similar pairs decreases rapidly and the effect of the noise in the data dominates. Second, the information about the rainfall included in the prediction is mostly related to events corresponding to the mean response of the system, i.e. rainfalls, whose major intensity is located near the "center" of the catchment's geometry.

Increasing the embedding dimension incorporates more and more information about the rainfall back in time, thus accounting more precisely for events, whose major intensity is located far upstream from the streamgage. This type of events will contribute to a significant discrepancy between the accuracy of predictions including, versus not including information about the spatio-temporal distribution of rainfall. Such a trend is clearly shown in Fig. 9 for embedding dimensions larger then 6/9.

Our results are in good agreement with the conclusions of *Porporato and Ridolfi* [2001], who found that the optimal embedding dimension of rainfall was 2 days, i.e. the time which enables one to capture the entire subsequent runoff.

4.5. Effects of the treatment of catchment geomorphology and uncertainty

Including a geomorphologically derived response function as a tool for comparison of the rainfall events definitely improves the accuracy of streamflow prediction (see Fig. 10). Improvement is superior in the case where the prediction is based on streamflow and mean rainfall over the basin, which could be expected considering the fact that the IUH is derived according to the "mean" response of the (sub)catchment and does account for the change of hydrologic response with the change in the forcing.

If in addition we include treatment of uncertainty due to unobserved processes in the catchment by means of an uncertainty model, an additional improvement in prediction accuracy is observed, especially in the case where the mean rainfall over the catchment is considered. This result is a consequence of the fact that our uncertainty model is also chosen according to the mean response of the system. Independently of the physical process by which a particular rainfall event contributes to the streamflow at the outlet, the uncertainty (coming from all the interactive processes in the catchment) is much higher when the event occurs far from the outlet. This, in turn, is projected to an uncertainty that depends on the spatial and temporal position of a rainfall event relative to the catchment outlet.

4.6. Effect of underlying geology

The improvement in prediction accuracy by including information about the spatial distribution of rainfall is different for the two catchments (~10 to 15 % for Chikaskia River and ~3 to 8 % for Gasconade River). This difference is explainable by the fact that the limestone aquifer underlying the Gasconade basin, acts as a large reservoir and smooths the spatial and temporal fluctuations of the input over large space-time scales. As a result, the temporal correlation of the streamflow time series increases at the Gasconade River (see Fig. 5) and the improvement in streamflow prediction due to the information about the spatial distribution of rainfall decreases. This fact suggests that underlying geology is an important factor which has to be taken into account in long-term phase-space nonlinear prediction of streamflow.

4.7. Predictability as a function of scale of aggregation

In Fig. 10 we analyze the improvement in streamflow predictability due to information about the rainfall input as a function of the scale of aggregation of the input. Obviously, incorporating information about the spatial distribution of the rainfall significantly improves the accuracy of nonlinear phase-space prediction. The accuracy is affected by two opposite trends: (1) Including more detail about the spatial distribution of rainfall improves streamflow prediction; (2) With an increase in dimensionality the chance to find "similar" rainfall events, even with measures such as the MSRMS and HD decreases. Thus, the best method appears to be a compromise between the detail, necessary to explain the space-time variability of the process, and the complexity and uncertainty in the data. The best result is achieved by the subdivision scheme of 3 and 9/10 subcatchments, followed by MSRMS, Hausdorff distance and uniform error (RMSE) based norms.

Based on the results of this work, we can conclude that incorporating information about spatially distributed forcing and catchment geomorphology in the phase-space approach, together with a treatment of the uncertainty due to unobserved processes in the catchment is a promising approach, worth of further investigation.

5. SUMMARY AND CONCLUSIONS

This work proposes a new approach for incorporating the spatial variability of rainfall and basin geomorphology into rainfall-runoff nonlinear analysis and prediction. Specifically, it proposes a framework for nonlinear prediction of streamflow, based on "hydrologically relevant" rainfall-runoff phase-space reconstruction acknowledging the fact that rainfall-runoff is a stochastic and spatially extended rather than deterministic and multivariate system. In contrast to subsampling the spatially variable forcing (rainfall) at a few fixed locations and applying

“classical” multivariate techniques for nonlinear prediction, the proposed methodology incorporates information about the spatially distributed forcing and the hydrologic response of the system to study the interplay of spatio-temporal rainfall, catchment geomorphology and underlying geology as they affect the dynamics (and thus predictability) of streamflow.

A geomorphologically derived response function was used as a tool for comparison of rainfall events in terms of their “hydrologic” similarity. A catchment specific uncertainty model was introduced in order to account for the hydraulic dispersion in the catchment occurring a posteriori to the rainfall input. The proposed framework can be seen as one that allows the incorporation of a simple conceptual rainfall-runoff model into a phase-space approach for nonlinear prediction of streamflow. It can also be seen as blending deterministic and stochastic considerations via appropriate “tolerance” distance functions in the phase-space of the system.

The developed methodology was applied to streamflow and rainfall data from the Chikaskia River basin, Oklahoma and the Gasconade River, Missouri, USA. The streamflow data used represent 6 hour average flows and the rainfall data consisted of radar rainfall images of 6 hour rainfall accumulations. In order to analyze the impact of the detail at which the spatial distribution of rainfall is specified on the quality of streamflow prediction, several types of time-series with different spatial resolution were analyzed: from 6-hr accumulations at 4x4 km pixel level, through aggregation over subcatchments of different spatial extent, to the average rainfall over the whole catchment. In order to analyze the effect of geologic control on the quality of streamflow prediction, the two basins were chosen to have completely different geologic environments: deep limestone aquifer underlying the Gasconade River basin and absence of dominating aquifer in the case of Chikaskia River.

The most important results based on analyses of data from the two catchments are:

- Including information about rainfall improves streamflow predictability (up to 18% in the case of Chikaskia River and up to 6% in the case of Gasconade River) compared to prediction based on only streamflow data.
- The information about the spatial distribution of rainfall increases the prediction accuracy (up to 12% for Chikaskia River and up to 5% for Gasconade River), compared to prediction based on mean rainfall over the whole basin.
- Best results are achieved by catchment subdivision schemes of 3 and 9/10 subcatchments respectively, for the two basins. Thus rainfall averaging over subcatchments of different subdivisions is the optimal tradeoff between the detail, necessary to explain the rainfall variability in space and time, the complexity of the problem and uncertainty in the data.
- Including information about (sub)catchment’s response as a tool for comparison of rainfall events improves additionally the accuracy of streamflow prediction by 2 to 10 %.
- If in addition we include treatment of uncertainty due to unobserved processes in the catchment (stochasticity), we increase prediction accuracy by an additional 2 to 5 %.

- The difference in prediction accuracy due to the information about the spatial distribution of rainfall depends on the geologic characteristics of the basin. This is suggested by the fact that more improvement is achieved in the case of Chikaskia River (~10 to 15 %), while in the case of Gasconade River the prediction accuracy is significantly affected by the smoothing effect of the underlying limestone aquifer (improvement ~3 to 8 %).

