WATER RESOURCES RESEARCH, VOL. 49, 7048–7063, doi:10.1002/wrcr.20529, 2013

On hydrological model complexity, its geometrical interpretations and prediction uncertainty

Liselot Arkesteijn[1] and Saket Pande[1]

Received 9 January 2012; revised 11 September 2013; accepted 11 September 2013; published 28 October 2013.

[1] Knowledge of hydrological model complexity can aid selection of an optimal prediction model out of a set of available models. Optimal model selection is formalized as selection of the least complex model out of a subset of models that have lower empirical risk. This may be considered equivalent to minimizing an upper bound on prediction error, defined here as the mathematical expectation of empirical risk. In this paper, we derive an upper bound that is free from assumptions on data and underlying process distribution as well as on independence of model predictions over time. We demonstrate that hydrological model complexity, as defined in the presented theoretical framework, plays an important role in determining the upper bound. The model complexity also acts as a stabilizer to a hydrological model selection problem if it is deemed ill-posed. We provide an algorithm for computing complexity of any arbitrary hydrological model. We also demonstrate that hydrological model complexity has a geometric interpretation as the size of model output space. The presented theory is applied to quantify complexities of two hydrological model structures: SAC-SMA and SIXPAR. It detects that SAC-SMA is indeed more complex than SIXPAR. We also develop an algorithm to estimate the upper bound on prediction error, which is applied on five different rainfall-runoff model structures that vary in complexity. We show that a model selection problem is stabilized by regularizing it with model complexity. Complexity regularized model selection yields models that are robust in predicting future but yet unseen data.

Citation: Arkesteijn, L., and S. Pande (2013), On hydrological model complexity, its geometrical interpretations and prediction uncertainty, *Water Resour. Res.*, *49*, 7048–7063, doi:10.1002/wrcr.20529.

1. Introduction

[2] Hydrological models are conceptualizations and need to be assessed [*Gupta et al*., 1998] against observations of a variable of prediction interest. By prediction of a variable of interest, we mean the deterministic, simulated output of the model in response to the measured inputs alone (i.e., in the example of section 5, the precipitation and evapotranspiration). Models are estimated on a finite data sample which even when uncorrupted by measurement errors can lead to uncertainty about the model, out of the available candidates, that best approximates the underlying processes. A model estimated on a finite sample may significantly differ from a model estimated on a sufficiently large sample size due to sampling uncertainty. This leads to uncertainty in predicting future events. However, even at large sample sizes where prediction uncertainty due to sampling uncertainty vanishes, two entirely different model structures or conceptualizations may yield similar predictions. These issues are closely linked to the issue of ill-posed hydrological model selection problems. Issues of uniqueness and stability limit the possibility of well-posed model identification [*Gupta and Sorooshian*, 1983; *Vapnik*, 2002; *Renard et al*., 2010]. While the former is a result of model predictive equation specification leading to nonunique global optima, the latter is linked to a model's capacity to recreate data with little or no hydrological information or the model's complexity relative to the amount of available data [*Vapnik*, 2002; *Pande et al*., 2009, 2012].

[3] A hydrological model selection problem is ill-posed (in Hadamard's sense) if the optimal solution of the selection problem either does not exist, is not unique or is not stable. Here by optimal model we imply that the estimated model is closest in its predictions of a variable of interest to the observed in some notion of closeness. By stability, we here mean parametric stability and distinguish it from its use in dynamical systems. A solution is stable if small perturbations in the parameters of the (solution) model result in small perturbations in its predictions of a variable of interest. We here note that the parameters can also represent combinations of various subcomponents of a model, thus this definition of stability is applicable in a broader context of model structures. We posit a regularized hydrological model selection approach that restricts the set of solutions, where the regularization is with respect to the complexity

Additional supporting information may be found in the online version of this article.

[1]Department of Water Management, Delft University of Technology, Delft, Netherlands.

Corresponding author: L. Arkesteijn, Department of Water Management, Delft University of Technology, Stevinweg 1, NL-2628 CN Delft, Netherlands. (e.c.m.m.arkesteijn@student.tudelft.nl)

0043-1397/13/10.1002/wrcr.20529

of the problem, and can "correct" the ill-posedness of a model selection problem (in Tikhonov's sense) [*Vapnik*, 1982]. The regularization achieves this correction by restricting the set of solutions to a smaller set where stability is ensured. The use of regularization methods is one of several ways for solving ill-posed problems and hydrological model complexity is one possible basis for stabilization (to regularize). However, the role of hydrological model complexity, as it is in this paper, is closely tied to prediction uncertainty due to sampling uncertainty. The paper therefore studies the issues of ill-posed hydrological model selection problems and hydrological prediction uncertainty through its treatment of model complexity.

[4] We define the empirical risk ξ (also called empirical error, finite sample prediction error, or finite sample model performance) as the mean absolute difference between observed and model predictions of a variable of interest. For small sample sizes N, the empirical risk can significantly differ from its expected value, the expected empirical risk. The selection of a model that performs best in expected sense on future unseen data depends on the expected empirical risk (i.e., the expected risk in validation) rather than the empirical risk estimated on a single data realization of finite length. Therefore, we use expected empirical risk to assess the prediction error or model performance. Since an infinite number of realizations is needed to calculate a mathematical expectation, the expected empirical risk cannot be calculated directly and has to be approximated.

[5] We express the expected empirical risk in terms of the empirical risk and an upper bound on the deviation of the empirical risk from its expected value. The size of this deviation depends on the convergence rate of the empirical risk to the expected empirical risk. Model complexity influences this rate. An upper bound for the expected empirical risk can be given by a sum of empirical risk and a function of complexity and sample size [*Pande et al*., 2012].

[6] The related issues of prediction uncertainty, that we thus address, are associated with the predictability problem of second or third kind of *Kumar* [2011] since the deviation of the empirical risk from the expected risk can either be due to uncertain boundary conditions, inadequate model structure or changes in the error of the observations of the output being assessed against. Novel techniques for efficient parameter uncertainty estimation, data assimilation, numerical integration, and multimodel ensemble prediction have been introduced to better describe or tame hydrological prediction uncertainty [such as *Vrugt et al*., 2009; *Moradkhani et al*., 2005; *Kavetski and Clark*, 2010; *Parrish et al*., 2012]. Bayesian approaches to hydrological model selection, prediction uncertainty, model complexity, and regularization have also been well studied [*Schwarz*, 1978; *Jakeman and Hornberger*, 1993; *Young et al*., 1996; *Cavanaugh and Neath*, 1999; *Ye et al*., 2008; *Gelman et al*., 2008]. The use of prior distribution as a regularization term in a log-likelihood maximization is similar in form to the regularization proposed in this paper [*Gelman et al*., 2008]. *Ye et al*. [2008] compared AIC, BIC, and KIC measures and showed that an effective complexity measure (and thus regularization based on it) in KIC, being a finite (though asymptotically large) sample version of BIC [*Ye et al*., 2008], depends on the Hessian of the likelihood function at the optimum under certain regularity conditions [*Cavanaugh and Neath*, 1999; *Ye et al*., 2008]. Meanwhile in BIC it depends on model parameter dimensionality. The regularity conditions are used to replace the need for full specification (that the observations are generated by a member of the model space specified by a likelihood function). These conditions exploit the second-order Taylor series expansion of a log-likelihood function, certain assumptions on the prior and large sample size arguments to justify the use of KIC for model selection [*Cavanaugh and Neath*, 1999]. The use of KIC for model selection may however not be accurate for finite sample sizes. This is because it is a good approximation for posterior model probability (integral of the likelihood function over the parameter space) with an error of $O(N^{-1})$, where N is the sample size, when the likelihood function is normally distributed [*Slate*, 1994; *Tierney and Kadane*, 1986] or when the log-likelihood function is highly peaked near its maximum even for small N [*Kass and Raftery*, 1995]. Such conditions rarely hold on the likelihood functions when N is finite, in particular when it is small.

[7] *Jakeman and Hornberger* [1993] and *Young et al*. [1996] used complexity measures related to the information matrix. In particular, the seminal work of *Young et al*. [1996] on model complexity is quite different from the notion of complexity discussed in this paper. They identify a model with lower complexity than another model by identifying the "dominant modes" of the more complex model. The lower order model is identified on the basis of noise-free simulated data from the higher order more complex model. The identification is based on YIC measure that refers to the inverse of the instrumental product matrix and is related to the information matrix. The lower order model explains the output of the more complex model almost exactly and without ambiguity.

[8] The treatment of prediction uncertainty here excludes numerical inadequacies in computing the states of a system under consideration [*Kavetski and Clark*, 2010]. Further, the aim is not to discuss hydrological model structure improvements since we only analyze the convergence of the empirical risk of a hydrological model to its expected value (for a given hydrological variable of prediction interest) and its dependence on model complexity and available number of observations. This in turn is conditional on the set of candidate hydrological models or on a given model structure and elucidates the relationship between hydrological prediction uncertainty, data finiteness and model (structure) complexity [*Ye et al*., 2008; *Clement*, 2011; *Pande et al*., 2012].

[9] Here the role of model complexity relative to data availability in ill-posed hydrological problems (in Hadamard's sense) and in bounding the expected empirical risk is recognized. The ill-posedness in hydrological model selection problems appears due to the possibility of the many-to-one mapping from a set of hydrological processes to a response variable such as streamflow. The many-to-one mapping can yield solutions to hydrological model selection, which in turn is a selection of hydrological processes, that either are unstable, nonunique, or nonexistent. Solutions to model selection problems are deemed unstable when a small variation in the observed variable of interest, with respect to which the process of model selection

(of process conceptualizations) is being undertaken, results in large variation in the preferred sets of hydrological process conceptualizations.

[10] We emphasize that model complexity plays the role of a stabilizer to restrict the set of solutions of an ill-posed hydrological model selection problem to a subset of the original set that is compact. A compact set is a set that is bounded and closed. The restriction of the set of solutions to a compactum treats the issues of nonexistence of a solution. Thus, this restriction regularizes the model selection problem, correcting the ill-posedness by restricting the set of solutions to a subset where the problem has a solution that exists, is unique and stable for any set of observations such as streamflow or evaporation (or any other hydrological variable of interest) with respect to which the model selection problem is defined. The hydrological model selection is then well posed in Tikhonov's sense.

[11] The role of model complexity as a stabilizer has been undertaken in other type of problems such as density estimation problems [*Vapnik*, 1982]. However, a stabilizer does not have to be a measure of complexity; other choices for stabilization are available. But prediction uncertainty crucially depends on model complexity as defined in this paper. We make minimal assumptions on the data and the underlying distributions. These assumptions are explicitly stated. Nonetheless, it is the issue of obtaining unstable solutions which translates into finding a widely different process conceptualization as the number of observations for model selection increases, that unsettles a modeler the most. Selecting widely different process conceptualizations also implies different model complexities, affecting our confidence in its predictions. This is because the uncertainty in model prediction, in the sense of the probability in exceedance by an arbitrary positive number of the deviation of empirical risk from its expectation, is bounded from above by a function of model complexity and sample size. Widely different model complexities for similar sample sizes would imply different prediction uncertainties and hence a lack of confidence in model predictions.

