WATER RESOURCES RESEARCH, VOL. 44, W00B03, doi:10.1029/2008WR006836, 2008

Model complexity control for hydrologic prediction

G. Schoups,[1] N. C. van de Giesen,[1] and H. H. G. Savenije[1]

Received 15 January 2008; revised 14 May 2008; accepted 10 June 2008; published 28 August 2008.

[1] A common concern in hydrologic modeling is overparameterization of complex models given limited and noisy data. This leads to problems of parameter nonuniqueness and equifinality, which may negatively affect prediction uncertainties. A systematic way of controlling model complexity is therefore needed. We compare three model complexity control methods for hydrologic prediction, namely, cross validation (CV), Akaike's information criterion (AIC), and structural risk minimization (SRM). Results show that simulation of water flow using non-physically-based models (polynomials in this case) leads to increasingly better calibration fits as the model complexity (polynomial order) increases. However, prediction uncertainty worsens for complex non-physically-based models because of overfitting of noisy data. Incorporation of physically based constraints into the model (e.g., storage-discharge relationship) effectively bounds prediction uncertainty, even as the number of parameters increases. The conclusion is that overparameterization and equifinality do not lead to a continued increase in prediction uncertainty, as long as models are constrained by such physical principles. Complexity control of hydrologic models reduces parameter equifinality and identifies the simplest model that adequately explains the data, thereby providing a means of hydrologic generalization and classification. SRM is a promising technique for this purpose, as it (1) provides analytic upper bounds on prediction uncertainty, hence avoiding the computational burden of CV, and (2) extends the applicability of classic methods such as AIC to finite data. The main hurdle in applying SRM is the need for an a priori estimation of the complexity of the hydrologic model, as measured by its Vapnik-Chernovenkis (VC) dimension. Further research is needed in this area.

Citation: Schoups, G., N. C. van de Giesen, and H. H. G. Savenije (2008), Model complexity control for hydrologic prediction, *Water Resour. Res.*, *44*, W00B03, doi:10.1029/2008WR006836.

1. Introduction

[2] Hydrologic prediction usually relies on incomplete and uncertain process descriptions that have been deduced from sparse and noisy data sets. A direct consequence is that water management decisions based on those models are subject to significant uncertainty. The uncertainty stems not only from conceptual errors in the models we use to simulate natural systems, but also from data estimation errors when inferring model parameter values from limited and noisy data. Two main modeling philosophies have been developed to tackle these problems, namely the upward or mechanistic approach, and the downward or data-driven approach [*Wagener et al.*, 2007]. Spatially distributed modeling is a typical example of the upward approach to construct a model that explicitly accounts for as much of the small-scale physics and the natural heterogeneity as computationally possible [*Freeze and Harlan*, 1969; *Loague and VanderKwaak*, 2004]. The approach has been criticized for resulting in models that are overly complex, leading to problems of overparameterization and equifinality [*Beven*, 1993, 2006], which may manifest itself in large prediction uncertainty [*Van der Perk*, 1997; *Uhlenbrook et al.*, 1999]. On the other end of the modeling spectrum, a downward or data-driven approach has been advocated, where complexity is added to the model only when it improves description of the data, without using an a priori defined model structure [*Jakeman and Hornberger*, 1993; *Young*, 2003; *Sivapalan et al.*, 2003; *Fenicia et al.*, 2008]. The idea is to arrive at models that are complex enough to explain the data, but not more complex than necessary, a strategy often referred to as Occam's razor or the principle of parsimony. For example, simple models may suffice to describe spatially integrated watershed response, despite significant heterogeneity within the watershed [*Savenije*, 2001].

[3] A key question that then arises is how to decide when the model has sufficient complexity. In computer science, specifically in the field of pattern recognition, it was recognized early on that some form of complexity control is needed [*Akaike*, 1970]. Several methods have been developed for this purpose, all of them relying on two ingredients: (1) specification of two or more model structures of varying complexity, and (2) evaluation and comparison of the ability of the models to mimic observed data. The a priori selection of various alternative model structures embodies the assumptions and previous physical knowledge about the hydrologic processes deemed to be important. Each model is calibrated to observed data and its ability in

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

0043-1397/08/2008WR006836$09.00

fitting data is compared to other models. A problem is that complex models are more flexible in fitting calibration data, so this approach typically favors complex models. Therefore, a better approach is to compare hydrologic models on the basis of performance in a split calibration-validation test, and select the model that gives the smallest prediction error for the validation data [*Klemeš*, 1986; *De Wit and Pebesma*, 2001; *Perrin et al.*, 2001]. In order to avoid problems with arbitrarily dividing data into calibration and validation sets, the split test can be repeated on different parts of the data set, resulting in resampling procedures such as bootstrapping and cross validation. A drawback is that for larger data sets and complex models, this approach can become computationally very intensive, since it involves solving many smaller calibration (parameter optimization) subproblems.

[4] An alternative to (cross) validation is to rely on parameter identifiability as a basis for model selection, by selecting the most complex model that still allows unique parameter values to be identified [*Jakeman and Hornberger*, 1993]. Here, the obvious advantage is that the selected model structure is guaranteed to have identifiable parameters.

[5] Another alternative is to rely on statistical model selection or model complexity control methods [*Cherkassky and Mulier*, 2007]. These are based on statistical theory and provide estimates of model prediction error by multiplying the calibration error by a factor that penalizes increasingly complex models. As such, an optimal model structure is identified through a balance between the data fitting ability of the model and its complexity. These statistical model selection methods differ in the functional form and assumptions of the analytical expressions for the penalization factor. Examples include Akaike's information criterion (AIC) [*Akaike*, 1970], and Bayes information criterion (BIC) [*Schwartz*, 1978]. These methods have been applied to model selection problems in rainfall time series modeling [*Gregory et al.*, 1992], groundwater modeling [*Honjo and Kashiwagi*, 1999; *Knotters and De Gooijer*, 1999], and flood frequency estimation [*Mutua*, 1994]. A common assumption in the development of both AIC and BIC is that an (infinitely) large data set is available for estimating prediction error. However, in many situations (e.g., hydrologic prediction) one is faced with limited and noisy data sets, in which case these methods do not strictly apply. An alternative model complexity control method that has been developed for use with finite data sets is structural risk minimization (SRM) [*Cherkassky and Mulier*, 2007]. Here an analytical upper bound (worst case) on prediction error is used to select a model of optimal complexity. In hydrology, SRM has been applied to model selection in rainfall-runoff modeling [*Dibike et al.*, 2001] and groundwater transport modeling [*Khalil et al.*, 2005], both studies using support vector machines instead of physically based hydrologic models.