The main implications of the presented work, given the above results are:

- The information about the spatial distribution of rainfall definitely increases the accuracy of nonlinear streamflow prediction, but has to be considered in the context of its relation to such important factors as catchment geomorphology and underlying geology.
- The proposed methodology for nonlinear prediction based on a catchment-specific distance function and treatment of uncertainty gives the best quality of prediction and overcomes the shortcomings of the "classical" multivariate technique for streamflow prediction as applied to a spatially extended system such as rainfall-runoff. The improvement (of approximately 3 to 15%) suggests that this line of research needs to be explored further. We envision that the proposed methodology can be extended to a generalized framework, by which catchment-specific geomorphologic information, uncertainty in the runoff generation process and channel hydraulic geometry can be meaningfully considered in the context of nonlinear rainfall-runoff analysis and prediction.

In view of the enormous amount of information about the spatial distribution of rainfall collected every day by ground radars and the still unresolved question as to how much this information enhances streamflow prediction when used in distributed rainfall-runoff models [e.g. see *Obled et al.*,1994; *Bell and Moore*, 2000], an approach such as the one proposed herein might provide an alternative worth of further investigation, as it is not specific to a particular rainfall-runoff model and its scale dependent calibration.

Acknowledgments

This research was partially supported by a NASA/NOAA grant (NAG8-1519) and by an NSF grant (EAR-0120914) as part of the National Center for Earth Surface Dynamics (NCED) at the University of Minnesota. Computational resources were kindly provided by the Minnesota Supercomputing Institute (MSI). We thank two anonymous reviewers for their thoughtful comments which resulted in an improved presentation.

APPENDIX Measures of Similarity Between Rainfall Patterns

The most common way to evaluate similarity between two image functions is to use the uniform error function [*Sendov*,1990] and its associated distance measures:

$$e_i = \left| f_i - g_i \right|$$

where i is the point where the functions f and g are defined (considering for simplicity the 1D case) and e is the uniform error function. The corresponding distance norm (known also as L_P norm) is:

$$L_p = \left\{ \frac{1}{N} \sum_{i=0}^{N-1} e_i^p \right\}^{1/p}$$

where N is the support length. The most commonly used norms are for p = ¥, $p = 1$ and $p = 2$ (the last known as root mean square error - RMSE), that is:

$$L_\infty = \sup_{i=0,1,\ldots,N-1} e_i,\ L_1 = \frac{1}{N} \sum_{i=0}^{N-1} e_i,\ L_2 = \left\{ \frac{1}{N} \sum_{i=0}^{N-1} e_i^2 \right\}^{1/2}$$

It is easy to see that the L_2 distance norm between two images is equivalent to distance between two points in N dimensional Euclidean space, where N is the number of pixels in the images. Thus, even a small translation between a rainfall image and its copy will produce significant distance. Obviously, the L_P norm is not the best choice for comparison of rainfall images. In order to meet the specifics of rainfall "similarity" relative to runoff production, we apply two other types of distances between rainfall images: (i) Multiscale version of the L_2 norm for the pixels covering the basin, and (ii) L_2 norm defined for Hausdorff metric. These distances are explained below.

<u>Multi-Scale Root Mean Square Error (MSRMS)</u>

We define a multiscale version of the uniform error as $e_i^{y_S} = \left| f_i^{y_S} - g_i^{y_S} \right|$, where $f^{y_S} = \Psi_S * f$ and $g^{y_S} = \Psi_S * g$ are respectively the convolutions between the functions f and g and some smooth, symmetric function Y_S with compact support determined by some scale factor S, satisfying the condition $\int \Psi_S(x)\ dx = 1$. Since we are comparing only the part of the images covering the catchment area, the only way for properly normalizing the smoothing function without including information outside the area of interest is to apply a simple "Box" function satisfying $\int B_S(x)\ dx = 1$. In other words, we apply a moving average of varying support S from a set of desired supports $S_k := \left\{ 2^i \times 2^i ; i = 0,\ldots,n_S \right\}$, with respect to the two functions and then compare them by means of the distance norm

$$L_p = \left\{ \frac{1}{(n_S+1)N_p} \sum_{S \in S_k} \sum_{i=0}^{N_p-1} \left(e_i^{y_S} \right)^p \right\}^{1/p}$$

where n_S is the number of different support lengths of the moving average and N_p the number of pixels covering the basin. The intuitive reasoning for applying this similarity measure is that performing smoothing at different scales will "redistribute" spatially the "mass" of the sharp anomalies and thus will decrease the distance between similar but slightly translated identical geometries.

<u>Hausdorff distance based norm (HD)</u>

Given two sets of points $A = \{a_1, \ldots, a_m\}$ and $B = \{b_1, \ldots, b_m\}$, the Hausdorff distance is defined as

$$\mathrm{H(A,B)} = \max\{\, \mathrm{h(A,B)}, \mathrm{h(B,A)} \,\}, \text{ where } \mathrm{h(A,B)} = \max_{a \in A} \left\{ \min_{b \in B} \|a - b\| \right\}$$

The function $h(A,B)$ is called the directed Hausdorff distance from *A* to *B*. It identifies the point $a \in A$ that is farthest from any point of *B*, and measures the distance from *a* to its nearest neighbor in *B*. Thus the Hausdorff distance, $H(A,B)$, measures the degree of mismatch between two sets, as it reflects the distance of the point of *A* that is farthest from any point of *B* and vice versa. Intuitively, if the Hausdorff distance is *d*, then every point of *A* must be within a distance *d* of some point of *B* and vice versa. It is well known that the Hausdorff distance is a metric over the set of all closed, bounded sets - it obeys the properties of identity, symmetry and triangle inequality [*Sendov*,1990].

The concept of Hausdorff distance was extended in *Sendov* [1990] and *Marinov* [1993], to comparison of bounded functions, considering their "completed graphs". Since we are interested in comparing rainfall images let us consider the completed graph of such an "image function". The completed graph consists of horizontal pixel areas and vertical walls between pixels of different rainfall intensity. Let us assume that every horizontal and vertical area is represented by an infinitely dense set of points covering that area. Then the completed graph of the image function will be the union of all subsets covering such flat areas. Thus the rainfall image is represented as a set of points in 3D and the "distance" between two images is measured not only in the intensity, but also in the spatial directions. *Sendov* [1990] introduced the "Hausdorff error" for quantifying local similarity:

$$h_i = \max\{ \min \left\| I_i^f - C_g \right\|, \min \left\| I_i^g - C_f \right\| \}$$

where I^f and I^g are "image" functions and C_f and C_g are their completed graphs.

Norms similar to those defined for the regular metric can be applied for the Hausdorff metric:

$$R_\infty = \max_{i=0,1,\ldots,N-1} h_i,\ R_1 = \frac{1}{N}\sum_{i=0}^{N-1} h_i,\ R_2 = \left\{\frac{1}{N}\sum_{i=0}^{N-1} h_i^2\right\}^{1/2}$$

respectively for $p = \infty$, 1 and 2 and for the general case

$$R_p = \left\{\frac{1}{N}\sum_{i=0}^{N-1} h_i^p\right\}^{1/p}$$

REFERENCES

[1] Babovic, V., and M. Keijzer, Forecasting of River discharges in the presence of chaos and noise. In *Flood issues in contemporary water management*, edited by J. Marsalek, Kluwer, Dordrecht, pp. 405-419, 2000.