[12] We identify an upper bound on the expected empirical risk for any hydrological model as a function of empirical risk, model complexity, and sample size. This upper bound for any given sample size serves to distinguish between models. This is akin to regularized hydrological model selection wherein a model with minimal complexity is selected from those which have lower empirical risk [*Pande et al.*, 2009]. Many concentration inequalities (inequalities that can bound the deviation of a random variable from its expected value) exist to estimate such bounds [*Boucheron et al.*, 2004], but most are applicable in hydrological model estimation only when model predictions are assumed to be independent between any two time steps. Since such model predictions are never independent between time steps, we use Markov's inequality that does not require independence in model predictions.

[13] Further, since the treatment of ill-posedness and prediction uncertainty crucially depend on the estimation of model complexity, we look at the computation of model complexity along with its geometric interpretation. We use mean absolute error as a measure of the empirical risk that can be interpreted as a measure of distance between the observed and predicted in a N-dimensional space, where N is the sample size. Here the sample size is the number of data points of a time series of a hydrological variable of interest such as streamflow or evaporation. Under a mild assumption, we show that the empirical risk depends on the distance of a prediction from its mathematical expectation, whose probability of exceedance is a function of model complexity and sample size. Using the same probability of exceedance, we show that model complexity, within the framework presented, is the expected absolute deviation of model prediction from the expectation of model prediction. This is not an assumption but it is a consequence of the theory presented in the paper. This geometrically describes model complexity as a summary statistic (expectation) of the size of model output space (measured by mean absolute deviation of predictions from expected values).

[14] The paper is organized as follows. Section 2 deals with prediction uncertainty, ill-posedness, and the role of model complexity. Section 3 then further discusses the notion of model complexity while section 4 provides a geometrical interpretation of model complexity. Section 5 then presents two algorithms to implement the theory presented and applies it on SAC-SMA and SIXPAR model structures. Section 6 presents a third algorithm to estimate an upper bound on expected empirical error. It is applied on five other nonlinear rainfall-runoff model structures using Guadalupe river basin data set (of daily streamflow, precipitation, and potential evapotranspiration) to determine models with optimal complexity on different sample sizes. It is also used to rank the model structures in terms of its (complexity regularized) suitability for the study area and compare it with the rankings provided by model selection without complexity regularization and BIC criterion. Finally, section 7 concludes.

2. Prediction Uncertainty and Ill-Posedness

[15] We define ξ as the absolute deviation of model prediction from the observed at time t, $\xi(t) = |y_0(t) - y(t)|$, where $y_0(t)$ is an observation of a hydrological variable of interest at time t, such as streamflow $Q(t)$, and $y(t)$ the model prediction. By prediction error, we mean the error that a model makes in predicting a variable of interest at some unobserved time t. It is assumed that its value is observed after the prediction has been made. Thus in case of streamflow, $\xi(t)$ measures the deviation of the predicted hydrograph $Q(t)$ from the observed hydrograph $Q_0(t)$. It follows that $|\xi(t) - E[\xi(t)]| = ||y_0(t) - y(t)| - E[|y_0(t) - y(t)|]|$ where E is an expectation operator, formally defined as $E[\xi(t)] = \int \xi(t) P(\xi(t)) \mathrm{d}\xi$. Here $P(.)$ is a probability distribution function. Since $\xi(t) = \xi(t; y_0, u)$, it then follows that $E[\xi(t)] = \iint \xi(t; y_0, u) P(y_0, u) \mathrm{d}y_0 \mathrm{d}u$ where u is a time series of input forcings. Similarly the expectation of the model output is defined as $E[y(t)] = \int y(t; u) P(u) \mathrm{d}u$. Since the distribution of u affects the expectation operator, we note that the expectation operator of $y(t)$ depends on $P(u)$. However, the sensitivity of the expectation operator to $P(u)$ has been suppressed in the remainder of the paper for notational convenience.

[16] We can obtain the expected value of $\xi(t)$ by $E[\xi(t)] = \lim_{M\to\infty} \sum_{j=1}^{M} \xi(t)_j/M$, where M is the number of realizations and $\xi(t)_j$, $t = 1, \ldots, N$ is the jth realization of a N-dimensional prediction vector. In section 5, we describe an algorithm to generate such a set of realizations. We assume that the absolute deviations $|\xi(t) - E[\xi(t)]|$ are of the order of the absolute deviation of model prediction from the expected prediction $|y(t) - E[y(t)]|$ at time t, i.e., **Assumption 1:** For some $\eta > 0$, let $|\xi(t) - E[\xi(t)]| \leq \eta|y(t) - E[y(t)]|$ for any admissible observed sequence of outputs $y_0(t)$. The interpretation of the assumption and η is discussed in a broader context, at the end of this section.

[17] By using the triangle inequality (that states that $|a + b| \leq |a| + |b|$ for any two real numbers a and b) and Assumption 1 we can bound the absolute deviation of empirical risk $\left(\sum_{t=1}^{N} \xi(t)/N\right)$ from expected empirical risk $\left(\sum_{t=1}^{N} E[\xi(t)]/N\right)$:

$$\frac{|\sum_{t=1}^{N}\xi(t) - \sum_{t=1}^{N}E[\xi(t)]|}{N} \leq \frac{\sum_{t=1}^{N}|\xi(t) - E[\xi(t)]|}{N} \leq \frac{\sum_{t=1}^{N}\eta|y(t) - E[y(t)]|}{N} \tag{1}$$

As shown later, this last term introduces a tradeoff between model complexity and sample size.

[18] For any $\gamma \geq 0$, let A and B be two events such that

$$\eta\gamma < \frac{|\sum_{t=1}^{N}\xi(t) - \sum_{t=1}^{N}E[\xi(t)]|}{N} \tag{A}$$

$$\eta\gamma < \frac{\sum_{t=1}^{N}\eta|y(t) - E[y(t)]|}{N} \tag{B}$$

[19] Since the right-hand side (RHS) of event A is less than or equal to the RHS of event B, it follows that event B is true whenever event A is true (or $A \Rightarrow B$). Thus $P(A) \leq P(B)$. Here $P(A)$ denotes the probability that event A is true. then

$$P\left(\frac{|\sum_{t=1}^{N}\xi(t) - \sum_{t=1}^{N}E[\xi(t)]|}{N} > \eta\gamma\right) \leq P\left(\frac{\sum_{t=1}^{N}|y(t) - E[y(t)]|}{N} > \gamma\right) \tag{2}$$

[20] Using the inequalities in (1), we have now devised an upper bound on the probability of exceedance for the absolute deviation of empirical risk from expected empirical risk in inequality (2) above. We here note that no assumptions have been made on the nature of the distribution from which ξ is being sampled. There exist upper bounds, other than the one presented later in the paper, for the LHS (left-hand side) of inequality (2), had ξ been independently distributed (i.e., if $P(\xi_t|\xi_{t'\neq t}) = P(\xi_t)$, ξ_t, and $\xi_{t'}$ are independently distributed). This is most often not the case for hydrological models, underlying the need for devising an upper bound on the LHS in inequality (2) that does not rely on the independence assumption.

[21] The RHS probability of inequality (2) is estimated by Markov's inequality [*Boucheron et al.*, 2004].
Lemma 1: (Markov's inequality). If X is an arbitrary positive random variable and $t > 0$, then

$$P(X \geq t) \leq \frac{E[X^2]}{t^2}$$

[22] By applying Markov's inequality on the RHS of inequality (2) we obtain inequality (3) below. The RHS can be split into two terms by expanding the quadratic term and using the linearity of the expectation operator. We obtain:

$$P\left(\sum_{t=1}^{N}|y(t) - E[y(t)]| > N\gamma\right) \leq \frac{E\left[\left(\sum_{t=1}^{N}|y(t) - E[y(t)]|\right)^2\right]}{N^2\gamma^2} \tag{3}$$

$$= \frac{1}{N^2\gamma^2}E\left[\sum_{t=1}^{N}|y(t) - E[y(t)]|^2\right] + 2\sum_{t=1}^{N}\sum_{t'=1}^{t-1}E[|y(t) - E[y(t)]||y(t') - E[y(t')]|] \tag{3a}$$

[23] From this equation, we note that the first term in (3a) contains the sum of variances of $y(t)$. The second term is a sum of $\frac{N(N-1)}{2}$ positive terms. Hence, the RHS of inequality (3) is always positive.

[24] Further we note that the numerator of the RHS, i.e., $E\left[\left(\sum_{t=1}^{N}|y(t) - E[y(t)]|\right)^2\right]$, is of $O(N^2)$ or less. From this we can conclude that the numerator can be bounded from above by a polynomial of N with a maximum order of 2. If we maximize $P\left(\sum_{t=1}^{N}|y(t) - E[y(t)]| > N\gamma\right)N^2\gamma^2$ with respect to γ for each N and denote the value of γ that corresponds to that maximum by $\gamma^N_{\max}$, the inequality in (3) holds with equality. A function to estimate the RHS can therefore be obtained by fitting a second-order polynomial of N to the maximum $P\left(\sum_{t=1}^{N}|y(t) - E[y(t)]| > N\gamma^N_{\max}\right)N^2{\gamma^N_{\max}}^2$.

[25] Let $h = \{\beta_0, \beta_1, \beta_2\}$ be a parameter set that defines the coefficients of the second-order polynomial $f(h, N)$ describing the RHS of inequality (3). Also, let $F(h, N) = f(h, N)/N^2$ and let γ be any nonnegative value. We can then rewrite inequality (3) to:

$$P\left(\sum_{t=1}^{N}|y(t) - E[y(t)]| > N\gamma\right) \leq \frac{f(h,N)}{N^2\gamma^2} = \frac{F(h,N)}{\gamma^2} \tag{4}$$

[26] We now note that by substituting this new upper bound into inequality (2) we obtain an upper bound for the probability:

$$P\left(\frac{|\sum_{t=1}^{N}\xi(t)-\sum_{t=1}^{N}E[\xi(t)]|}{N}>\eta\gamma\right)\leq\frac{F(h,N)}{\gamma^2} \tag{5}$$

[27] Then, if we denote ξ_N as the empirical risk on a sample set of size N $\left(\xi_N=\sum_{t=1}^{N}|y_0(t)-y(t)|/N\right)$ and equating $\chi=\frac{F(h,N)}{\gamma^2}\geq 0$, it holds with probability $(1-\chi)$ that $|\xi_N-E[\xi_N]|\leq\eta\gamma$. Substituting $\sqrt{\frac{F(h,N)}{\chi}}$ for γ gives:

$$|E[\xi_N]-\xi_N|\leq\eta\sqrt{\frac{F(h,N)}{\chi}} \tag{6}$$

[28] We now have an upper bound on the allowable range for the deviation of the empirical risk from the expected empirical risk. Model complexity is embedded in this inequality containing expected empirical risk and empirical risk. In presence of minimal data the upper bound (RHS of equation (5)) on the range is crucial. The problem is stable if the upper bound in the inequality is small for all N since the solutions such as selected process conceptualizations do not vary widely as N increases. Here "small" may be defined relative to measurement errors present in the data set. We also note that the RHS bounds the deviation of the empirical risk from its expected value and the capacity to have such larger deviations depends on the richness or complexity of the underlying model structure. Thus, two estimation problems can be ordered based on the respective magnitudes of the RHS for any N. Since for the same N and a fixed η, what distinguishes the RHS of the two problems is the parameter set h, which identifies model complexity or complexity of model estimation.