[6] The purpose of this paper is to explore the usefulness of several statistical model complexity control methods for hydrologic prediction. In particular, we are interested in the performance of SRM compared to more traditional methods such as AIC, as it is based on assumptions that appear more realistic in hydrologic applications, i.e., finite and noisy data sets with unknown error distribution. We do this using several case studies. First, the various model complexity control methods considered in this paper are briefly discussed. This is followed by a short description of the case studies examined here. Then we present results of our comparison studies and discuss implications for hydrologic model selection.

2. Methods

2.1. Problem Formulation

[7] We are concerned with making predictions of some hydrological variable Y (e.g., river discharge, soil moisture content, groundwater level) as a function of a number of input variables X (e.g., rainfall, potential evaporation). The problem is tackled using a combination of observation and simulation. First, n data samples $Z = \{(x_i, y_i), i = 1\ldots n\}$ are obtained by simultaneously measuring inputs $X = \{x_i, i = 1\ldots n\}$ and outputs $Y = \{y_i, i = 1\ldots n\}$. Second, an approximating function or model $f(X, \theta)$ is introduced to infer system behavior from available observations,

$$Y = f(X;\theta) + \varepsilon_e \tag{1}$$

where θ is a set of model parameters, and ε_e is an empirical error between measured and simulated values, which includes observation and model errors. The goal is then to estimate values for the model parameters θ such that model f provides the best approximation of true system behavior, which is represented by an unknown "target" function g. The quality of this approximation can be quantified by prediction risk or error R_p, which measures expected error between true and approximate system behavior [*Cherkassky and Mulier*, 2007]. Mathematically,

$$R_p = \int L[g(x), f(x,\theta)]p(x)dx \tag{2}$$

where L is a general error or loss function describing the discrepancy between g and f, and $p(x)$ is the unknown probability density function (pdf) of input values. Since target function g is unknown, we estimate R_p on the basis of observations $Z = \{(x_i, y_i), i = 1\ldots n\}$ and define empirical error R_e as

$$R_e = \frac{1}{n}\sum_{i=1}^{n} L[y_i, f(x_i, \theta)] \tag{3}$$

[8] Loss function L can take many forms depending on the application and assumptions about the model error structure. Under maximum likelihood estimation, and for Gaussian errors ε_e that are independent and identically distributed (i.i.d.), L becomes the squared loss function, and empirical error R_e is given by the mean squared error (MSE) between observed and simulated values [*Cherkassky and Mulier*, 2007].

[9] For large and accurate data sets, i.e., $n \to \infty$ and measurements y_i closely follow true system behavior g, we find that empirical error R_e in (3) provides a good estimate of true prediction error R_p in (2). However, for finite and noisy data sets, it tends to underestimate true prediction error. The reason for this is that model f can usually be made to fit the data quite well by varying model parameter values, resulting in small values for ε_e and R_e. This is especially the

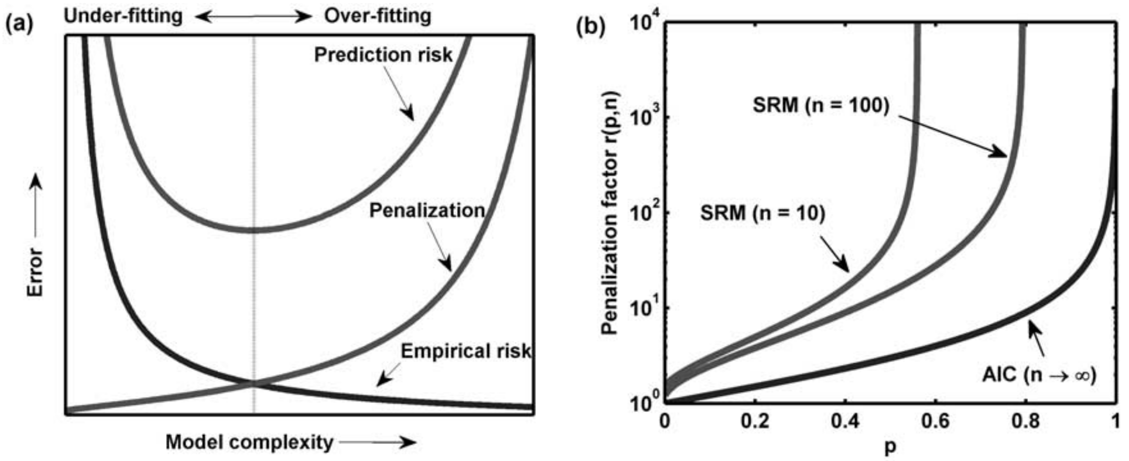

Figure 1. (a) Conceptual diagram of the trade-off between empirical fit to data and penalization of model complexity, with resulting effect on prediction risk or error according to (4), and (b) penalization factor r as a function of p and sample size n, according to two statistical theories, i.e., AIC in (5) and SRM in (6).

case for complex models with many adjustable parameters. When only limited data are available for estimating model parameters, the model may not perform well on new data, as it is only representative of a small subset of the entire data population. This is even more so if the data are noisy. In other words, for small noisy data sets, model selection through minimization of (3) results in models that may provide a poor approximation of true system behavior. In the following section, we review methods for model selection when available data are sparse and noisy.

2.2. Model Complexity Control Methods

[10] Overconditioning of the model on a small and noisy data set can be avoided by reducing the complexity (number of parameters) of the model. For this reason, model selection is often referred to as model complexity control. In this section, we briefly describe existing methods used for model complexity control. These all rely on (1) specifying a range of model structures to be considered, (2) estimating prediction error R_p in equation (2) for each model structure, (3) selecting the model structure with smallest value for R_p as optimal, and (4) estimating parameters of the optimal model structure by minimizing empirical error in equation (3) using all the data. The main differences between the various methods lie in the way that prediction error is calculated, i.e., the second step. In general, we can distinguish between analytical and numerical methods for estimating prediction error R_p in equation (2).