[2] Bell, V. A. and R. J. Moore, The sensitivity of catchment runoff models to rainfall data at different spatial scales, *Hydrology and Earth System Science*s, 4, 653-667, 2000.

[3] Bohr, T., M. H. Jensen, G. Paladin and A. Vulpiani, *Dynamical systems approach to turbulence*, Cambridge University Press, 1998.

[4] Broomhead, D., and G. P. King, Extracting qualitative dynamics from experimental data, *Physica D*, 20, p. 217, 1986.

[5] Cao, L., A. Mees and K. Judd, Dynamics from multivariate time series, *Physica D*, 121, pp. 75-88, 1998.

[6] Casdagli, M., S. Eubank, J. D. Farmer and J. Gibson, State space reconstruction in the presence of noise, *Physica D*, 51, p. 52, 1991.

[7] Chow, V. T., Maidment, D. R. and L. W. Mays, *Applied hydrology*, McGraw-Hill Higher Education, 1988.

[8] Dodov, B. and E. Foufoula-Georgiou, Incorporating catchment geomorphology into nonlinear analyses of streamflow dynamics, paper presented at *AGU Fall meeting*, 2001.

[9] Dodov, B., and E. Foufoula-Georgiou, Generalized hydraulic geometry: Derivation based on a multiscaling formalism, *Water Resour. Res.*, 40 (6), W06302, doi:10.1029/ 2003WR002082, 2004.

[10] Farmer, J.D., and J. J. Sidorowich, Predicting chaotic time series, *Phys. Rev. Letters*, 59, pp. 845-848, 1987.

[11] Georgakakos, K.P., Sharifi, M.B., Sturdevant, P.L., 1995. Analysis of high-resolution rainfall data. In: Kundzewicz, Z.W. (Ed.), *New Uncertainty Concepts in Hydrology and Water Resources*, Cambridge University Press, New York, pp. 114–120.

[12] Hense, A., On the possible existence of a strange attractor for the southern oscillation, *Beitr. Phys. Atmos., 60 (1)*, 34–47, 1987.

[13] Islam, M. N., and B. Sivakumar, Characterization and prediction of runoff dynamics: a nonlinear dynamical view, *Adv. Water Resour.*, 25, pp. 179–190, 2002.

[14] Jayawardena, A.W., and F. Lai, Analysis and prediction of chaos in rainfall and stream flow time series, *J. Hydr., 153*, 23–52, 1994.

[15] Kantz, H., and T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press, Cambridge, 1997.

[16] Koutsoyiannis, D., and D. Pachakis, Deterministic chaos versus stochasticity in analysis and modeling of point rainfall series, *J. Geophys. Res.*, 101 (D21), 26 441–26 451, 1996.

[17] Leopold, L. B. and T. Maddock, The hydraulic geometry of stream channels and some physiographic implications, *U.S Geol. Survey Prof. Paper*, 252, 57p., 1953.

[18] Liu, Q., S. Islam, I. Rodriguez-Iturbe, and Y. Le, Phase-space analysis of daily streamflow: characterization and prediction, *Adv. Water Resour., 21*, 463–475, 1998.

[19] Maidment, D.R., F. Olivera, A. Calver, A. Eatherall and W. Fraczek, A unit hydrograph derived from a spatially distributed velocity field, *Hydrological Processes*, Vol. 10, 831 - 844, 1996.

[20] Marinov, P., *Hausdorff approximations of functions*, Ph.D. Thesis, Mathematical Institute, Bulgarian Academy of Sciences, 1993.

[21] Obled, Ch., J. Wardling and K. Beven, The sensitivity of hydrological models to spatial rainfall patterns: an evaluation using observed data. *J. of Hydrol.*, 159, 305-333, 1994.

[22] Parlitz, U. and C. Merkwirth, Prediction of spatiotemporal time series based on reconstructed local states, *Physical Review Letters*, V 84(9), 2000.

[23] Pasternack, G. B., Does the River run wild? Assessing chaos in hydrological systems, *Adv. Water Resour.*, 23, pp. 253-260, 1999.

[24] Phoon, K. K., M. N. Islam, C. Y. Liaw, and S. Y. Liong, A practical inverse approach for forecasting of nonlinear time series analysis. *J. Hydrol. Engg., ASCE*, 7(2), 116-128, 2002.

[25] Porporato, A., and L. Ridolfi, Clues to the existence of deterministic chaos in River flow, *Int. J. Mod. Phys., B 10*, 1821–1862, 1996.

[26] Porporato, A., and L. Ridolfi, Nonlinear analysis of River flow time sequences, *Water Resour. Res., 33 (6)*, 1353–1367, 1997.

[27] Porporato, A., and L. Ridolfi, Multivariate nonlinear prediction of River flows. *J. Hydrol.*, 248, 109-122, 2001.

[28] Rodriguez-Iturbe, I., F. B. De Power, M. B. Sharifi, and K.P. Georgakakos, Chaos in rainfall, *Water Resour. Res., 25 (7)*, 1667–1675, 1989.

[29] Rodriguez-Iturbe, I., and A. Rinaldo, *Fractal River Basins*, Cambridge Univ. Press, New York, 547 p., 1997.

[30] Sauer, T. 1993 *Time series prediction using delay coordinate embedding.* Time Series Prediction, ed. by A.S. Weigend and N.A. Gershenfeld, SFI Studies in the Sciences of Complexity, Proc. XV, 175, Addison-Wesley, Reading, Mass.

[31] Schertzer, D., I. Tchiguirinskaia, S. Lovejoy, P. Hubert, H. Bendjoudi and M. Larchevêque, Which chaos in the rainfall–runoff process? Discussion of "Evidence of chaos in the rainfall–runoff process", *Hydrol. Sci. J.*, 47(1), p. 139-148, 2002.

[32] Sendov, B., *Hausdorff approximations*, Kluwer Academic, 1990.

[33] Sharifi, M.B., K. P. Georgakakos, and I. Rodriguez-Iturbe, Evidence of deterministic chaos in the pulse of storm rainfall, *J. Atmos. Sci., 47*, 888–893, 1990.

[34] Sivakumar, B., S.-Y. Liong and C.-Y. Liaw, Evidence of chaotic behavior in Singapore rainfall, *J. Am. Water Resour. Assoc.*, 34 (2), 301–310, 1998.

[35] Sivakumar, B., R. Berndtsson, J. Olsson, and K. Jinno, Evidence of chaos in rainfall-runoff process, *Hydrological Sciences Journal, 46(1)*, 131-146, 2001.

[36] Sivakumar, B., A. W. Jayawardena, and T. M. G. H. Fernando, River flow forecasting: Use of phase-space reconstruction and artificial neural networks approaches. *J. Hydrol.*, 265(1-4), p. 225, 2002.

[37] Stall J. B., and Y. S. Fok, Hydraulic geometry of Illinois streams, *Ill. Water Resour. Res. Rep.*, 15, 47 p., 1968.

[38] Takens, F. Detecting strange attractors in turbulence, *Lect Notes Math. no.898*, Springer Verlag, Berlin, 1981.

[39] Tsonis, A. A., J. B. Elsner, and K.P. Georgakakos, Estimating the dimension of weather and climate attractors: important issues about the procedure and interpretation, *J. Atmos. Sci.*, *50*, 2549-2555, 1993.