[29] We note that the inequality (6) provides an upper bound on the expected empirical risk:

$$E[\xi_N]\leq\xi_N+\eta\sqrt{\frac{F(h,N)}{\chi}} \tag{7}$$

[30] Minimizing the upper bound in the RHS of inequality (7) yields a model with smaller expected empirical risk than most of the other potential models. Hence it is preferred for simulating the unknown future. Thus, a trade-off between empirical risk and a measure of complexity, as in RHS of (7), bounds the prediction uncertainty of a preferred model. It also demonstrates the role that model complexity plays in bounding prediction uncertainty in addition to its role of inducing stability.

[31] We note that $F(h,N)$ in the RHS also acts as stabilizer to a potential ill-posed hydrological model selection problem where $F(h,N)$ is a continuous mapping from a model space (of potential hydrological process conceptualizations) to a positive real line ($F(h,N)$ is nonnegative for any model by definition, see inequality (4)). The minimization of the RHS of (7) is a Lagrangian equivalent of minimizing the empirical risk subject to a constraint on $F(h,N)$ of type $\sqrt{F(h,N)}\leq c$ where c is some positive constant. A Lagrangian formulation represents a constrained minimization (or a maximization) problem as an unconstrained problem (the Lagrangian), where the constraints enter the objective function in penalized form. The penalty is defined by Lagrange multipliers that in turn quantify how binding the constraints are to the problem. The constrained problem, whose Lagrangian is equivalent to the RHS of inequality (7), is to minimize the empirical risk with respect to the model parameters and h and subject to $\sqrt{F(h,N)}\leq c$. Such a constrained minimization ensures that two selected models with close empirical risk are not arbitrarily different (in terms of parameterization, including process conceptualizations). Uniqueness and existence of a solution to a hydrological model selection problem can be ensured by a certain choice of η such that the constrained model selection is restricted to a certain subset of the original hypothesis space as long as it can be ensured that the global minimizer lies in this subset. Thus, the RHS of inequality (7) poses any ill-posed hydrological selection problem as a well-posed one [*Vapnik*, 1982, pp. 23 and 308].

[32] Thus, the role of η is to control the degree to which a hydrological model selection problem is regularized. However, it is a consequence of Assumption 1. Regarding the latter, $\eta>0$ can be shown to exist for any hydrological model selection problem based on the triangle inequality such that Assumption 1 holds. Therefore, Assumption 1 can be stated as a proposition under minimal assumption (boundedness of hydrological model prediction in the variables of interest). However inequality (1), which is a direct consequence of Assumption 1, is not tight due to the minimalist nature of Assumption 1. Consequently inequalities (6) and (7) are weak (though definition and computation of model complexity based on inequality (4) remains unharmed).

[33] Finally, if we let η be some function of N such that $\eta\rightarrow 0$ as $N\rightarrow\infty$, the convergence of $|E[\xi_N]-\xi_N|\rightarrow 0$ as $N\rightarrow\infty$ is ensured. Thus Assumption 1 does not appear to be a strict assumption if η as an appropriate function of N can be found such that the preferred model that minimizes the empirical risk is stable for all N and limits to the model that minimizes the expected empirical risk. We relegate its more formal treatment in hydrology to future research.

3. Model Complexity

[34] In the previous section, we suggested that the function that bounds $P\left(\sum_{t=1}^{N}|y(t)-E[y(t)]|>N\gamma\right)N^2\gamma^2$ (inequality (4)) is a second-order polynomial of data size N, depending on complexity h. Let $f(h,N)=\beta_2N^2+\beta_1N+\beta_0$, where $h=\{\beta_0,\beta_1,\beta_2\}$. We now formulate an answer to the question as to why this function indeed depends on complexity. First, we show why $f(h,N)$ informs us about the rate of convergence of P_N (for brevity reasons, we define $P_N=P\left(\sum_{t=1}^{N}|y(t)-E[y(t)]|>N\gamma\right)$), i.e., how P_N converges to an asymptote with increasing N. In the next section, we show why h is a measure of complexity, by using its geometric interpretation of a statistic measuring the size of model output space.

[35] We start by taking a closer look at inequality (4). A smaller value of $f(h,N)$, for a given value of N, implies a tighter upper bound and hence allows smaller values for P_N. For increasing N, P_N will reach a certain asymptote,

and the rate at which this takes place is no larger than the change of $f(h,N)$ for increasing N. Further, the rate at which P_N reaches an asymptote for a particular model, the convergence rate, depends on complexity of a model [*Vapnik*, 1982] and allows an intercomparison between any two models. We note that a more complex model intuitively requires more observations to have credible predictions than a less complex model. This translates to a notion that probability with which empirical risk of a more complex model deviates from its expected value by a certain threshold (called the probability of error) is higher than that of a less complex model for a given number of observations (sample size). The rate at which the probability of error approaches an asymptote (here to 0), i.e., the rate of convergence, is therefore faster for a less complex model. We note that the rate of convergence is defined on the left-hand side (LHS) of inequality (5). However, if the RHS of inequality (5) meaningfully controls the rate of convergence, it should depend on model complexity. Based on our construct, we note that any measure of model complexity appearing in the RHS of (5) should also bound the rate of convergence of model predictions to its expected value as shown in inequality (4). If then, $f(h,N)$ is represented by a polynomial of maximum order 2, the values of the coefficients of the polynomial between two models of sufficiently different complexity should be different. Hence h (the coefficients of the polynomial) is a measure of complexity.

[36] The parameter set h determines the rate of increase between two maximum values of $P_N N^2\gamma^2$ for two subsequent N, i.e., it measures

$$\max_{\gamma}|P_{N+1}(N+1)^2\gamma^2| - \max_{\gamma}|P_N N^2\gamma^2| \tag{8}$$

We note that the rate of convergence of P_N is the rate at which $|P_{N+1}-P_N| \to 0$ with increasing N and for any positive γ. This rate of convergence is also embodied in the behavior of $\max_{\gamma}|P_{N+1}-P_N|\gamma^2$ with N. Further, if $N^2 \max_{\gamma}|P_{N+1}-P_N|\gamma^2$ diverges faster for one model compared to the other, the more divergent model is more complex. This is because a faster rate of divergence of the above quantity implies a slower rate of convergence of $\max_{\gamma}|P_{N+1}-P_N|\gamma^2$ (since the N^2 term in $N^2 \max_{\gamma}|P_{N+1}-P_N|\gamma^2$ contributes to its divergence and this contribution is the same for any model, given that $P_{N+1}-P_N$ converges to zero for any γ and for any model). This in turn embodies the rate at which $|P_{N+1}-P_N| \to 0$ with increasing N and for any positive γ.

[37] An equivalence between (8) and $\max_{\gamma}(P_{N+1}-P_N)\gamma^2$ is now shown in the following (in equations (9) and (10)) for large N. We note that for $N >> 1$, the following holds,

$$N^2 \max_{\gamma}|P_{N+1}-P_N|\gamma^2 \approx \max_{\gamma}|P_{N+1}(N+1)^2 - P_N N^2|\gamma^2 \tag{9}$$

[38] This approximate equality is interpreted and shown to hold for a simple example. We consider a mapping $y: \mathbb{R}^d$ x $\mathbb{R}^d \to \mathbb{R}$, where $y(\mathbf{c},\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ if $\mathbf{c}$ and $\mathbf{x}$ are defined as column vectors. Further, let $\mathbf{c}$ be a vector with constant components and $\mathbf{x}$ be a vector with i.i.d. (independently and identically distributed) stochastic components with 0 mean and variance 1. We note that such a mapping represents a class of linear functions on $\mathbf{x}$ with parameters $\mathbf{c} \in \mathbb{R}^d$.

[39] Then, since $\mathbf{x}$ has zero mean and using the linearity of the expectation operator (defined on the distribution of $\mathbf{x}$):

$$\begin{aligned}\mathrm{Var}(y) &= E[y-E[y]]^2 = E[\mathbf{c}^T\mathbf{x} - E[\mathbf{c}^T\mathbf{x}]]^2 \\ &= E[\mathbf{c}^T\mathbf{x}]^2 = E\left[\sum_{i=1}^{d} c_i x_i\right]^2 = E\left[\left(\sum_{i=1}^{d} c_i x_i\right)^2\right] \\ &= E\left[\sum_{i=1}^{d} c_i^2 x_i^2 + 2\sum_{i=1}^{d}\sum_{j=1}^{i-1} c_i x_i c_j x_j\right] = \sum_{i=1}^{d} c_i^2 E[x_i]^2\end{aligned}$$

Since $\mathrm{Var}(x_i) = E[x_i - E[x_i]]^2 = E[x_i]^2 = 1$, it follows that:

$$\mathrm{Var}(y) = \sum_{i=1}^{d} c_i^2 E[x_i]^2 = \sum_{i=1}^{d} c_i^2 = ||c||$$

We now use inequality (3) to define an upper bound on the probability P_N. Substituting the above gives:

$$P_N \le \frac{N\mathrm{Var}(y)}{N^2\gamma^2} = \frac{||c||}{N\gamma^2}$$

Applying this to equation (9), we get for the LHS:

$$N^2 \max_{\gamma}|P_{N+1}-P_N|\gamma^2 = N^2\left|\frac{||c||^2}{N+1} - \frac{||c||^2}{N}\right| = N^2\left|\frac{-||c||}{N(N+1)}\right| \approx ||c||$$

For the RHS of (9) we note that all variables in the equation are positive and therefore the absolute value operator may be removed:

$$\begin{aligned}&\max_{\gamma}|P_{N+1}(N+1)^2\gamma^2| - \max_{\gamma}|P_N N^2\gamma^2| \\ &= \max_{\gamma}|||c||(N+1)| - \max_{\gamma}|||c||\mathrm{N}| = ||c||(N+1-N) = ||c||\end{aligned}$$

[40] Hence LHS $\approx$ RHS. The example allows an interpretation of equation (9), that either side of the equality estimates the norm of the parameters of the class of linear functions (or more generally the norm of the constants of the defined mapping). The norm of the parameters of linear regressors is used as stabilizers, one example being of ridge regression to correct ill-posedness issues such as the presence of multicollinearity in linear regression problems [*Marquardt and Snee*, 1975]. Further, we note that the RHS of equation (9) is the quantity defined in (8) that is expected to measure model complexity. Indeed, the norm of the parameters of linear regressors is often used as a measure of complexity that affects prediction uncertainty (see, e.g., Theorem 5.1 of *Vapnik* [2002]).

[41] Finally, we show that the expression in (8) (that is measured by h) is related to the rate of convergence embodied in the LHS of (9). We note that $\max|a+b| \le \max|a| + \max|b|$ where a and b are two arbitrary variables. ($\forall a,b$ it holds that $|a| \le \max|a|$ and $|b| \le \max|b|$. Hence $|a+b| \le \max|a| + \max|b|$, since $|a+b| \le |a|+|b| \le \max|a| + \max|b| \forall a,b$, but then also $\max|a+b| \le \max|a| + \max|b|$.)