2.2.1. Numerical Methods

[11] A common numerical technique is k-fold cross validation (CV), whereby the entire data set is split into k learning (or calibration) and validation subsets, and prediction error is quantified on the basis of performance of the calibrated models on k validation subsets. The procedure is repeated for various model structures, and the model structure that yields the smallest validation error, averaged over all k subsets, is selected as optimal. The k-fold cross-validation method is simple and robust, and does not rely on any assumptions about the residuals, such as independence and normality. The main disadvantage is that it is computationally intensive. Cross validation can be considered to be a general form of the common practice of split testing in hydrology, i.e., dividing the available data set into two parts, one for calibration, and one for validation. The advantage of cross validation is that it (1) avoids ad hoc decisions on how to split the data and (2) makes use of all available data. The algorithm for k-fold cross validation is detailed by *Cherkassky and Mulier* [2007], including several other methods for generating cross-validation data sets.

2.2.2. Analytical Methods

[12] Alternatively, one can derive analytical estimates of prediction error on the basis of statistical theory. In general, these estimates take the following form:

$$\hat{R}_p = rR_e \tag{4}$$

where r is a penalization factor [*Cherkassky*, 2002], $r \geq 1$. Alternatively, equation (4) can be formulated as the sum of empirical error R_e and a regularization term, as in Tikhonov regularization and its variations. Penalization factor r increases as a function of model complexity relative to size of the data set, resulting in a trade-off (Figure 1a) between (1) empirical fit to the data, as quantified by R_e, and (2) model complexity relative to data availability, as incorporated in r. Various analytical forms for penalization factor r have been proposed, including Akaike's information criterion (AIC) [*Akaike*, 1970], and Bayes information criterion (BIC) [*Schwartz*, 1978]. In the case of AIC, the penalization factor can be written as,

$$r_{AIC} = \frac{1+p}{1-p} \tag{5}$$

where $p = d/n$, n is sample size, and d is degree of freedom or number of parameters, a measure of model complexity. Equation (5) leads to a form of AIC that is known as final prediction error (FPE) [*Akaike*, 1970].

[13] A disadvantage of these classic analytical approaches for model selection and penalization is that they are based on an assumption of large sample size, or $n \rightarrow \infty$. For this reason they may overestimate model complexity in the case

Table 1. Assumptions of Model Complexity Control Methods[a]

	AIC	SRM	CV
Sample size	large	any	any
Noise distribution	Gaussian, i.i.d.	any, i.i.d.	any
Needs estimate of error variance	yes	no	no
Supported models	linear	(non)linear	(non)linear
Assumes true model included in set of candidate models	yes	no	no
Penalization method	analytical	analytical	numerical

[a] AIC, Akaike's information criterion; SRM, structural risk minimization; CV, cross validation; i.i.d., independent and identically distributed.

of small sample sizes [*De Ridder et al.*, 2005]. An alternative method that is also valid for finite sample sizes uses structural risk minimization (SRM) [*Cherkassky and Mulier*, 2007]. SRM is rooted in statistical learning theory (SLT) [*Vapnik*, 1998] and relies on the following expression for penalization factor r [*Cherkassky and Mulier*, 2007],

$$r_{SRM} = \left[1 - \sqrt{p - p\ln p + \frac{\ln n}{2n}}\right]_{+}^{-1} \qquad (6)$$

where $p = h/n$, and h is the Vapnik-Chervonenkis (VC) dimension of the model as a measure of model complexity. The VC dimension h of a model is related to its data-fitting flexibility, with larger values for h indicative of more complex models that can describe a larger number of random data points. As an example, consider a first-order polynomial with two parameters, slope and intercept. This model can be fitted exactly through any two, but not three, data points. Hence its VC dimension equals 2. In general, VC dimension of models linear in the parameters, such as polynomials, equals the number of parameters. However, VC dimension of nonlinear models can be either greater or smaller than the number of parameters. Therefore, in general VC dimension h of a model is unknown and needs to be determined through numerical experiments [*Shao et al.*, 2000]. Basically, such experiments empirically measure the model's ability or flexibility of describing randomly generated data sets of increasing size.

[14] According to (4) and (6), the SRM estimate of prediction error equals $r_{SRM}R_e$. Vapnik's statistical learning theory guarantees that this estimate provides an upper bound on true prediction error with probability equal to $1 - 4/\sqrt{n}$ [*Cherkassky and Mulier*, 2007, chapter 4]. Hence, in the limit of $n \rightarrow \infty$, penalization factor r_{SRM} converges to 1, and estimated prediction error converges to empirical error with probability equal to 1. Figure 1b illustrates the effects of model complexity h and sample size n on the values of r_{SRM} and r_{AIC}. Penalization is small for large n and for small values of p (h/n). Note that AIC penalizes complex models less than SRM ($r_{AIC} \leq r_{SRM}$), since AIC assumes infinite sample size.

[15] The goal of this paper is to compare performance of three model complexity control methods for hydrologic prediction, namely CV, AIC, and SRM. Table 1 gives an overview of the assumptions on which the methods are based. Generally speaking, CV is the most general and flexible method as it makes no assumptions about the error distribution in (1). On the other hand, AIC makes the strongest assumptions about both sample size and error distribution, and was developed for situations where the true model is contained in the set of approximating functions; that is, one of the models in the set describes the system behavior without error. SRM falls in the middle of these two extremes, as it has been especially designed for model selection from limited and noisy data sets, which is a more realistic situation in hydrologic applications. The focus in our comparison study is on identifying model structures that have a level of complexity that matches the amount and quality of available data, and that lead to identifiable parameters.

2.3. Case Studies

[16] In order to illustrate and compare various model complexity control methods, we consider flow from a measuring device for low flows, consisting of a tower with many small holes at various threshold heights, as shown in Figure 2 [*Stomph et al.*, 2002]. This setup forms the basis for generating synthetic data sets by simulating discharge Q (output Y) as a function of water height h_w (input X) in the tower using the known physics and threshold locations, and adding random noise. In this case, true system behavior is given by a discharge-height relationship,

$$Q(h_w) = c\sum_{i=1}^{n_h}\sqrt{\max(0, h_w - t_i)} \qquad (7)$$

where summation is over n_h holes (n_h = 20), t_i is height or threshold of the ith hole, and c is a constant. Each term in (7) represents a flow contribution from a single hole. This problem was chosen for illustration purposes because (1) it is of limited complexity, (2) the physics is exactly known, and (3) it involves threshold behavior often observed in hydrological processes.