[40] Yakowitz, S., Nearest neighbor methods for time series analysis, *J.Time Series Analysis, 8(2)*, 235-247, 1987.

[41] Yakowitz, S., and M. Karlsson, Nearest neighbor methods for time series with application to rainfall-runoff prediction, In *Stochastic Hydrology*, edited by J. B. MacNeil & G. H. Humphries, 149-160, Reidel, Hingham, Massachusetts, 1987.

[42] Wang, Q., and T. Y. Gan, Biases of correlation dimension estimates of streamflow data in the Canadian prairies, *Water Resour. Res., 34 (9)*, 2329-2339, 1998.

Table 1. State vectors and similarity distances used in phase-space reconstruction and nonlinear prediction.

Notation	Description	
	State vector	Similarity measure
F	Successive streamflow measurements	Euclidean distance (RMSE)
$\mathbf{FR_{A}}$	Successive streamflow measurements and basin average rainfall	Euclidean distance (RMSE)
$\mathbf{FR_{3C}}$	Successive streamflow measurements and 3 subcatchments average rainfall	Euclidean distance (RMSE)
$\mathbf{FR_{9/10C}}$	Same as above with 9 or 10 subcatchments	Euclidean distance (RMSE)
$\mathbf{FR_{A,M}}$	Successive streamflow measurements and basin average rainfall	Euclidean distance (RMSE) with morphology
$\mathbf{FR_{3C,M}}$	Successive streamflow measurements and 3 subcatchments average rainfall	Euclidean distance (RMSE) with morphology
$\mathbf{FR_{9C,M}}$	Same as above with 9 or 10 subcatchments	Euclidean distance (RMSE) with morphology
$\mathbf{FR_{RMS}}$	Successive streamflow measurements and radar rainfall images at 4 km resolution	RMSE
$\mathbf{FR_{MSRMS}}$	Same as above	Multiscale RMSE (MSRMS)
$\mathbf{FR_{HD}}$	Same as above	Hausdorff distance based norm

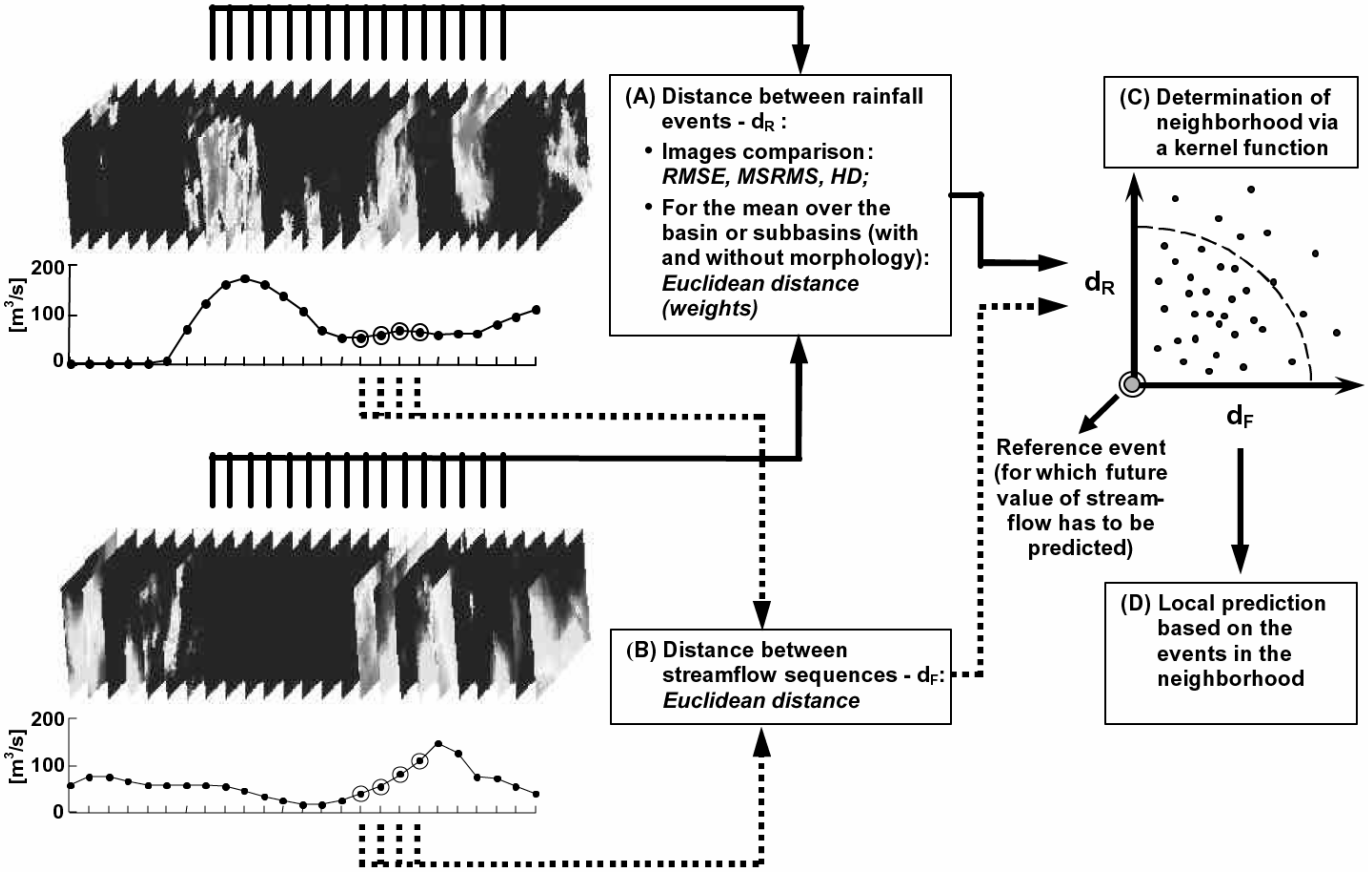


Fig. 1. General scheme (consisting of steps A, B, C and D) for rainfall-runoff phase-space reconstruction and nonlinear prediction. After the distances between rainfall states (step A) and streamflow states (step B) are computed, they are projected in a ”distance space” in which neighborhoods are defined and used for local prediction. See text for further details.