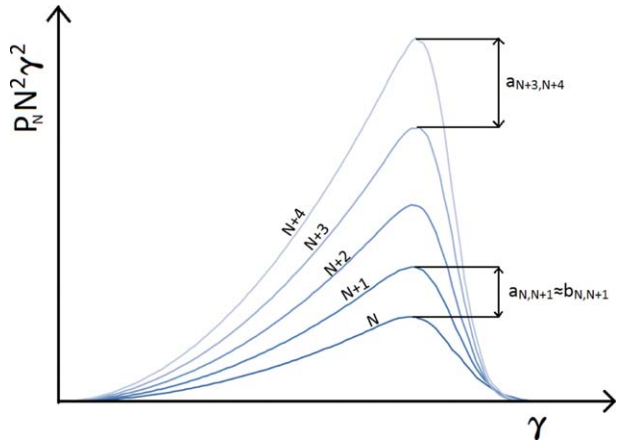

Figure 1. Multiple curves $P_N N^2\gamma^2$ are drawn for subsequent N. The distance between the maxima with respect to γ of two subsequent curves is denoted by $a_{j,j+1}$ (LHS of (10)), where j is the sample size. For larger N this distance increases due to the second-order polynomial that fits these maxima. The RHS of (10) is indicated by $b_{j,j+1}$ and is approximately equal to $a_{j,j+1}$.

By substituting $a+b$ by a, it can be shown that $\max|a|-\max|b|\leq\max|a-b|$. It then follows that:

$$\begin{aligned}&\max_{\gamma}|P_{N+1}(N+1)^2\gamma^2| - \max_{\gamma}|P_N N^2\gamma^2| \\ &\leq \max_{\gamma}|P_{N+1}(N+1)^2 - P_N N^2|\gamma^2\end{aligned} \tag{10}$$

[42] We note that the inequality (10) holds with equality for the example of linear mappings $y(\mathbf{c},\mathbf{x})$ with $\text{LHS}=\text{RHS}=||\mathbf{c}||$. In Figure 1, multiple curves $P_N N^2\gamma^2$ are drawn for subsequent N. The model used to generate these curves is a conceptual hydrological model. More details on these calculations can be found in section 5. The maxima with respect to γ of these curves are used for the fitting of $f(h,N)$. As one can see, the distance between the maxima of two subsequent curves (LHS of (10)), denoted by "a," increases for increasing N. The RHS of (10) is indicated by "b" and should never be smaller than a. In this figure the maximum γ's for the different curves are very close to each other and therefore note that $a\approx b$.

[43] However, we note that $a\approx b$ holds for any model when N is large. Since then, $P_N\rightarrow P_{N+1}$ and thus $\gamma^N_{\max}$ that maximizes $P_N N^2\gamma^2$ converges to $\gamma^{N+1}_{\max}$ that maximizes $P_{N+1}(N+1)^2\gamma^2$. This is because for large N, P_N is no longer a function of N and therefore the $\gamma^N_{\max}$ that maximizes $P_N N^2\gamma^2$ is independent of N. Thus for large N it follows that:

$$\begin{aligned}&P_{N+1}(N+1)^2\left(\gamma^{N+1}_{\max}\right)^2 - P_N N^2\left(\gamma^N_{\max}\right)^2 \\ &\approx \left(P_{N+1}(N+1)^2 - P_N N^2\right)\left(\gamma^N_{\max}\right)^2\end{aligned}$$

or,

$$\begin{aligned}&\max_{\gamma}\left(P_{N+1}(N+1)^2\gamma^2\right) - \max_{\gamma}\left(P_N N^2\gamma^2\right) \\ &\approx \max_{\gamma}\left(P_{N+1}(N+1)^2 - P_N N^2\right)\gamma^2\end{aligned}$$

or, $a\approx b$.

[44] Here we note that a model with larger complexity will have a value of h such that the curve $f(h,N)$ will be pointwise greater than that of a less complex model. Thus, the LHS of inequality (10) will be larger, which may imply a larger RHS in inequality (10) at least for a significantly different LHS. Finally, we note that if the LHS is significantly different for two models, then the RHS will be also significantly different. From approximation (9), the larger the RHS is, the higher is the model's complexity. Meanwhile the LHS is the derivate of $f(h,N)$ with respect to N and depends on h. Thus significant differences in h measure differences in complexity.

4. Geometric Interpretation

[45] A geometric interpretation exists for the function $F(h,N)$ in inequality (4). We note that the expected value of model output is a centroid of model output space (populated by model output points with certain probability) while a model output point itself can be anywhere in model output space. Both are points in a N-dimensional space where the model output space defines a region wherein a model prediction point may lie. The probability that the distance between those values exceeds a threshold is larger when the size of model output space is larger. In this case an average of such a distance for a finite sample of size N will also be larger. If γ represents the threshold and $\left(\sum_{t=1}^{N}|y(t)-E[y(t)]|\right)/N$ represents the distance between two N-dimensional vectors $\mathbf{y_N}(t)=(y(1), y(2),\ldots,y(N))$ and $E[\mathbf{y_N}(t)]=(E[y(1)],\ E[y(2)],\ \ldots,\ E[y(N)])$, this should imply that the probability on the LHS of inequality (4) is larger if the size of model output space is larger. For a sufficiently tight upper bound in (4), this leads to a larger RHS for any N and γ.

[46] Two model output spaces are exemplified in Figure 2. Both output spaces have the same shape but the sizes differ. For both models the probability $P_N=\left(\sum_{t=1}^{N}|y(t)-E[y(t)]|/N>\gamma\right)$ in inequality (4) can be calculated by dividing the number of points $\mathbf{y_N}(t)$ outside the circle by the total number of realizations of $\mathbf{y_N}(t)$. Since model output space 1 is significantly larger than model output space 2, more realizations of $\mathbf{y_N}(t)$ will lie outside the circle and thus the probability is larger for model 1. In case of a sufficiently tight upper bound, the function of complexity, $F(h,N)$ should also be larger. By sufficiently tight we mean that while comparing two models, a smaller LHS implies a smaller RHS.

[47] In the previous section, we defined $f(h,N)$ such that for any N there exists $\gamma^N_{\max}$ such that inequality (4) holds with equality. Also, we note that it holds that $P\left(\lim_{N\rightarrow\infty}\frac{\sum_{t=1}^{N}|y(t)-E[y(t)]|}{N}>\gamma^N_{\max}\right)=\lim_{N\rightarrow\infty}P\left(\frac{\sum_{t=1}^{N}|y(t)-E[y(t)]|}{N}>\gamma^N_{\max}\right)$. Details on this can be found in the supporting information. The following then holds for large N:

$$\begin{aligned}P\left(\lim_{N\rightarrow\infty}\frac{\sum_{t=1}^{N}|y(t)-E[y(t)]|}{N}>\gamma^N_{\max}\right)&=\lim_{N\rightarrow\infty}\frac{1}{{\gamma^N_{\max}}^2}F(h,N)\\ &=\frac{\beta_2}{{\gamma^N_{\max}}^2}\end{aligned} \tag{11}$$

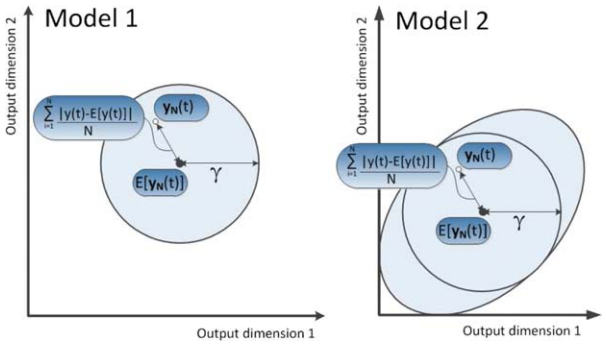

Figure 2. Determination of model complexity by measuring the size of model output space. In two model output spaces of different size a circle with radius γ is drawn. The vector $E[\mathbf{y_N}(t)]$ and one instantiation of $\mathbf{y_N}(t)$ are indicated with points. For a larger output space, the probability of points $\mathbf{y_N}(t)$ lying outside this circle is larger which implies the model's complexity is larger.

We here note that this last fraction is always less or equal to 1 because of the way β_2 is constructed. Further, we define β_2 as the asymptotic complexity. The LHS of the inequality inside the probability can be rewritten as an expected value, $E[|y(t) - E[y(t)]|]$, a constant. Denoting $\gamma^* = E[|y(t) - E[y(t)]|]$, we have from (11):

$$P(E[|y(t) - E[y(t)]|] > \gamma) = \begin{cases} 1 & \text{for} \quad \gamma \leq \gamma^* \\ 0 & \text{for} \quad \gamma > \gamma^* \end{cases} \tag{12}$$

[48] We note that γ^* maximizes $P_N N^2 \gamma^2$ as $N \to \infty$, since $P_N = 1$ for γ^* and $P_N = 0$ for any $\gamma > \gamma*$. Thus we have, $\gamma^* = \gamma^N_{\max}$ as $N \to \infty$. Finally using (11) and (12), we have $\frac{\beta_2}{\gamma^{*2}} = 1$ and thus

$$\beta_2 = \gamma^{*2} = E[|y(t) - E[y(t)]|]^2 \tag{13}$$

[49] From (13) we can make the following conclusion: if the model output space is large, we expect the absolute deviation of a model prediction point from its expected value to be large as well (RHS of (13)). Thus β_2, the asymptotic complexity, is large if the size of model output space is large.

Table 1. Parameter Ranges for SAC-SMA Model

Parameter	Range
UZTWM (mm)	1–150
UZK (day^{-1})	0.1–0.5
ADIMP	0–0.4
ZPERC	1–250
LZTWM (mm)	1–1000
LZFPM (mm)	1–1000
LZPK (day^{-1})	0.0001–0.025
RSERV	0.3
UZWFM (mm)	1–150
PCTIM	0–0.1
RIVA	0
REXP	1–5
LZFSM (mm)	1–1000
LZSK (day^{-1})	0.01–0.25
PFREE	0.0–0.6
SIDE	0.0

Table 2. Parameter Ranges for SIXPAR Model

Parameter	Range
UM (mm)	0–300
BM (mm)	0–3000
z	0–1
UK (day^{-1})	0–0.5
BK (day^{-1})	0–0.0796
x	0–10

5. Quantification of Model Structure Complexity: A Comparison of Complexities of SAC-SMA and SIXPAR Model Structures

[50] We now explicitly present the algorithm to quantify model complexity based on the theory presented and apply it on two hydrological model structures, SAC-SMA and SIXPAR at daily time steps. These model structures have been extensively studied in the literature with the latter model structure used as a simplification of the former [*Burnash*, 1995; *Duan et al*., 1992]. In the supporting information, short descriptions of both model structures are given. Tables 1 and 2 display the parameter ranges used in this study for SAC-SMA and SIXPAR, respectively.

[51] The objective of this application is to show that the theory distinguishes between the complexity of the two model structures when they have equivalent parameter ranges (with similar upper and lower zone capacities and similar corresponding recession parameters). In order to compute the complexity, the probability of exceedance in (4) has to be estimated. Therefore, M realizations of samples of size N are needed, with N ranging from low to "sufficiently" high values. For this application, we choose $M = 2000$ (number of realizations) and let the maximum value of N be 5000 ($=N_{\max}$). Smaller values of N are then obtained by subsampling data sequences of smaller sizes. Thus, a total 2000 sequences of 5000 data points for daily precipitation and evapotranspiration are sampled at once.