[17] We consider several case studies to evaluate performance of the three complexity control methods (Table 2). First, we investigate how the type of approximating function or model f influences prediction errors and model selection. Two different types of approximating functions are used, namely polynomials of increasing order and physically based models with increasing number of threshold parameters t_i in (7). Second, the effects of both data quantity (in terms of sample size) and data quality (in terms of signal-to-noise ratio (SNR)) are investigated. In each case, synthetic data are generated as follows: n values for water height h_w are uniformly drawn between 0 and 30 cm, and corresponding values for outflow Q are obtained using true system behavior (7) perturbed by an additive noise term, randomly drawn from a normal distribution N (0, σ_0^2). Error variance σ_0^2 is obtained as a function of specified signal-to-noise ratio, since $SNR = \sigma/\sigma_0$ with σ^2 the variance

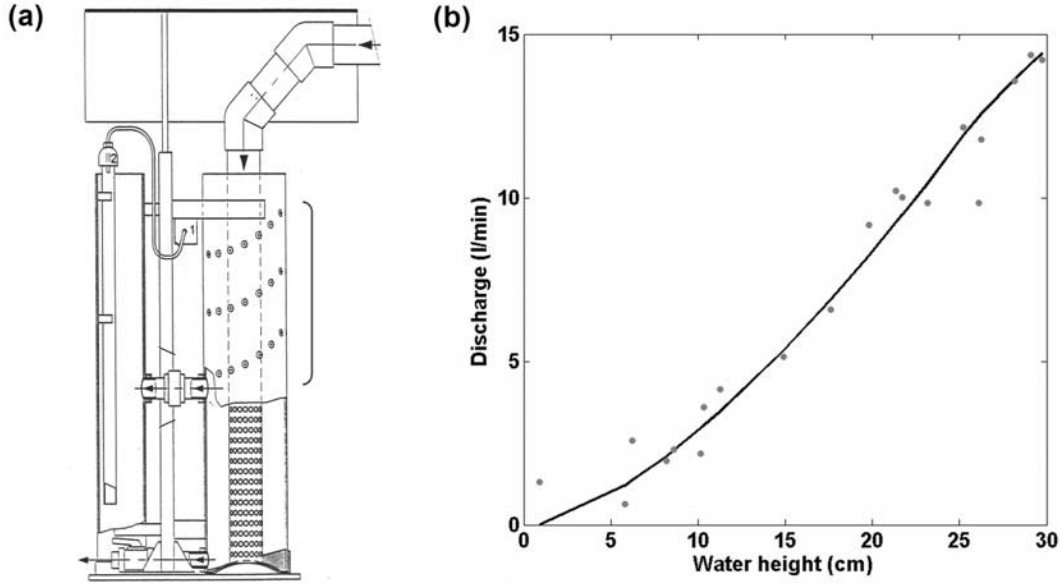

Figure 2. (a) Tower with holes at different threshold heights for measuring flow as a function of water height and (b) example of a synthetic data set, along with the underlying true relationship (solid line) between water height in and discharge from the tower.

of true discharges. Figure 2 shows an example of a synthetic data set of water heights and discharges, including the underlying true relation between the two variables.

[18] Parameter identification relies on minimizing empirical error in (3) using the squared loss function L. For polynomial models, this is done using linear least squares, whereas nonlinear physically based models are calibrated using a gradient-based nonlinear parameter optimization algorithm. In order to obtain continuous first derivatives we smooth out the original threshold-based function in (7) using the Chen-Harker-Kanzow-Smale max function [*Kavetski and Kuczera*, 2007],

$$\max(0, h_w - t_i) \approx 0.5\left[h_w - t_i + \sqrt{(h_w - t_i)^2 + m}\right] \tag{8}$$

where m is a smoothing parameter, which is chosen to be small enough to provide an excellent approximation to the original function (here, $m = 10^{-4}$). Model selection is performed with three complexity control methods, namely CV, AIC, and SRM. For AIC and SRM, we use the number of model parameters, i.e., number of polynomial coefficients or number of threshold heights t_i, as an estimate of both degrees of freedom d in (5) and VC dimension h in (6). Since for nonlinear models exact values for d and h are not known [*Cherkassky and Mulier*, 2007], this is an approximation. Rigorous estimation of VC dimension of nonlinear models can be done using the approach of *Shao et al.* [2000]. Following *Cherkassky et al.* [1999], we compare the methods using several performance indices: (1) prediction error R_p (2) of the selected model: measures how well the optimal model approximates true system behavior in (7); (2) error estimation accuracy: measures how well true prediction error R_p in (2) is estimated by each model selection method; (3) model complexity of the selected model: indicates how complex the optimal model is, i.e., how many parameters the model has; and (4) total parameter variance of the selected model: measures parameter identifiability or equifinality of the optimal model. Effects of sampling variation are accounted for by repeating model selection on 100 synthetic data sets, and reporting results as box plots of the four performance indices.

[19] In order to assess performance of model complexity control under more realistic conditions, where assumptions of data independence and Gaussian errors may not hold, we also consider a real-world case study using hydrologic data, namely daily rainfall, potential evaporation, and streamflow data from the Leaf River basin in Mississippi [*Duan et al.*, 2006]. We apply a nonlinear soil moisture accounting model to transform rainfall and evaporation into effective rainfall, which is further transformed into streamflow using a simple, linear reservoir lumped rainfall-runoff model [*Wagener et al.*, 2001]. Complexity is introduced by allowing model

Table 2. Different Case Studies in the Synthetic Flow Problem Depicted in Figure 2

Case	Approximating Functions	Sample Size, n	Signal-to-Noise Ratio	Error Distribution
1	polynomials	20	5	Gaussian
2	physically based models	20	5	Gaussian
3	physically based models	5	5	Gaussian
4	physically based models	20	1	Gaussian