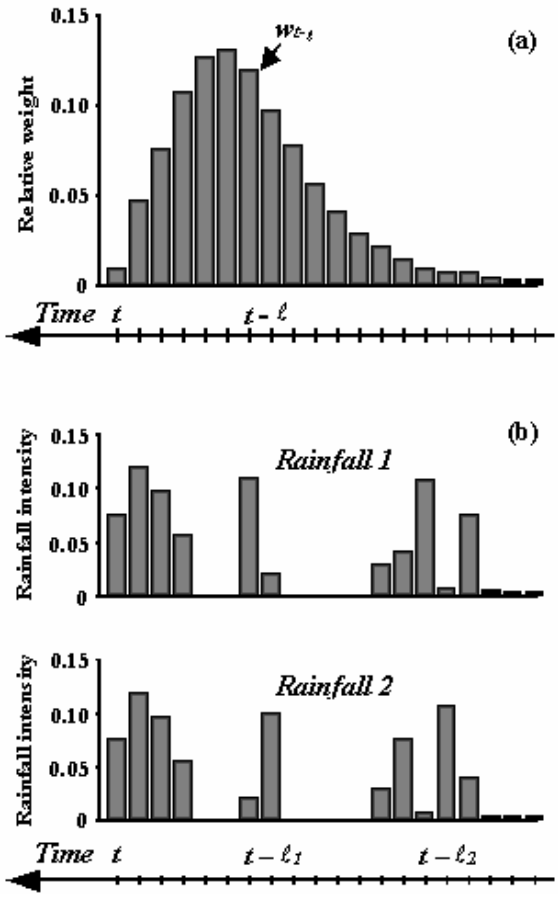
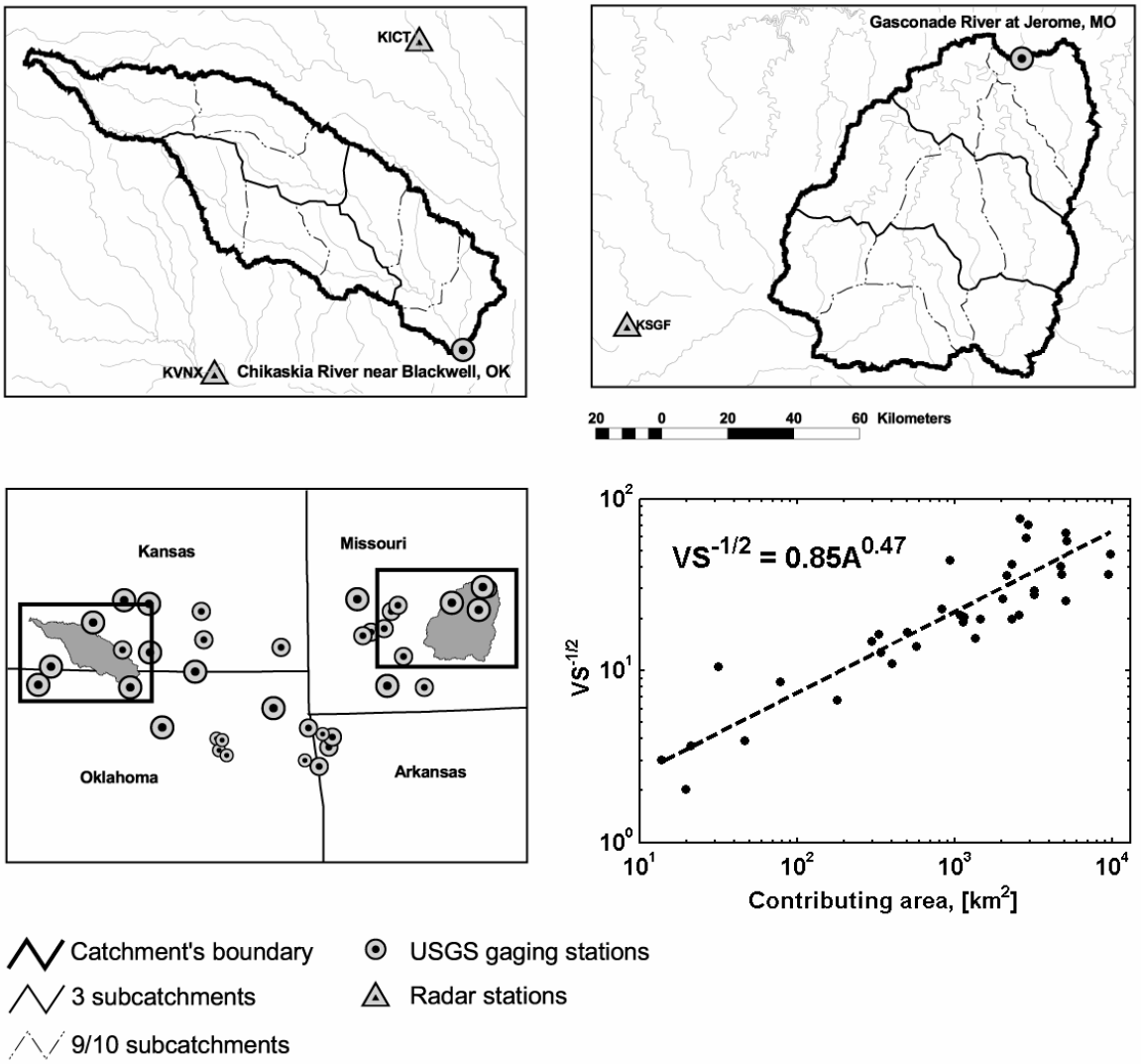
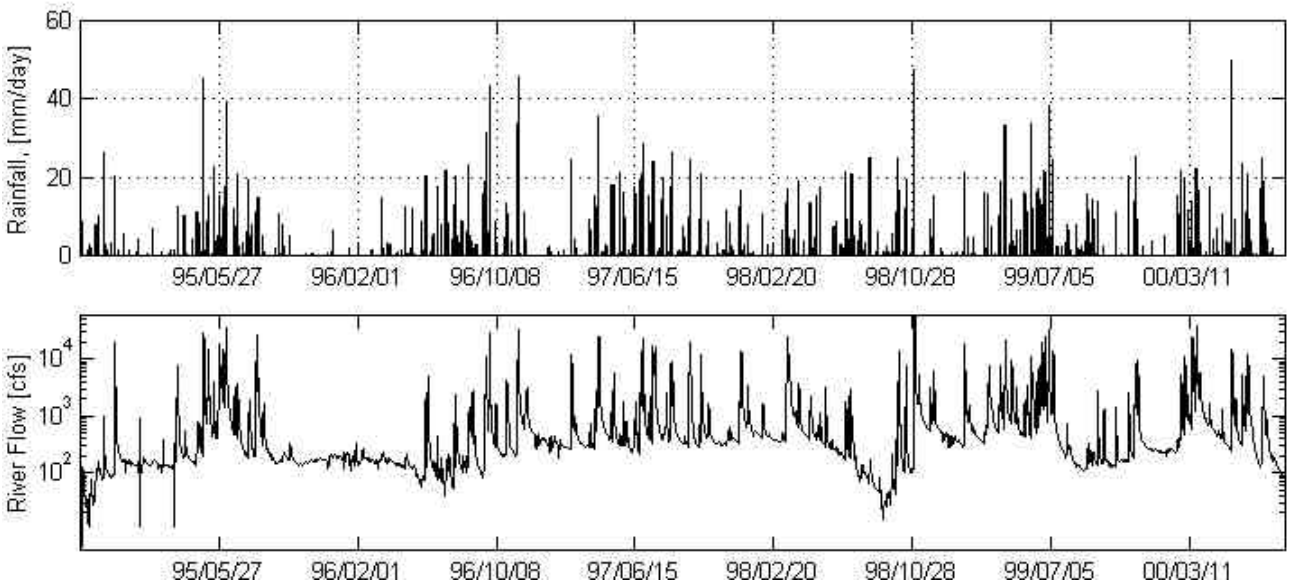


Fig. 2. (a) Example of normalized unit hydrograph which acts as a weighting kernel on the rainfall sequences; (b) Example of two rainfall sequences which are "similar" with respect to their effect on the runoff at moment *t*, but different, if considered in terms of their Euclidean distance in a multidimensional space (e.g. RMSE between the two sequences).