[52] In order to randomly sample data sets that are realistic (in hydrologic sense), a simple weather "resampler" is constructed and used. The weather resampler is such that it can at least preserve a basin specific correlation structure between evapotranspiration and precipitation. For the application presented here, we use over 30 years of daily precipitation and potential evapotranspiration data from Guadalupe river basin in the United States [*Duan et al*., 2006] from which the weather resampler generates the required matrix of data sequences.

[53] The weather "resampler" is described in the following algorithm.

Algorithm 1. (A simple weather resampler):

1. Obtain daily precipitation and potential evapotranspiration data for a basin.

2. Identify wet (a set of contiguous days with positive precipitation) and dry (a set of contiguous days with zero precipitation) spell pairs for each month: determine the

amount and length of spell pairs and attach an identifier to each spell.

3. Construct a 1 month sample for each month: conditioned on a selected month, randomly sample (with replacement) spell pairs, along with evapotranspiration values for the same days, across different years for the same month, appending these wet-dry spells till the total length of the sequence exceeds 30 days.

4. Repeat step 3 for all 12 months of a year.

5. Permute the months (if correlation between months is to be removed), while maintaining the order of sequences within each month, to create one year sample.

6. Repeat steps 4 and 5 and create one realization data sequence at daily time steps with $N_{\max}$ data points.

7. Repeat step 6 to create M realizations of $N_{\max}$ data points.

[54] Using the weather resampler, we obtain M sequences of $N_{\max}$ data points for daily precipitation and potential evapotranspiration. For each realization, data sequences of smaller sample sizes $N = 200 : 50 : N_{\max}$ are obtained by sampling its first N data points.

[55] We here note that the performance of the weather generator in replicating the statistical properties of the original time series crucially depends on the preservation of the wet/dry spell characteristics [*Lall et al*., 1996; *Mehrotra et al*., 2012; *Lee and Ouarda*, 2012]. We note that it is a multivariate uniform kernel resampler conditioned on a month that assumes independence of one wet/dry spell pair from another. This assumption can be restrictive but it can be relaxed by introducing resampling weights based on proximity in time or uniformly resampling blocks of wet/dry spells, each containing more than one wet/dry spell pair. See, for example, *Yu* [1994], *Meir* [2000], and *Kundzewicz and Robson* [2004] on the statistical properties of the class of weakly "mixing" processes, which are the processes for which the future depends only weakly on the past (such as ARMA process that is an exponential mixing process) and for the justification of using block resampling along the same lines as Algorithm 1. Thus, the simple weather resampler as detailed in Algorithm 1 can be improved by extending the definition of a block to contain more than one wet/dry spell pair. A study of the sensitivity of complexity quantification to a weather resampler is left for future research.

[56] Further, we note that our weather resampler is just one out of many possible algorithms to generate realistic time series of input forcings. The characteristics of the algorithm attempts to replicate $P(u)$ of a particular basin in the definition of the expectation operators defined in section 2 and improving this algorithm will improve the precision of the analysis.

[57] In order to evaluate the LHS of inequality (4) for a model structure either of SAC-SMA or SIXPAR, we need to sample its parameter sets from feasible ranges. Since the choice for a particular parameter set influences the empirical risk and its expected value, multiple parameter sets for both models are sampled. Table 1 shows the ranges that are used for the parameters of SAC-SMA. The ranges of SIXPAR model are adapted (shown in Table 2) to get equivalent ranges, e.g., the total lower/upper zone storage capacity of SAC-SMA is the same as the upper and lower zone storage capacity of SIXPAR and the geometric means of the upper and lower zone recession coefficients are the same as the upper and lower zone recession coefficients of SIXPAR. This is done so that the effect of magnitude of parameters on model complexity can be removed before comparing model complexities of SAC-SMA and SIXPAR [*Pande et al*., 2012]. Five hundred different parameter sets are then sampled from the respective ranges using hypercube sampling.

[58] Finally, Algorithm 2 presented below is applied on SAC-SMA and SIXPAR using the data generated by Algorithm 1 to estimate the respective model complexities over 500 parameter sets based on inequality (4).

Algorithm 2. (Quantification of model complexity):

1. For each parameter set of a model, estimate the left-hand side (LHS) probability in inequality (4), for different values of N and γ using M samples of data set of size N, resampled using Algorithm 1.

2. Find $\tilde{f}(N)$, a maximum of $PN^2\gamma^2$ with respect to γ for each N. Let the γ that maximizes $PN^2\gamma^2$ be $\gamma^N_{\max}$.

3. Repeat steps 1 and 2 for $N = 200 : 50 : N_{\max}$.

4. Determine the set of coefficients $h = \{\beta_0, \beta_1, \beta_2\}$ of $f(h, N) = \beta_2 N^2 + \beta_1 N + \beta_0$ that fits data points $\{\tilde{f}(N), N = 200 : 50 : N_{\max}\}$, where model complexity is represented by $h = \{\beta_0, \beta_1, \beta_2\}$.

5. Repeat steps 1–4 to estimate complexity for different parameter sets of a model structure.

[59] The first four steps of Algorithm 2 estimate the complexity of one parameter set only. Taking the median values of the ranges from Tables 1 and 2 (for SAC-SMA and SIXPAR, respectively), two equivalent parameter sets are obtained. Figure 3a plots the probability of exceedance, P_N, from (4) against γ for these parameter sets for N(sample size) = 200 and 4000. The rate at which this probability of exceedance converges as sample size increases, is the rate of convergence. As noted before, a slower rate of convergence implies higher complexity. An estimate of γ^* can also be obtained from this figure. As $N \to \infty$ the range of γ in which the transition of P_N from 1 to 0 takes place shrinks, eventually converging to the Heaviside function of (12). The value of γ^* thus lies in the range of this "transition" in Figure 3a. But then, also a range of β_2 (asymptotic complexity) can be estimated from it, since $\beta_2 = \gamma^{*2}$. Figure 3a therefore also suggests that the asymptotic complexity of SIXPAR model structure is lower than of SAC-SMA model structure.

[60] Algorithm 2 is applied to obtain 500 estimates of $f(h, N)$, corresponding to 500 parameter sets that are sampled for each model structure. Figure 3b shows the distributions of probability of exceedances for SIXPAR and SAC-SMA at $\gamma^N_{\max}$ (estimated in step 2 of Algorithm 2) for different values of N, using the model predictions based on 500 parameter set samples for each model structure. It also shows the medians of model predictions (solid lines). The observation that the median probability of exceedance for SIXPAR is pointwise lower than SAC-SMA indicates that SAC-SMA is more complex than SIXPAR. The boxplots at $N = 500, 1000, ..., 4000$ give an indication of the spread of the probability of exceedance across the different parameter sets. A second observation that the interquantile ranges

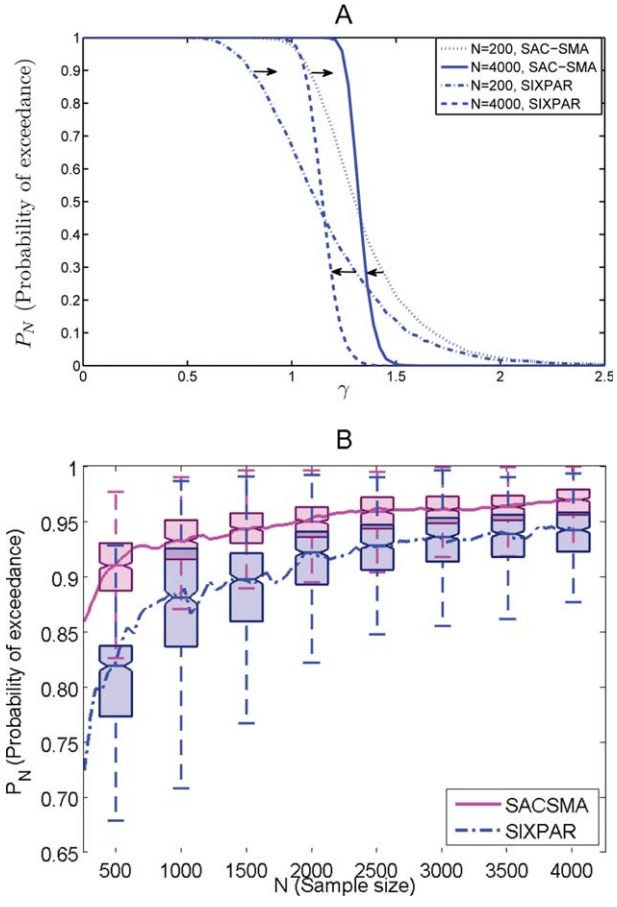

Figure 3. Probability of exceedance against γ and N. (a) $P\left(\sum_{t=1}^{N}|y(t)-E[y(t)]|>N\gamma\right)$ against γ for SAC-SMA and SIXPAR model are shown for $N=200$ and $N=4000$. The probability of exceedance converges for any γ as $N\rightarrow\infty$. (b) The spread of the probability of exceedance versus N at $\gamma=\gamma_{\max}^{N}$, across 500 different parameter samples of respective model structures. The lines show the median value at each sample size.

of SAC-SMA and SIXPAR for the same N do not overlap significantly, further supports the claim that the median values of the probability of exceedances, P_N, of both models are different.

[61] Further the coefficient β_2, that is estimated in step 4 of Algorithm 2, is the asymptotic complexity and can be used to compare complexities of SAC-SMA and SIXPAR. Figure 4 shows the boxplots of β_2 for each model structure (over 500 parameter sets). It shows that various quantile values of β_2 of SIXPAR are smaller than those of SAC-SMA. It therefore suggests that the asymptotic complexity of SIXPAR is lower than that of SAC-SMA.

[62] 6. Complexity Regularized Model Selection: Intercomparison Between Different Rainfall-Runoff Model Structures

[63] Section 5 quantified and compared the complexities of two model structures SAC-SMA and SIXPAR. In this section, we provide another algorithm that estimates the upper bound on prediction error given by inequality (7). The algorithm is then implemented for five different model structures of varying complexities (that are quantified by Algorithm 2). The algorithm is presented below as Algorithm 3.

Algorithm 3. (Estimation of upper bound on prediction error):

1. Sample P parameter sets for a model structure $\mathfrak{M}_l$.
2. For K values of $c=\frac{\eta}{\sqrt{\lambda}}$ between $(c_{\min}, c_{\max})$ on a logarithmic scale, calculate $T_1=\xi_N+c\sqrt{F(h,N)}$ for each parameter set on a data set D of length N ($F(h,N)$ is computed by Algorithm 2).
3. For each c, determine the minimum of T_1 over the P different parameter sets. The minimum of T_1 yields an optimal parameter set.
4. Calculate $T_2=\xi_{N'}$ for each optimal set corresponding to a value of c obtained in step 3 on another data set D', independent of D, of length N'.
5. Minimize T_2 and denote the parameter set and c corresponding to the minimum obtained in step 4 by $\theta_{l,N}^{*}$ and $c_{l,N}^{*}$.
6. Repeat steps 1–4 over the different model structures $l=1,\ldots,L$.
7. Calculate $T_3=\xi_{N''}$ for parameter set $\theta_{l,N}^{*}$ corresponding to each model structure $\mathfrak{M}_l$ on a third independent data set D'' of length N'' and rank the model structures 1 to L, where 1 is given to the structure that has the lowest value of T_3.