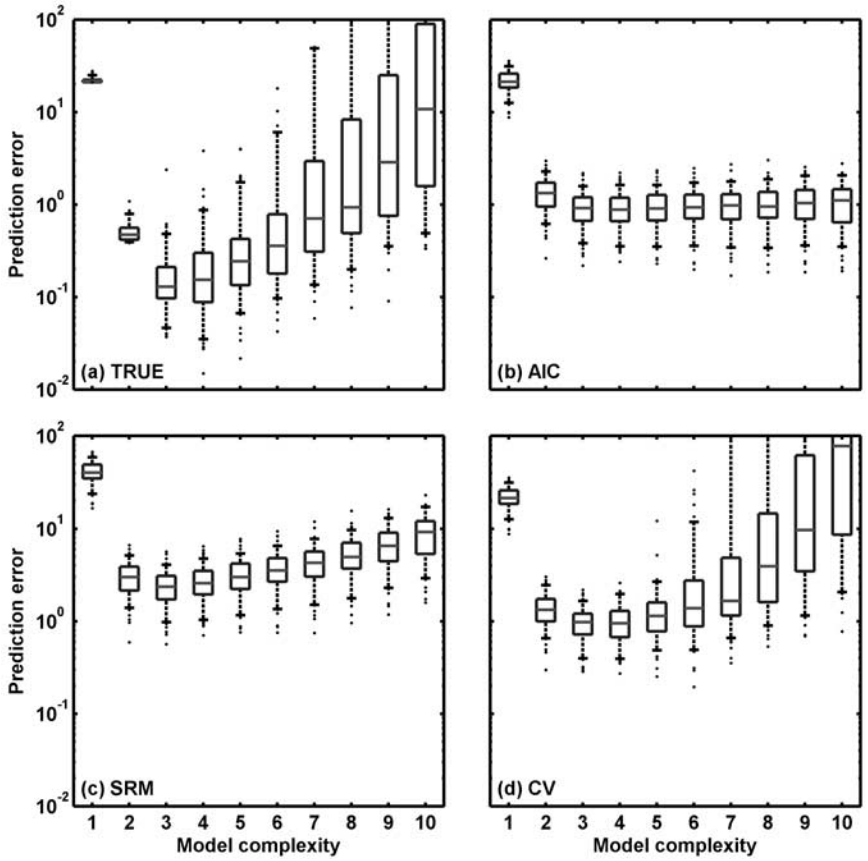

Figure 3. Prediction error as a function of polynomial model complexity: (a) true prediction error given by (2) and (b–d) prediction error estimated with (4) using AIC, SRM, and CV, respectively (AIC, Akaike's information criterion; SRM, structural risk minimization; CV, leave-one-out cross validation). Box plots summarize statistical results for 100 samples of size $n = 20$ and an SNR (signal-to-noise ratio) equal to 5.

parameters to be variable in time. Model selection results for this case are presented in section 3.5.

3. Results

3.1. Case 1: Polynomial Models

[20] The first case uses polynomials of increasing order to approximate true system behavior in (7). This corresponds to a case of no physical insight; that is, we assume no knowledge about the underlying process of generating outflow as a function of water height in the tower. The system is treated as a black box relating inputs to outputs, without any physical insight or physical meaning attached to model parameters, i.e., polynomial coefficients. This approach is typical in neural network applications.

[21] Figure 3 shows how well polynomials of increasing order are able to approximate system behavior on the basis of noisy data with sample size $n = 20$ and signal-to-noise ratio $SNR = 5$. True prediction error in Figure 3a initially decreases for polynomials of increasing order, but then increases as polynomials become more complex. There is an optimal model complexity where prediction error is minimal: lower-order models underfit the data, whereas higher-order models overfit the data. Hence, the danger of overfitting is clearly present for polynomial models in Figure 3a. Figures 3b–3d indicate how well different model complexity control methods are able to estimate true prediction error depicted in Figure 3a. Increasing prediction error with increasing polynomial complexity is only accurately reproduced by CV (Figure 3d), whereas SRM (Figure 3c) and especially AIC (Figure 3b) underestimate both the increasing trend and variation in prediction error for high-order polynomials. An important consequence of overfitting is that model parameters are not uniquely identifiable anymore, resulting in parameter equifinality. This is illustrated in Figure 4, where total variance of polynomial coefficients is plotted for increasing polynomial order. It is clear that parameter identifiability is optimal for 2 to 4 parameters and becomes worse for more complex models. This parallels trends in prediction error (Figure 3a).

[22] Model selection results for polynomial models are shown in Figure 5. Application of SRM results in models that generally have lower prediction error than either AIC or CV (Figure 5a). This is particularly true when considering the 95% percentile (upper whisker in box plots), which is lower for SRM than for AIC or CV. The accuracy with which prediction error of optimal models is estimated is fairly similar for the three methods (Figure 5b). SRM also selects polynomials with the smallest number of parameters (Figure 5c), compared to AIC and CV, and consequently yields the best parameter identifiability and lowest parameter variance in Figure 5d. Both AIC and CV select more

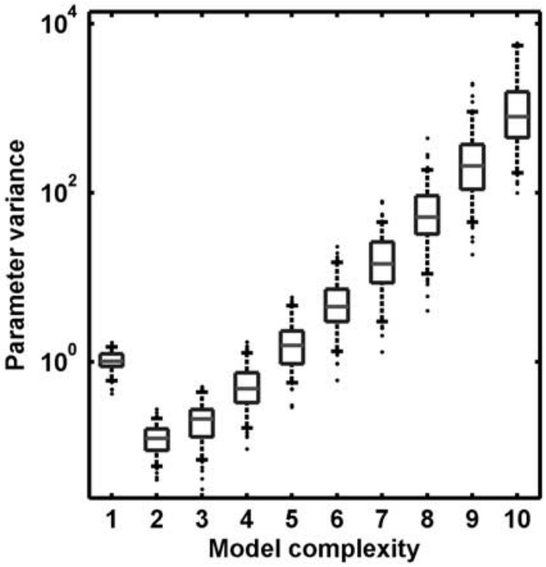

Figure 4. Parameter identifiability, expressed in terms of total parameter variance of optimal parameter values, as a function of polynomial model complexity (n = 20, SNR = 5).

complex models with greater prediction error and poorer parameter identifiability. Note finally that variance or spread of performance indices over 100 random samples is smallest for SRM, indicating robustness of SRM, i.e., its insensitivity to random sample variations.