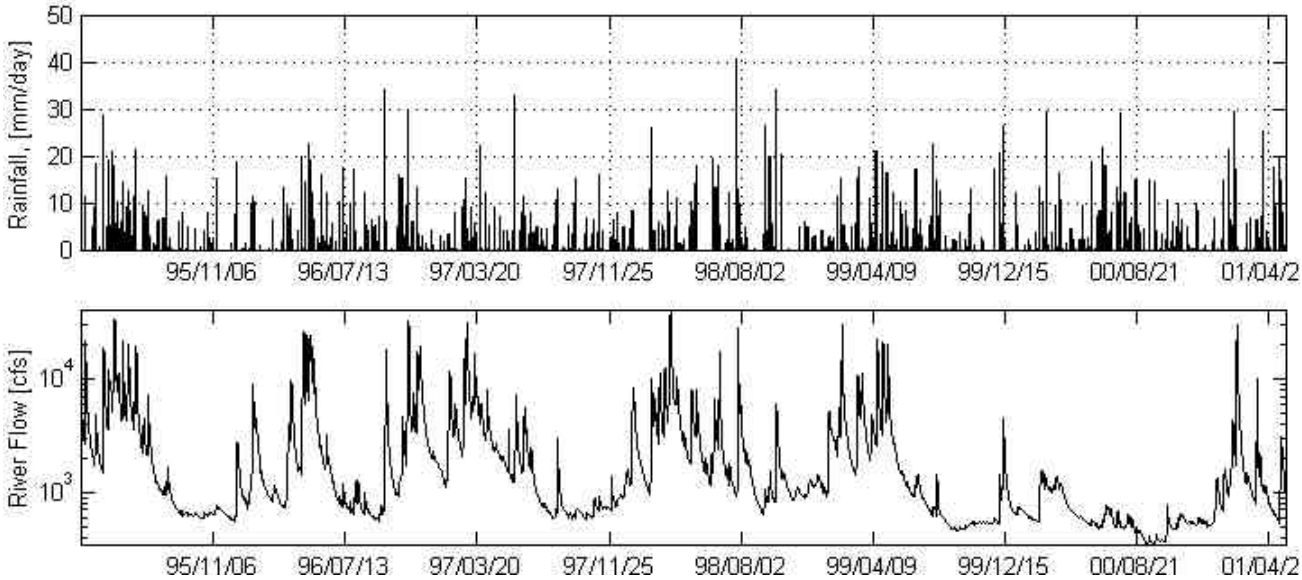


Station name	State	Station ID	Latitude	Longitude	Drainage Area km^2	Elevation m	HUC	Period of record	
								From	To
Chikaskia River near Blackwell	OK	7152000	36°48'41"	97°16'37"	4813.7	294.87	11060005	09/20/94	08/31/00
Gasconade River at Jerome	MO	6933500	37°55'47"	91°58'38"	7352.4	200.45	10290203	03/01/95	05/30/01

Fig. 3. Location and physical characteristics of the two basins (Chikaskia and Gasconade) studied. The scheme of subdivision in 3 subbasins and 9/10 subbasins (respectively for Chikaskia/Gasconade) is also shown.



(a)



(b)

Fig. 4. Display of the 6-hr mean precipitation over the catchment and the 6-hr average flow for: (a) station 07152000 (Chikaskia River near Blackwell) and (b) station 06933500 (Gasconade River at Jerome).

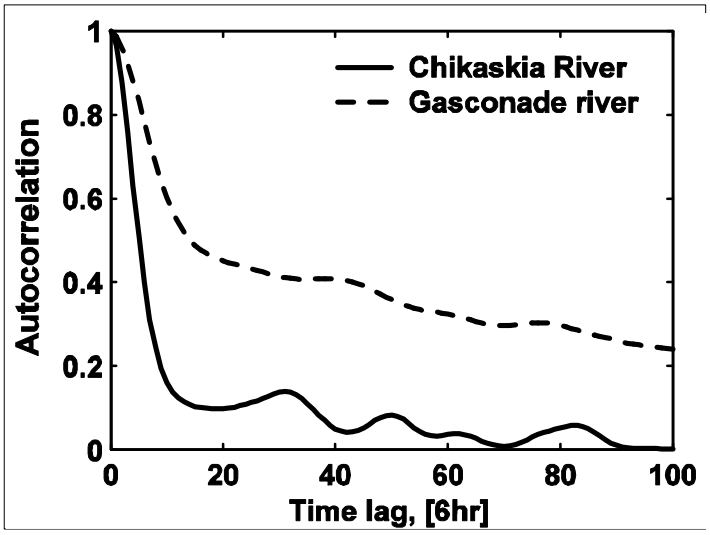
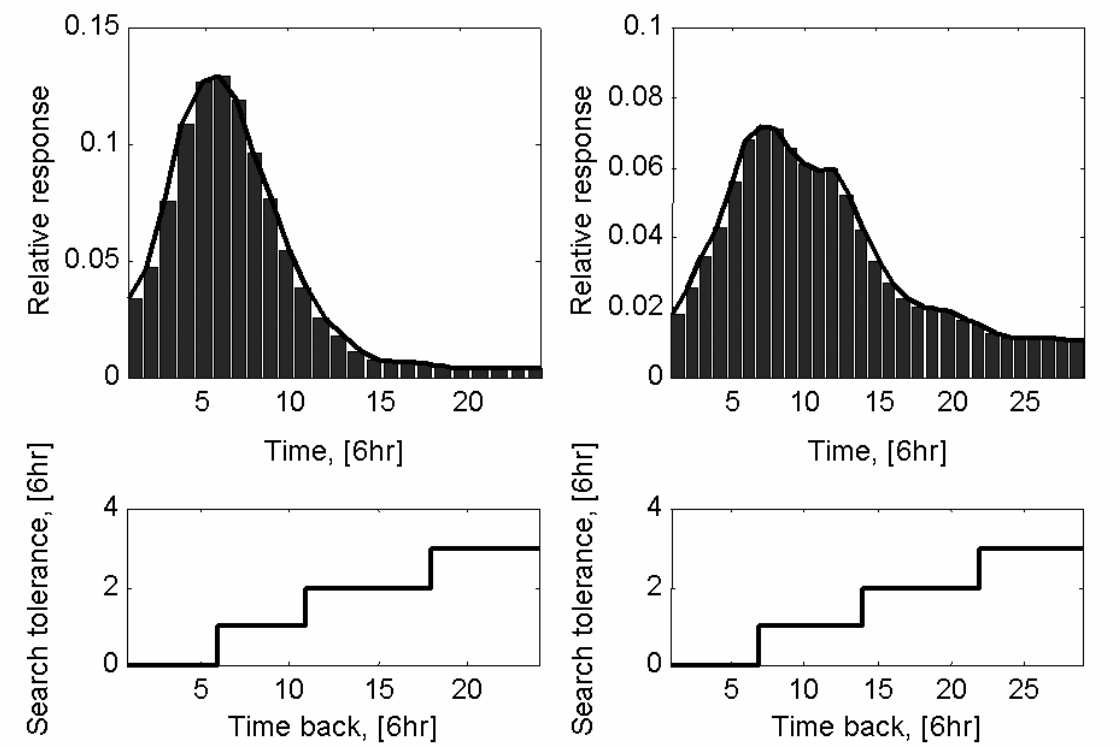
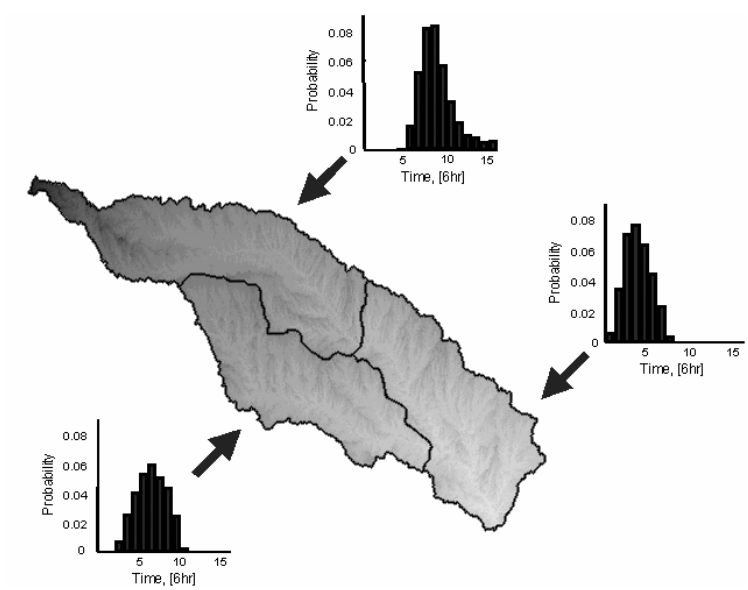