[64] The algorithm is implemented for $P=500$, $L=5$ (the five model structures are described in Appendix A), $K=10000$, $c_{\min}=10^{-3}$, $c_{\max}=10^{3}$, $N'=N''=5$ years and N takes values of $\frac{1}{3}$, $\frac{1}{2}$, and 1 year for three different experiments. Daily precipitation, evaporation, and streamflow data set of Guadalupe river basin [*Duan et al.*, 2006] is used to implement the algorithm. The data lengths $N'=N''=5$ years are sufficiently large such that sampling uncertainty is minimal.

[65] Algorithm 3 selects a model of optimal complexity for each model structure l in steps 2–4. The model for a

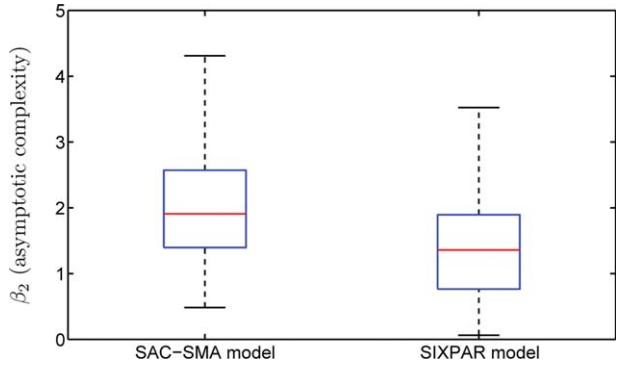

Figure 4. Boxplot of asymptotic complexity. Boxplot for β_2 (asymptotic complexity) of SAC-SMA and SIXPAR model for parameters sampled from the ranges in Table 1 (SAC-SMA) and ranges equivalent to these ranges (SIXPAR). This figure shows that the asymptotic complexity of SIXPAR model is lower than that of SAC-SMA model.

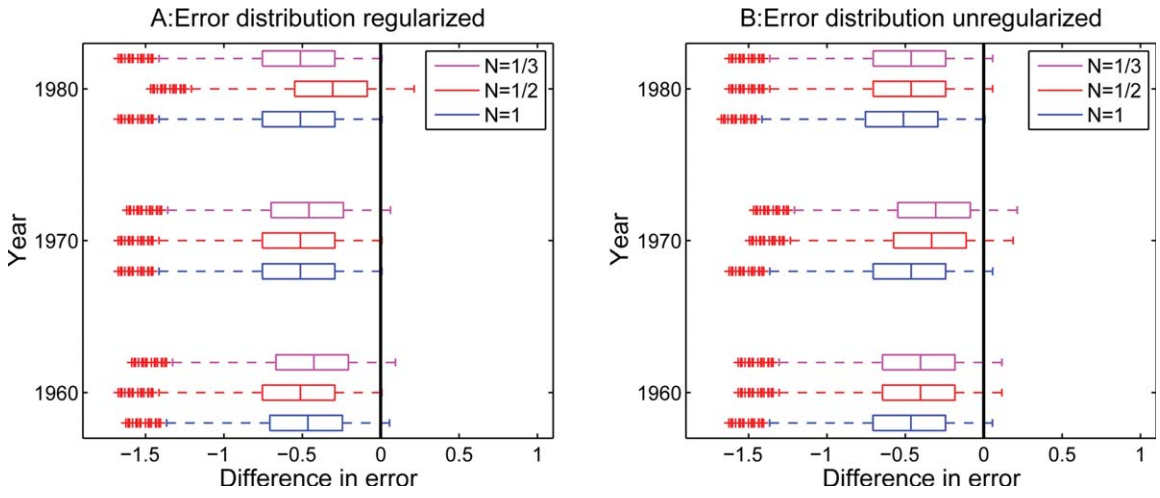

Figure 5. Distribution of the difference between the performance (empirical error) of models corresponding to a supposed optimal parameter set $\theta^*_{l,N}$ or $\tilde{\theta}_{l,N}$ and all 499 other parameter sets on a test set D'' from 1990 to 1994 for model structure 1. (a) Regularized model selection $(\theta^*_{l,N})$. (b) Unregularized model selection $\left(\tilde{\theta}_{l,N}\right)$.

given model structure $\mathfrak{M}_l$ is selected by a split sample test. The split sample test obtains a penalty $c^*_{l,N}$ such that the upper bound on prediction uncertainty (from equation (7)) is tightest for a model that minimizes this upper bound. The model that then minimizes the upper bound is the model selected from the model structure $\mathfrak{M}_l$ when the data size is N. The model selected from $\mathfrak{M}_l$ that corresponds to $\theta^*_{l,N}$ is then complexity regularized and has "optimal" complexity for the given data size N.

[66] The model corresponding to $\theta^*_{l,N}$ has better performance on future unseen data than a model, say corresponding to a parameter set $\tilde{\theta}_{l,N}$ that is selected by only minimizing ξ_N on data D of length N (unregularized model selection). This especially holds for small sample size N. The robustness in the performance of complexity regularized hydrological model is due to stability imparted by controlling for complexity. Further, it is a consistent estimator in the sense that $\theta^*_{l,N} \rightarrow \tilde{\theta}_{l,N}$ as N becomes large.

[67] For a given model structure $\mathfrak{M}_l$, a model selected based on complexity regularization $(\theta^*_{l,N})$ performs better than a model that is selected without regularization $\tilde{\theta}_{l,N}$ on future unseen data. The performance of $(\theta^*_{l,N})$ on an independent data set is a better representative of what a model structure $\mathfrak{M}_l$ is capable of than $\tilde{\theta}_{l,N}$. Hence, the performance of $(\theta^*_{l,N})$ on an independent data set D'' of length N'' is used to rank model structures in terms of their suitability to model the underlying processes of the study area (step 7).

[68] The performances of the models with parameters $(\theta^*_{l,N})$ (obtained from step 5 of Algorithm 3) and $\tilde{\theta}_{l,N}$ are compared against the model performance of models corresponding to all other $P-1$ parameter sets ($P=500$). This is done on a test data set $\hat{D}$ of size $\hat{N}=5$ years. The test data set does not overlap with the data sets D of size N that are used to estimate $(\theta^*_{l,N})$. The same data sets D are also used to estimate $\tilde{\theta}_{l,N}$. Figure 5a displays the boxplots of the differences between the empirical errors corresponding to $P-1$ parameter sets (excluding $(\theta^*_{l,N})$) and the empirical error computed by a model with $(\theta^*_{l,N})$. It is done so for three nonoverlapping data sets D of size N for model structure 1, i.e., $l=1$. The size of data set D takes values of $N=\frac{1}{3}, \frac{1}{2}$, and 1 year. Similarly, Figure 5b displays the boxplots of the differences between the empirical errors of models corresponding to $P-1$ parameter sets (excluding $\tilde{\theta}_{l,N}$) and the empirical error computed by a model with $\tilde{\theta}_{l,N}$.

[69] Figure 5a demonstrates that complexity regularized model performance is relatively stable with increasing sample size in the sense that the fraction of positive differences do not reduce or increase with increasing sample size. The fraction of positive differences in errors decreases with increasing sample size for unregularized model selection. However, regularized model selection is more often better than nearly all other $P-1$ models for all sample sizes than unregularized model selection. The distribution of differences is also shifted more to the left for regularized selection than for unregularized model selection for nearly all sample sizes.

[70] Figure 5 suggests that complexity becomes less relevant (or complexity regularized model selection converges to nonregularized model selection) when large data sets are used. This is a desirable property, often termed as consistency, since complexity regularized risk function such as on the RHS of equation (7) converges to expected empirical risk. Yet another observation that the distribution of the differences in error for complexity regularized model selection is shifted more to lower (negative) values than unregularized model selection is evidence of robust performance of complexity regularized model selection. This robust performance of complexity regularized model selection is due to stability (in Tihonov's sense) imparted to the model selection problem by penalizing model complexity.

[71] Figure 6 further demonstrates the stability (and thus robustness) introduced by complexity regularization in model selection problems. It plots the kernel cumulative density estimate of the difference between the performance (empirical error) of models corresponding to $\left(\theta^*_{l,N}\right)$ and $\tilde{\theta}_{l,N}$ for model structure 1 where 3 different lengths of D,

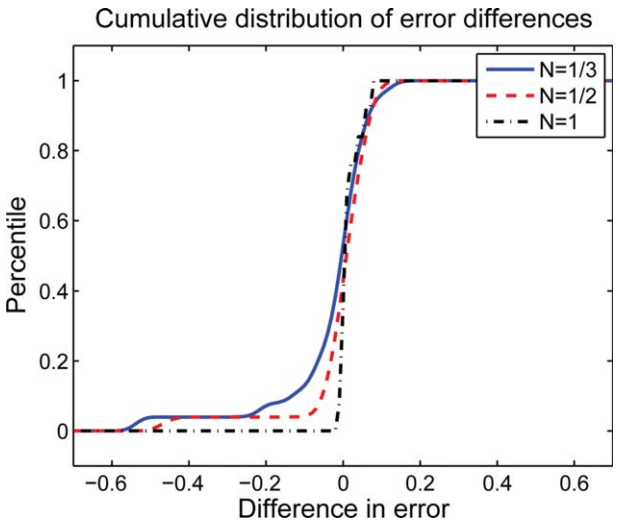

Figure 6. Kernel cumulative density estimate of the difference between the performance (empirical error) of models corresponding to $\theta^*_{l,N}$ and $\tilde{\theta}_{l,N}$ for model structure 1.

$N = \frac{1}{3}$, $\frac{1}{2}$, and 1 year are used to estimate $\left(\theta^*_{l,N}\right)$ and $\tilde{\theta}_{l,N}$. Twenty-four realizations of D for a given N are considered. These 24 realizations are all even years between 1948 and 1970 and between 1978 and 2000. The empirical errors are estimated on the same test data set of length 5 years from 1972 to 1976 such that it does not overlap with D. The distribution functions are fat tailed for negative values for sample sizes smaller than 1 year, while it is a Heaviside function at 0 for $N = 1$ year. The skewness in the distribution function reduces as the sample size increases, "converging" to the Heaviside function for $N = 1$ year. This demonstrates that complexity regularization is effective in producing robust performance for small sample size. Further, the figure demonstrates that complexity regularized model selection selects a consistent model.

[72] The model structures 1–3 are now assessed based on complexity regularized model selection. For a given data D of length N and a model structure $\mathfrak{M}_l$, steps 1–5 of Algorithm 3 provide a model corresponding to the parameter set $\theta^*_{l,N}$ that performs better than most other models corresponding to the other $P - 1$ parameter sets over future but yet unseen data. The performance of such a model on an independent data set D'' (from the same underlying but unknown distribution) therefore represents the best performance that the model structure $\mathfrak{M}_l$ can provide.

[73] Algorithm 3 is repeatedly applied using 5 years of data from 1973 to 1977 to construct 15, 10, and 5 data sets D of lengths $N = \frac{1}{3}$, $N = \frac{1}{2}$, and $N = 1$ year, respectively (for each N, various realizations of D are nonoverlapping) and eight data sets D'' of length $N'' = 5$ years spanning from 1948 to 1997 that do not overlap with D or D'. For each combination of D and D'', step 7 of algorithm 3 calculates the ranking of the three model structures. One realization of D' of length $N' = 5$ years is also required for regularized model selection (see steps 4 and 5 of Algorithm 3). A period from 1978 to 1982 is used for D'. This period is ignored for unregularized model selection since it only requires nonoverlapping data sets D and D''. This results in a total of 15*8, 10*8, and 5*8 orderings for $N = \frac{1}{3}$, $N = \frac{1}{2}$, and $N = 1$ year lengths of D, respectively. Note that a model is selected for a given model structure on each realization of D of length N. Thus, three models corresponding to the three model structures are selected on each D. These models represent the best that the corresponding model structures can do in replicating the observations. The performance of these models on a nonoverlapping data set D'' is therefore used to rank the corresponding structures in terms of their (complexity regularized) suitability for the study area. The frequency with which a model structure is ranked the best over the combinations of one realization of D and all eight realizations of D'' for a given N is then estimated.