3.2. Case 2: Physically Based Models

[23] The next step is to evaluate whether the conclusions of fitting with polynomials also hold when using physically based models to predict outflow from the tower on the basis of water height. In this case, we assume that the physical relationship between discharge from a single hole as a function of water height is known and given by an individual term in (7). However, the threshold heights are unknown and need to be determined on the basis of limited noisy data of discharge and water height. Model complexity can be gradually increased by including more holes with different threshold heights. Although the example is of limited complexity, this situation mimics hydrologic problems where large-scale behavior is estimated on the basis of knowledge of local-scale flow processes. This differs from the black box approach in the previous section.

[24] Figure 6 shows prediction errors for physically based models with an increasing number of holes. For reference, the synthetic data are based on a true model with 20 holes. Prediction error decreases as the model becomes more complex (more holes) and then flattens off indicating that

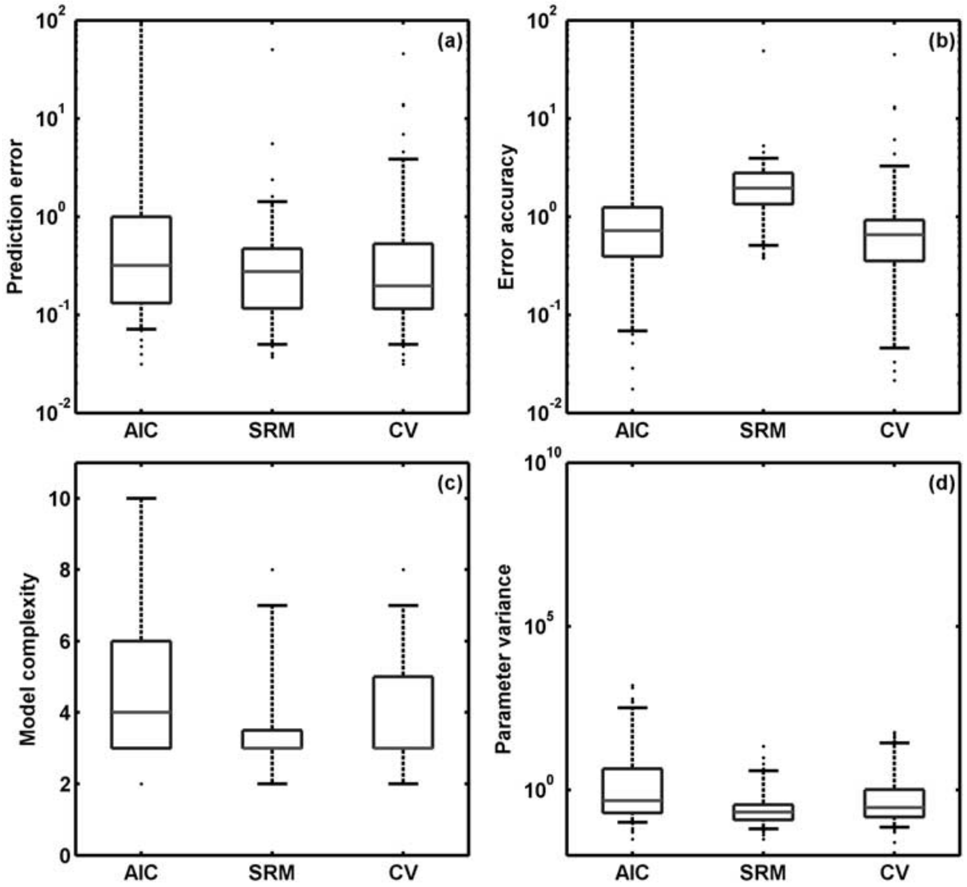

Figure 5. Model selection results for polynomial models: (a) true prediction error, (b) accuracy of estimated prediction error, (c) model complexity, and (d) parameter identifiability of optimal polynomial models according to three model complexity control methods (n = 20, SNR = 5).

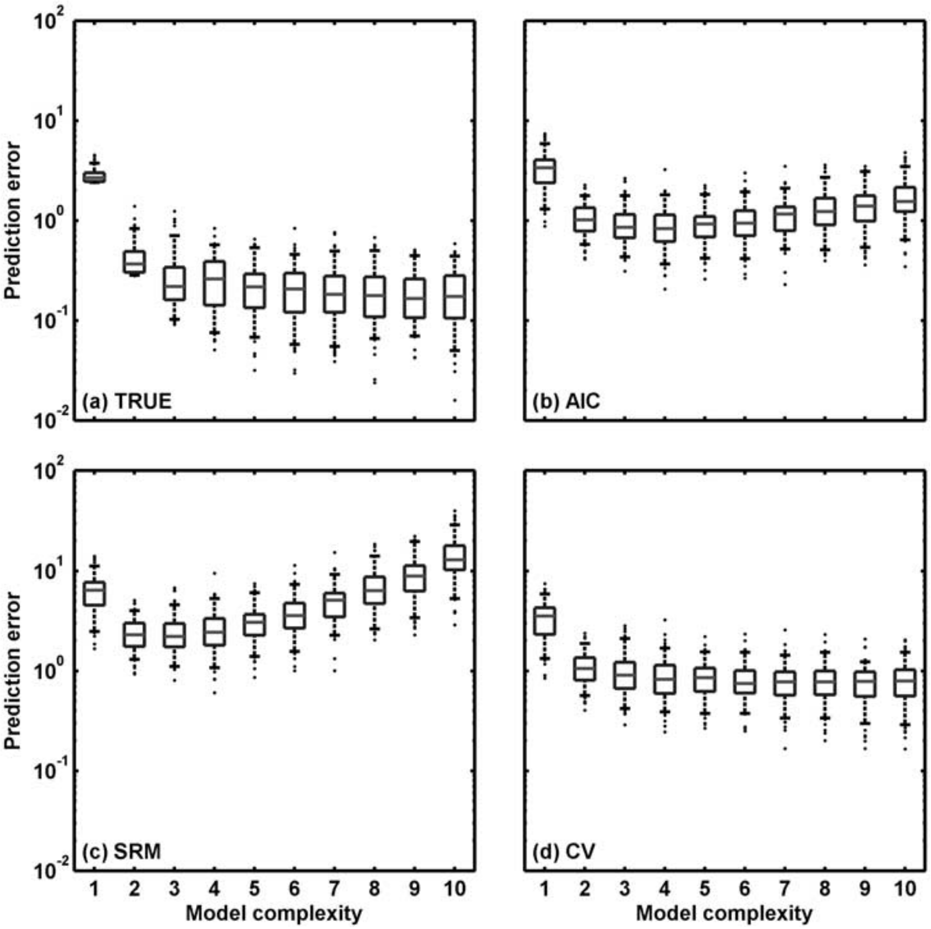

Figure 6. Prediction error as a function of complexity of physically based models: (a) true prediction error given by (2) and (b–d) prediction error estimated with (4) using AIC, SRM, and CV, respectively (n = 20, SNR = 5).