Fig. 5. Autocorrelation function of streamflow time series for the two rivers.

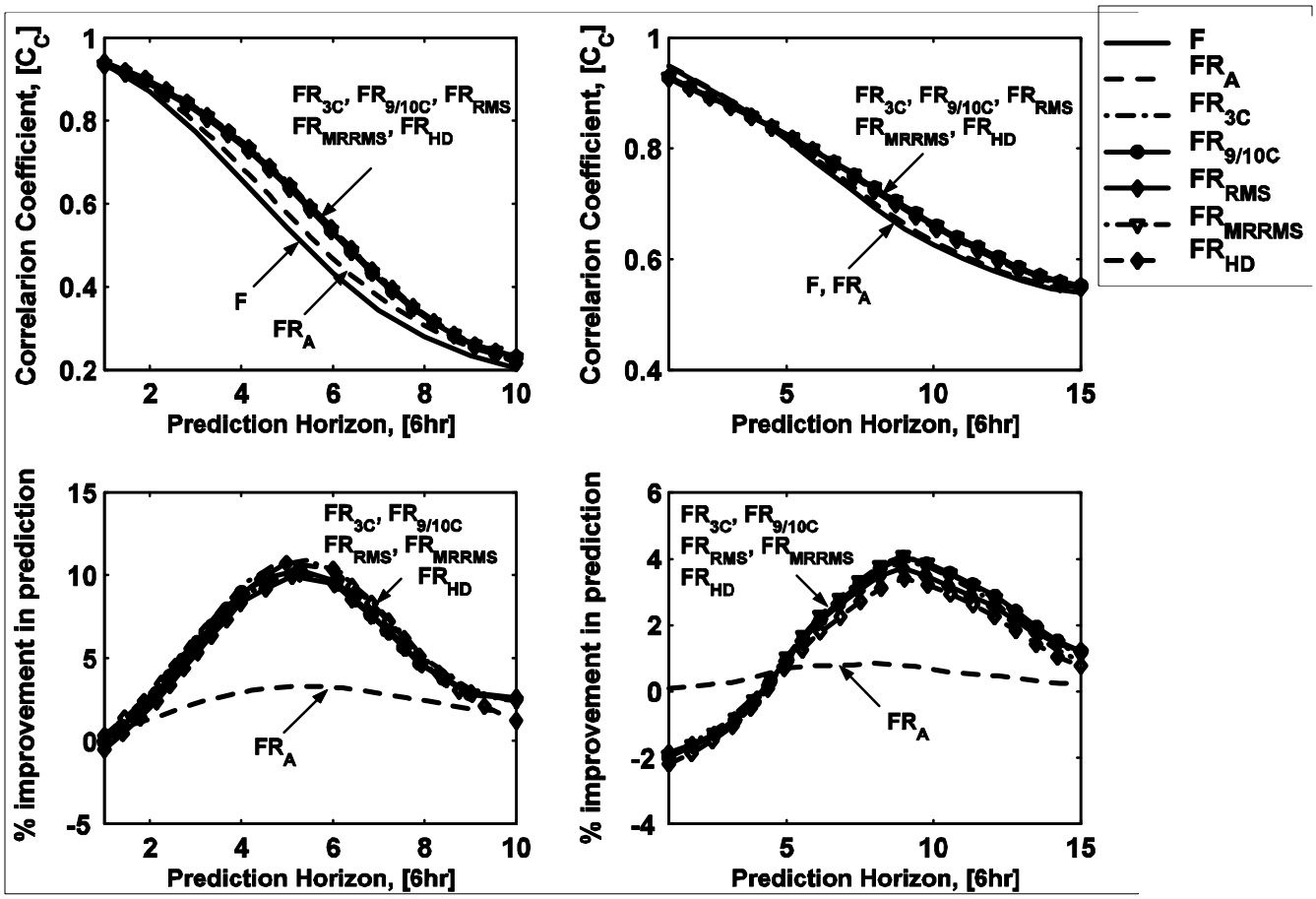


(a) Chikaskia River (b) Gasconade River



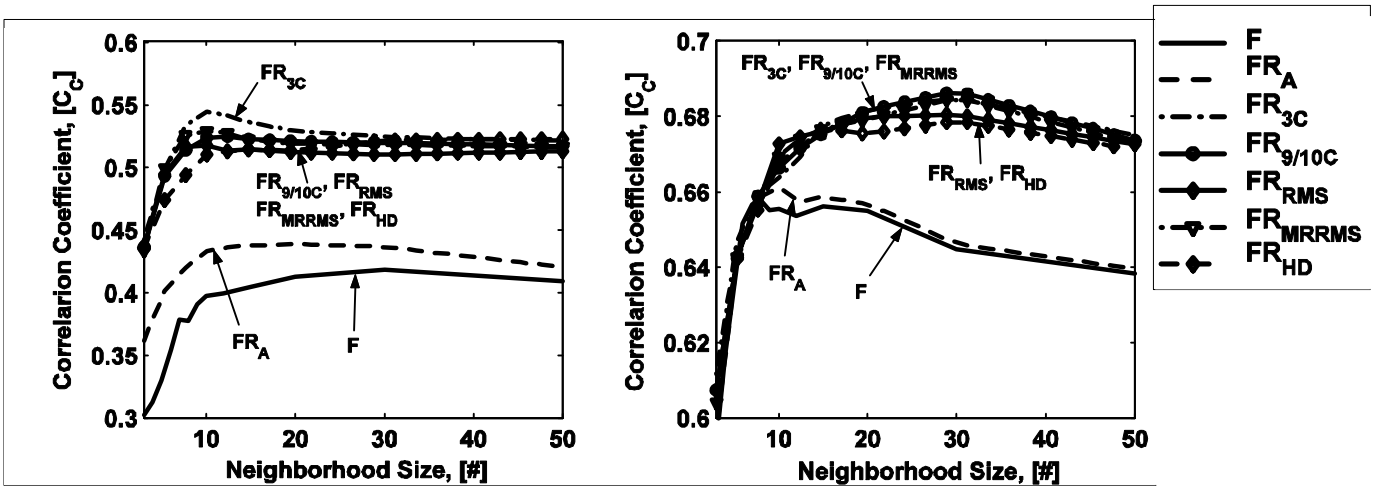
(c) Chikaskia River

Fig. 6. (a) and (b): Geomorphologically derived basin response functions (top) and temporal search tolerance thresholds ("uncertainty models") for each basin (bottom); (c): Geomorphologically derived response functions of the three subbasins of the Chikaskia River basin and gray-colored travel time to the outlet.