[74] The mean and standard deviation of these frequencies for each model structure and N is provided in Table 3. The table also provides the same statistics for unregularized model selection, i.e., when model complexity is not regularized when selecting a model for a given model using D. The table demonstrates that both regularized and unregularized model selection find model structure 2 to be the best structure for the study area at $N = 1$ year. The mean frequency is nearly the same and high for both. The standard deviation is low relative to the magnitude of mean frequency in both the cases. For $N = \frac{1}{2}$ year, the mean frequency of structure 2 for regularized model selection remains the same with standard deviation slightly higher than at $N = 1$ year. This is not the case for unregularized model selection; its mean frequency of structure being the best is lower at $N = \frac{1}{2}$ year than at $N = 1$ year. The standard deviation is also higher at $N = \frac{1}{2}$ year than at $N = 1$ year. Its standard deviation is also marginally higher than that of regularized model selection at $N = \frac{1}{2}$ year. Thus, at $N = \frac{1}{2}$ year, regularized model selection finds the winning model structure (i.e., 2, which is asymptotically the best given its converged performance at $N = 1$ year for both regularized and unregularized model selection) with higher confidence than unregularized model selection. By confidence here we mean that the mean frequency of structure 2 is 2 standard deviations away from 0 in the case of regularized model selection, unlike the unregularized case. At $N = \frac{1}{3}$ year, the

Table 3. Mean and Standard Deviation (in Square Brackets) of Winning Frequencies for a Given N[a]

	Regularized			Unregularized		
	$N = \frac{1}{3}$ year	$N = \frac{1}{2}$ year	$N = 1$ year	$N = \frac{1}{3}$ year r	$N = \frac{1}{2}$ year	$N = 1$ year
Structure 1	0.47 [0.35]	0.13 [0.32]	0.13 [0.26]	0.27 [0.38]	0.16 [0.32]	0.13 [0.14]
Structure 2	0.4 [0.37]	0.75 [0.41]	0.75 [0.36]	0.48 [0.47]	0.56 [0.43]	0.78 [0.36]
Structure 3	0.13 [0.21]	0.13 [0.31]	0.13 [0.09]	0.26 [0.39]	0.28 [0.41]	0.1 [0.13]

[a]Two cases of complexity regularized and unregularized model selection are contrasted. 15, 10, and 5 nonoverlapping data sets D of lengths $N = \frac{1}{3}$, $N = \frac{1}{2}$ and $N = 1$ year, respectively, and eight nonoverlapping data sets D'' of length $N'' = 5$ years are considered.

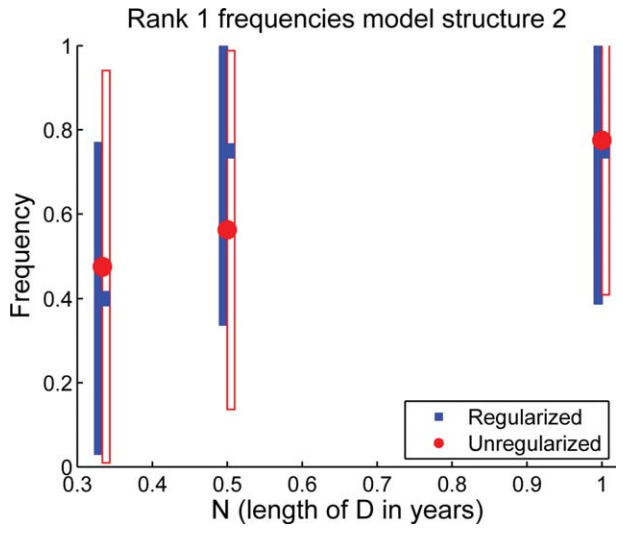

Figure 7. Mean and standard deviations of winning frequencies of model stucture 2 for $N = \frac{1}{3}$, $\frac{1}{2}$, and 1 year. In regularized model selection a faster convergence is seen than in unregularized model selection.

standard deviation of winning frequencies of model structure 2 for regularized model selection is still lower than corresponding standard deviation for the unregularized case. However, this time the regularized model selection finds model structure 1 to be a better choice for the study area based on mean wining frequency. Meanwhile unregularized model selection still chooses model structure 2 as the best although with higher standard deviation than regularized model selection at $N = \frac{1}{3}$ year and unregularized model selection at $N = \frac{1}{2}$ year.

[75] The standard deviation of winning frequencies of a model structure is higher for unregularized than regularized model selection at each N, except for model structure 1 at $N = 1$ year (where both regularized and unregularized model selection appear to have converged to each other in distribution sense). This indicates that complexity regularization stabilizes model selection since the variation in the rankings of the three model structures is lower for regularized model selection. However, stabilizing the variance of ranking introduces certain bias, especially at low sample size. This is probably the reason why regularized model selection at $N = \frac{1}{3}$ year finds structure 1 to be marginally better suited for the study area than structure 2. Nonetheless, for regularized model selection, all the model structures quickly converge to their asymptotic mean frequencies already at $N = \frac{1}{2}$ year. This is not the case for unregularized model selection.

[76] Figure 7 plots the mean and the standard deviation of the winning frequencies for model structure 2 for different values of N and for regularized and unregularized model selection. The faster convergence of mean frequency of being the best structure to its asymptotic value for regularized model selection than unregularized model selection is evident. Further, the difference in the standard deviation of the frequencies reduces with increasing sample sizes. It again demonstrates the role complexity as a stabilizer to model selection problems. It controls for potential ill-posedness in model selection by controlling the variance of selecting a model for a given model structure. Finally, the convergence of the ordering of model structures provided by regularized and unregularized model selection at $N = 1$ year (as shown in Table 3 and Figure 7) is evidence of consistent selection by the former.

[77] The ordering of model structures based on complexity regularized selection is also compared with the ordering estimated by BIC [*Kass and Raftery*, 1995]. The estimation of BIC requires maximum likelihood parameter estimation. We therefore acknowledge a weakness of such a comparison since we here limit ourselves to P samples of parameter sets for each model structure M_l. We also note that BIC tends to favor higher order models. BIC is estimated based on the following steps for each model structure: (1) A General Likelihood function and a Markov Chain Monte Carlo parameter sampler used in *Pande* [2013b] is used to obtain maximum likelihood parameter estimates that includes the parameters of the model structure and the parameters of the error model. (2) The maximum likelihood parameter estimates of the error model (after excluding the maximum likelihood parameters of the model) are then used alongside the P sampled parameter sets to estimate a model that has the maximum likelihood value amongst P candidates models corresponding to the P sampled parameter sets. (3) BIC is estimated using the parameter set out of the sampled P parameter sets, that maximizes the General Likelihood function.

[78] The General Likelihood function assumes a general distribution for the errors (residuals) between observations and model predictions. It accommodates autocorrelation and nonzero higher order moments of error (such as skewness and kurtosis). The parameters that describe the distribution of errors therefore include parameters related to the considered hydrological model structure and the parameters related to the general distribution function for errors that are not explained by the model structure.

[79] Table 4 shows the resulting frequencies for the ordering, using the estimation of BIC on the same eight test sets D'' as used in Table 3. BIC favors model structure 3 over the other two model structures.

[80] Algorithm 3 is applied again to order model structures 1, 4, and 5 (see Appendix A, for its description) using D and D'' realizations covering same periods for $N = \frac{1}{3}$, $\frac{1}{2}$, 1 years as for the analysis of model structures 1–3. Additional realizations of D of length $N = 2$ years are considered in order to demonstrate the convergence of regularized model selection to unregularized model selection. This is required since the complexities of considered model structures are different from the complexities of model structures 1–3 (see Figure 8). All D and D'' realizations are nonoverlapping except for $N = 2$ years where a moving window of 2 years from 1973 to 1978 is considered. This is required to avoid any overlap between D, D', and D'' for

Table 4. Frequencies of Rank Numbers Based on Eight Nonoverlapping Data Sets D'' Using BIC

	Rank 1	Rank 2	Rank 3
Model 1	0	0.375	0.625
Model 2	0	0.625	0.375
Model 3	1	0	0

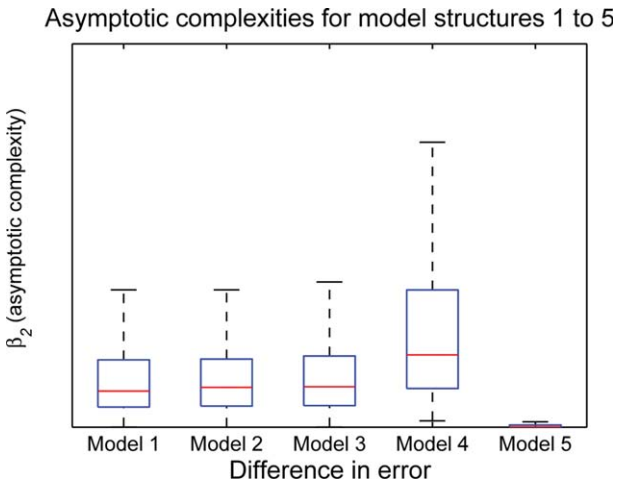

Figure 8. Boxplot of asymptotic complexity (β_2) for model structures 1–5. The figure shows a similar asymptotic complexity for model structures 1–3, but different values for model structures 4 and 5.

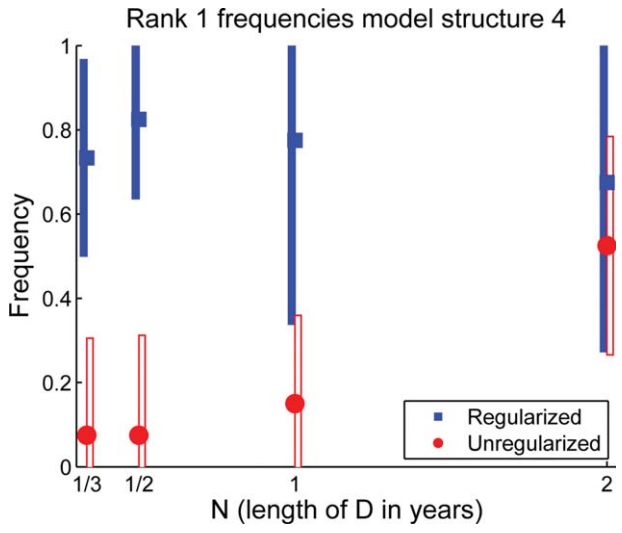

Figure 9. Mean and standard deviations of winning frequencies of model structure 4 for $N = \frac{1}{3}, \frac{1}{2}$, 1, and 2 year.

the case of regularized model selection (see steps 4 and 5 of Algorithm 3).