no further improvement is possible (Figure 6a). As opposed to polynomial models (Figure 3a), prediction error does not increase with increasing model complexity. The physical relationship between discharge and water height imposes a smoothing constraint on the "wiggliness" of the approximating function, leading to a pronounced smoothing effect not present when using polynomials. Hence, our results suggest that with physically based models, overfitting does not lead to increased prediction errors, and one may get away with using overly complex models, as long as physical laws (mass balance, constitutive relations) are included. Figure 6 also shows estimates of prediction error with the three model complexity control methods. In accordance with results for polynomials (Figure 3), CV provides the best estimates, whereas AIC and SRM now overestimate prediction error of complex models. Again, SRM penalizes complex models more than does AIC.

[25] Model selection results for physically based models are shown in Figure 7. Prediction error of selected models is similar for all three methods, with SRM yielding models with slightly greater prediction error (Figure 7a). The accuracy with which prediction error of the selected model is estimated is fairly similar to the previous case. (Figure 7b). In addition, Figure 7c shows quite a large range in optimal number of parameters (complexity), with SRM again selecting the least complex models (3 parameters on average), and CV selecting a relatively large range of complex models. These results indicate that model performance and prediction error are fairly insensitive to the number of parameters used, confirming results in Figure 6a that there seems to be no effect of overfitting on prediction error. As opposed to results for polynomials, Figure 7 suggests that more complex models (AIC, CV) result in slightly lower prediction error. Despite the robustness of prediction error with regard to model complexity, we see in Figure 7d that parameter identifiability is much worse for complex models selected by CV, as opposed to simpler ones obtained by SRM, and to a lesser extent, AIC. In other words, overfitting or overparameterization of physically based models does not lead to an increase in prediction error, but causes poor parameter identifiability and large parameter equifinality.

3.3. Case 3: Effect of Data Quantity (Sample Size)

[26] Next we investigate the effect of the amount of data, i.e., sample size n, to infer underlying system behavior. The analysis in the previous section using physically based models is repeated here, but now sample size n is decreased from 20 to 5. As before, model selection is performed with three methods for a total of 100 random data sets.

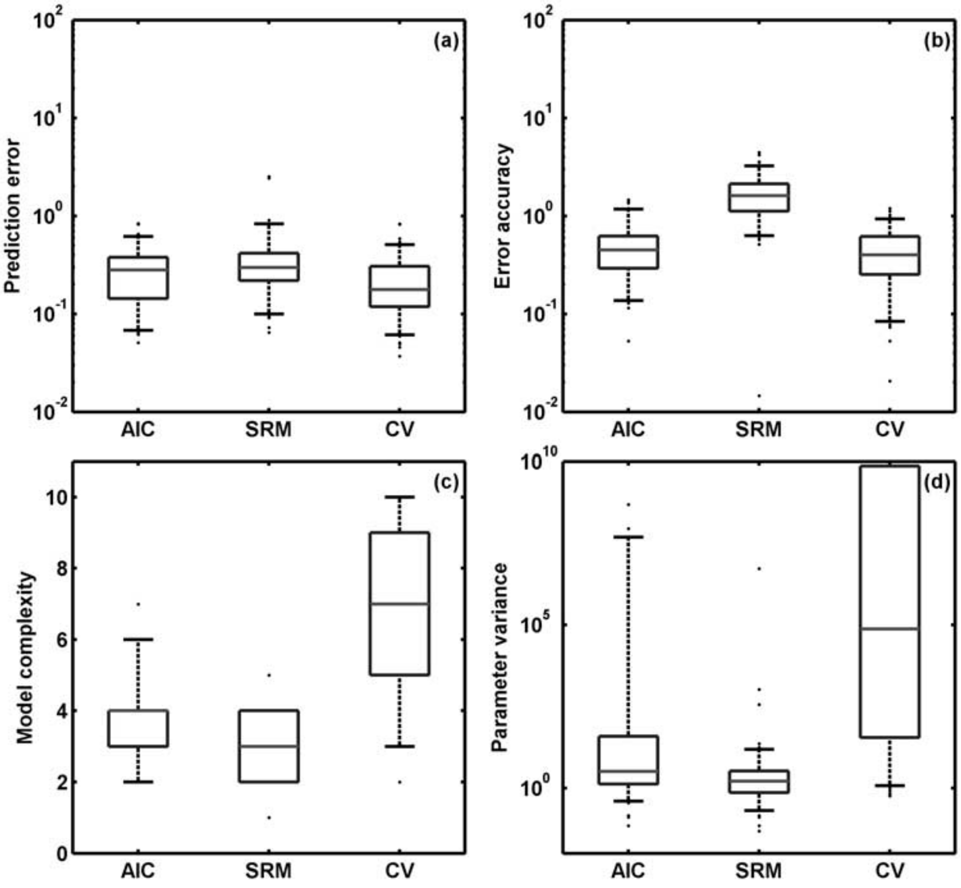

Figure 7. Model selection results for physical models: (a) true prediction error, (b) accuracy of estimated prediction error, (c) model complexity, and (d) parameter identifiability of optimal physically based models according to three model complexity control methods (n = 20, SNR = 5).

[27] Results for n = 5 are presented in Figure 8. Comparison with results for a larger sample size n = 20 in Figure 7 allows an assessment of the effect of sample size on model selection. First of all, prediction error of selected models shows a similar trend as before, with SRM selecting models with somewhat larger prediction error. Also note that prediction error in absolute value is greater in Figure 8a than in Figure 7a, which is to be expected when models have to be inferred from a smaller amount of data. Similarly, accuracy of estimated prediction error follows results in Figure 7. Furthermore, Figure 8c clearly illustrates the effect of sample size on optimal model complexity: when less information is available about the underlying system behavior, less complex models are supported by the data, and this is true for all three methods. In other words, model complexity control reflects the intuitive notion that there should be a balance between the amount of data available and the complexity of the model that can be inferred from it. Finally, note in Figure 8d that with less data the model parameters do not necessarily become less identifiable, as long as model complexity and number of parameters are decreased to reflect the decrease in useful information. As was observed for larger sample size, SRM again yields models with identifiable parameters.