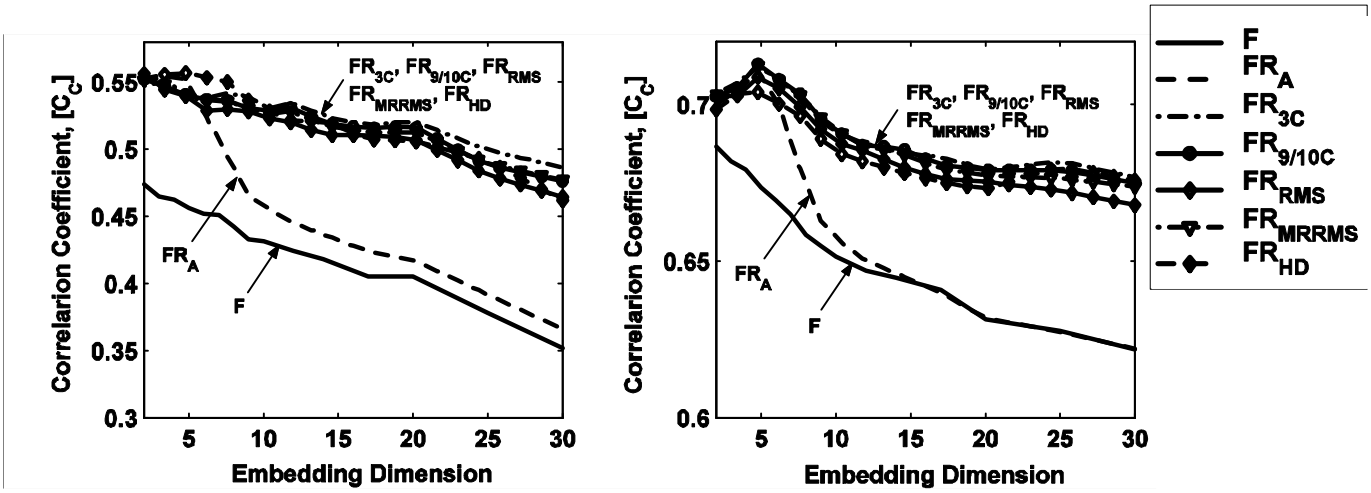
(a) Chikaskia River (b) Gasconade River

Fig. 7. Prediction accuracies, measured as correlation between the true and predicted series and the "improvement" in prediction (measured relatively to using the streamflow series only) based on the spatial information about rainfall for (a) station 07152000 (Chikaskia River near Blackwell) and (b) station 06933500 (Gasconade River at Jerome). It is clearly seen that incorporating the spatial distribution of rainfall instead of the basin average rainfall only, significantly improves streamflow predictability. The best improvement for the Chikaskia River is achieved at approximately 30 hours and for the Gasconade River at approximately 60 hours. See text for more explanation and Table 1 for definitions.



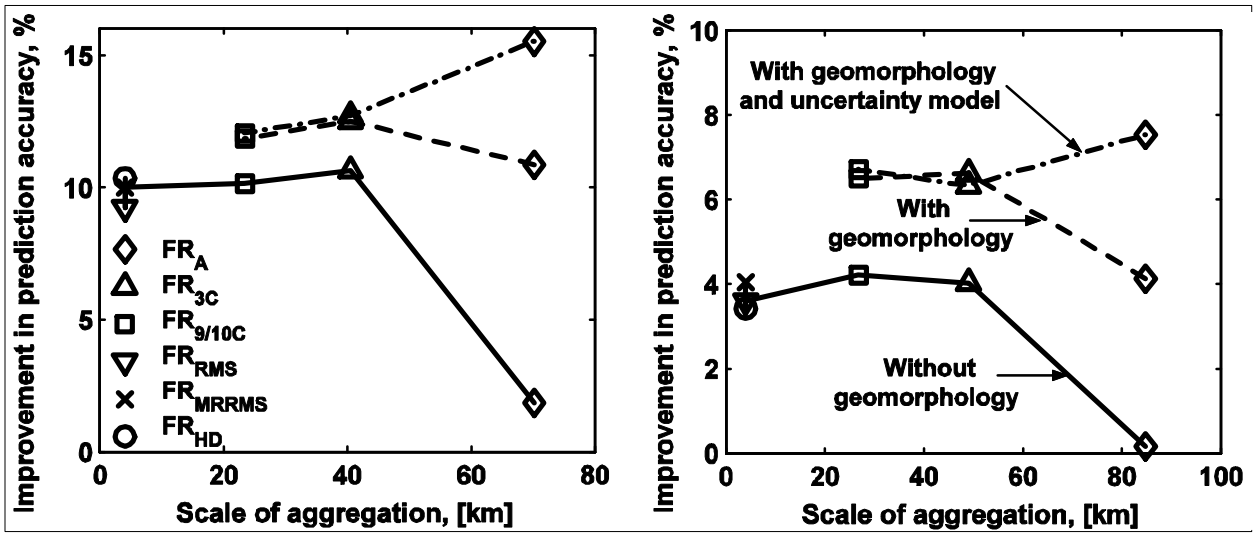
(a) Chikaskia River (b) Gasconade River

Fig. 8. Prediction accuracy as a function of neighborhood size. A neighborhood of size 30 was chosen for presenting the results of all analyses since for most cases this number provided best prediction accuracy. See Table 1 for definitions of symbols.



(a) Chikaskia River (b) Gasconade River

Fig. 9. Prediction accuracy for different distance measures as a function of the embedding dimension i.e., how far back in time the information about rainfall and streamflow is important for prediction of future streamflow. The prediction lead time is 6/9 time steps (for Chikaskia/Gasconade) and the neighborhood size is 30. See Table 1 for definitions of symbols.



(a) Chikaskia River (b) Gasconade River

Fig. 10. Improvement of predictability due to the spatial information about rainfall as a function of spatial rainfall aggregation scale (prediction horizon 6 and 9 time steps for Chikaskia and Gasconade Rivers respectively). Increased streamflow predictability is clearly seen when the catchment's morphology is incorporated and a tolerance threshold (uncertainty model) is introduced in assigning the contribution of a particular rainfall pulse to streamflow. See Table 1 for definition of symbols.