[81] Table 5 provides the results of model structure ordering for structures 1, 4, and 5. The table has the same construct as Table 3. Similar to Table 3, it also demonstrates that complexity regularization stabilizes model selection. The ranking based on regularized model selection is stable with standard deviation of frequencies of being the best model structure lower than unregualrized model selection. The unregularized model selection is highly unstable given that the rankings change till they convergence to the ranking of model structures under regularized model selection (in the sense of the best model structure) for $N = 2$ years (the rankings, as well as the mean and standard deviation of the frequencies of each model structure being the best, is similar for $N = 2$ years and $N = 5$ years for unregularized model selection).

[82] Figure 9 provides the mean and standard deviation of the frequencies of model structure 4 that is asymptotically the best structure amongst 1, 4, and 5 for both regularized and unregualrized model selection (see Table 5). The figure has the same construct as Figure 7. Similar to Figure 7, the mean of winning frequencies converge faster with increasing N to the asymptote for regularized model selection. Meanwhile, its standard deviation of the winning frequencies remains smaller than unregularized model selection.

Table 5. Mean and Standard Deviation (in Square Brackets) of Winning Frequencies for a Given N[a]

	$N = \frac{1}{3}$ year	$N = \frac{1}{2}$ year	$N = 1$ year	$N = 2$ year
Regularized				
Structure 1	0.18 [0.17]	0.11 [0.16]	0.03 [0.05]	0.1 [0.10]
Structure 4	0.74 [0.23]	0.83 [0.19]	0.78 [0.44]	0.68 [0.40]
Structure 5	0.09 [0.14]	0.06 [0.12]	0.20 [0.27]	0.23 [0.26]
Unregularized				
Structure 1	0.32 [0.41]	0.38 [0.43]	0.60 [0.42]	0.20 [0.17]
Structure 4	0.08 [0.23]	0.08 [0.24]	0.15 [0.21]	0.53 [0.26]
Structure 5	0.61 [0.47]	0.55 [0.42]	0.25 [0.22]	0.28 [0.28]

[a]Two cases of complexity regularized and unregularized model selection are contrasted. 15, 10, and 5 nonoverlapping data sets D of lengths $N = \frac{1}{3}$, $N = \frac{1}{2}$, and $N = 1$ year, respectively, and five overlapping data sets D of lengths $N = 2$ year are considered. For data sets D'' eight nonoverlapping sets of length $N'' = 5$ years are used.

7. Discussion and Conclusions

[83] This paper dealt with the problem of ill-posedness in hydrologic model estimation and hydrologic prediction uncertainty by expressing the latter as a trade-off between empirical risk and model complexity. We made no assumptions on the probability distribution of underlying processes and allowed dependency in model predictions over time. We formulated an expression for expected empirical risk in terms of empirical risk and a function of complexity, i.e., $E[\xi_N] \leq \xi_N + \eta\sqrt{F(h,N)/\chi}$. We also provided a geometric interpretation of model complexity as a statistic measuring the size of model output space. We however note that the notion of complexity used in this paper is not unique, several other notions exist [see, e.g., *Ye et al*., 2008; *Young et al*., 1996].

[84] We emphasized the need to consider model complexity if the expected empirical risk of two different models is to be compared given a finite sample for model estimation. In doing so, we provided an algorithm to calculate model complexity of an arbitrary hydrologic model and applied it to two hydrological model structures, SIXPAR and SAC-SMA. We found SIXPAR to have a smaller asymptotic complexity than SAC-SMA. We also provided an algorithm (Algorithm 3) based on the presented theory to calculate the prediction error. We applied it on five complex model structures with multiple states and fluxes that differed only in the number of routing reservoirs. The complexity regularized model selection based on Algorithm 3 was then compared with unregularized model selection that involved no penalization on model complexity. Both the selection problems were found to converge which provided evidence that complexity regularized model selection is a

consistent estimator. Further, it provided supporting evidence for the role of complexity regularization as a stabilizer of model selection problems. The regularized model selection was better able to pick the same model structure as the best approximation on small sample sizes. The variation in picking the winner and in calculating the frequencies of being a winner were also lower for regularized model selection than for unregularized model selection at almost all considered sample sizes.

[85] The theory presented is limited by Assumption 1 and restricted by the lack of assumptions on the underlying process distribution, data and on the type of hydrological models used since additional assumptions can facilitate tighter bounds. Assumption 1 simplifies the relationship between the deviation of empirical risk from its expected value and the deviation of prediction of a hydrological variable of interest from its expected value. It assumes that the former is a multiple of the latter, and therefore implicitly assumes that the effect of observed time series of a hydrological variable can be encapsulated by a multiplier. Such an assumption can result in weak upper bounds on rates of convergence such as in (5) which in turn may result in conservative assessment of prediction uncertainty. The lack of assumptions on the underlying process distribution such as the assumptions on the error structure and related probability distributions can also result in weak upper bounds on the rate of convergence. However, a lack of such assumptions is deliberate since it makes the presented theory generic and applicable to a wide variety of hydrological modeling problems.

[86] Our geometric interpretation of model complexity in part relies on that $E[|y(t) - E[y(t)]|] = \lim_{N\to\infty} \left(\sum_{t=1}^{N} |y(t) - E[y(t)]|\right)/N$, where model predictions are dependent over time. We intend to further investigate the validity of this statement and estimate complexity of various hydrological models to infer contribution of model relative to input data to prediction complexity and uncertainty.

Appendix A: Model Structures 1–5

[87] Five conceptual rainfall-runoff model structures are considered. All five structures have explicit representation of the unsaturated and saturated zones as nonlinear reservoirs. The evaporation is a nonlinear function of the storage (moisture) in the unsaturated zone in all the model structures. The overland flow is a nonlinear function of moisture in the unsaturated zone except for model structure 4 where it is also nonlinearly related to the inverse of lower zone (saturated) moisture content. Interception is not considered by any of the model structures except by model structure 5. Daily precipitation and potential evapotranspiration are nonlinearly transformed to overland flow and actual evaporation, respectively, by all the model structures. The unsaturated zone contributes to the saturated zone through percolation that itself is a nonlinear function of storage in the unsaturated zone in all the structures. The lower reservoir contributes subsurface runoff as a linear function of its storage. The overland flow and the subsurface flows are then routed through a set of linear reservoirs. The five structures also differ in the number of routing reservoirs. Model structure 1 has three routing reservoirs connected in series, model structures 2 and 4 have two reservoirs connected in series and model structures 3 and 5 have only one routing reservoir. For a general description of the model structures, readers are referred to *Pande* [2013b].

[88] **Acknowledgments.** The authors thank the Editor and four referees including Paul Smith and Bellie Sivakumar for their comments that helped to improve the quality of the paper.

References

Boucheron, S., G. Lugosi, and O. Bousquet (2004), Concentration inequalities, in *Advanced Lectures on Machine Learning, Lecture Notes in Comput. Sci.*, vol. *3176*, edited by O. Bousquet, U. von Luxburg, and G. Rätsch, pp. 208–240, Springer, Berlin.

Burnash, R. J. C. (1995), The NWS river forecast system-catchment modelling, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 311–366, Water Resour. Publ., Highlands Ranch, Colo.

Cavanaugh, J. E., and A. A. Neath (1999), Generalizing the derivation of the Schwarz information criterion, *Commun. Stat. Theory Methods*, *28*, 49–66.

Clement, T. P. (2011), Complexities in hindcasting models when should we say enough is enough?, *Ground Water*, *49*, 620–629, doi:10.1111/j.1745-6584.2010.00765.x.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*, 1015–1031.

Duan, Q., et al. (2006), The Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*, 317, doi:10.1016/j.jhydrol.2005.07.031.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008), A weakly informative default prior distribution for logistic and other regression, *Ann. Appl. Stat.*, *2*(4), 1360–1383.

Gupta, H. V., and S. Sorooshian (1983), Uniqueness and observability of conceptual rainfall-runoff parameters percolation process examined, *Water Resour. Res.*, *19*, 269–276.

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

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*, 2637–2649.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*(430), 773–795.

Kavetski, D., and M. P. Clark (2010), The ancient numerical daemons of conceptual hydrological modeling. 2: Impact of time stepping scheme on model analysis and prediction, *Water Resour. Res.*, *46*, W10511, doi:10.1029/2009WR008896.

Kumar, P. (2011), Typology of hydrologic predictability, *Water Resour. Res.*, *47*, W00H05, doi:10.1029/2010WR009769.

Kundzewicz, C. W., and A. J. Robson (2004), Change detection in hydrological records a review of the methodology, *Hydrol. Sci.*, *49*(1), 7–19.

Lall, U., B. Rajagopalan, and D. G. Tarboton (1996), A nonparametric wet/dry spell model for resampling daily precipitation, *Water Resour. Res.*, *32*, 2803–2823, doi:10.1029/96WR00565.

Lee, T., and T. B. M. J. Ouarda (2012), Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition, *Water Resour. Res.*, *48*, W02514, doi:10.1029/2011WR010660.

Marquardt, D. W., and R. D. Snee (1975), Ridge regression in practise, *Am. Stat.*, *29*(1), 3–20.

Mehrotra, R., S. Westra, A. Sharma, and R. Srikanthan (2012), Continuous rainfall simulation. 2: A regionalized daily rainfall generation approach, *Water Resour. Res.*, *48*, W01536, doi:10.1029/2011WR010490.

Meir, R. (2000), Nonparametric time series prediction through adaptive model selection, *Mach. Learning*, *39*, 534.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Houser (2005), Dual state parameter estimation of hydrological models using ensemble

Kalman filter, *Adv. Water Resour.*, *28*(2), 135–147, doi:10.1016/j.advwatres.2004.09.002.

Pande, S. (2013a), Quantile hydrologic model selection and model structure deficiency assessment: 1. Theory, *Water Resour. Res.*, *49*, doi:10.1002/wrcr.20411, in press.

Pande, S. (2013b), Quantile hydrologic model selection and model structure deficiency assessment: 2. Applications, *Water Resour. Res.*, *49*, doi:10.1002/wrcr.20422, in press.

Pande, S., M. McKee, and L. A. Bastidas (2009), Complexity-based robust hydrologic prediction, *Water Resour. Res.*, *45*, W10406, doi:10.1029/2008WR007524.

Pande, S., L. A. Bastidas, S. Bhulai, and M. McKee (2012), Parameter dependent convergence bounds and complexity measure for a class of conceptual hydrological models, *J. Hydroinformatics*, *14*(2), 443–463, doi:10.2166/hydro.2011.005.

Parrish, M., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, *48*, W03519, doi:10.1029/2011WR011116.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

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

Slate, E. H. (1994), Parameterizations for natural Exponential families with quadratic functions, *J. Am. Stat. Assoc.*, *89*(428), 1471–1482.

Tierney, T., and J. B. Kadane (1986), Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.*, *81*(393), 82–86.

Vapnik, V. (1982), *Estimation of Dependencies Based on Empirical Data*, Springer, New York.

Vapnik, V. (2002), *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. M. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*, 273–290.

Ye, M., Meyer, P. D., and Neuman, S. P. (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.

Young, P., S. Parkinson, and M. Lees (1996), Simplicity out of complexity in environmental modelling: Occam's razor revisited, *J. Appl. Stat.*, *23*(2–3), 165–210.

Yu, B. (1994), Rates of convergence for empirical processes of stationary mixing sequences, *Ann. Probab.*, *22*, 94–116.