3.4. Case 4: Effect of Data Quality (Signal-to-Noise Ratio)

[28] Apart from data quantity, model selection methods should also account for the effect of data quality. Here, we express quality of data in terms of signal-to-noise ratio (SNR), and investigate its impact on performance of the three model complexity control methods. The analysis from case 2 was repeated with n = 20, but with SNR = 1 instead of 5, indicating a case of decreased data quality.

[29] The results for SNR = 1 are presented in Figure 9. Comparison with results for SNR = 5 in Figure 7, allows an assessment of the effect of data quality on model selection. Decreasing SNR has a similar effect as decreasing sample size: in both cases prediction error of optimal models increases. Furthermore, all three methods select less complex models when SNR is lower and more noise is present in the data. Analytical methods, especially SRM, are more robust to sampling variation, as indicated by robust estimates of optimal model complexity and parameter variance, compared to a much greater sensitivity of CV to sampling variation, as evidenced by box plot heights. Such sensitivity may be due to the relatively flat shape of prediction errors for physically based models (see Figure 6a). It also leads to CV selecting much more complex models than the analytical methods. However, the benefit of a resulting decrease in

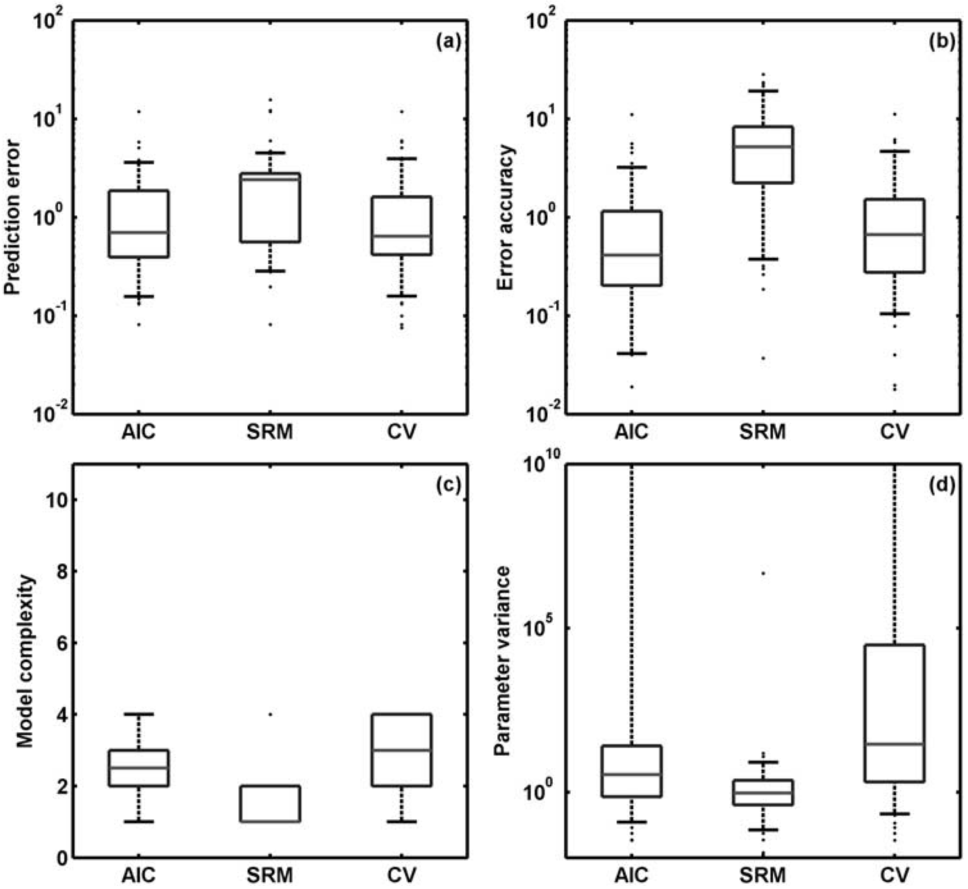

Figure 8. Model selection results for physical models: (a) true prediction error, (b) accuracy of estimated prediction error, (c) model complexity, and (d) parameter identifiability of optimal physically based models according to three model complexity control methods (n = 5, SNR = 5).

prediction error (Figure 9a) is small relative to a significant worsening of parameter identifiability (Figure 9d).

3.5. Case 5: Streamflow Prediction in Leaf River Basin

[30] Rainfall and streamflow data for this case are displayed in Figure 10. We focus on four nonconsecutive streamflow "events," two in winter and two in summer. Simulation is done using a lumped rainfall-runoff model with three parameters, as described in Table 3. Two additional models are considered: one whose parameters may vary by season (6 parameters total), and one whose parameter values may vary by event (12 parameters total). Each model is fitted to the data and both SRM and AIC are used to compute prediction errors according to (4).

[31] Results for this case are summarized in Figure 11 and Table 3. We conclude that temporally variable model parameters, by season or by event, result in increasingly better fits to the data, as shown by R_e values in Table 3. However, models with time-variable parameters exhibit parameter nonuniqueness and equifinality, as evidenced by parameter variances in Table 3. SRM prevents such overparameterization by selecting a model with constant parameters, whereas AIC selects a model that is too complex for accurate parameter identification. These results confirm our findings with the synthetic case studies above, and point to a relative insensitivity of model selection methods to assumptions about data independence, which is clearly violated in this case.

4. Discussion

[32] Next, we discuss some implications of our results for application of model complexity control to hydrologic prediction. Although our case studies are of limited complexity, they provide necessary insights into advantages and limitations of the different methods. Our results also point to key issues for further research.

4.1. Effect of Physical Constraints

[33] Model selection was performed using both polynomials and models based on physical principles. Comparison of results for these two cases suggests the following. First, models based on smooth physical laws, such as the monotonically increasing head-discharge relationship in our case, do not exhibit increased prediction errors as the model complexity or number of parameters increases. This is in contrast to flexible polynomial models, which are prone to overfitting noisy data, resulting in increased prediction error for complex polynomials. Hence, there is a clear benefit to physically constraining the model structure for making predictions. However, although overfitting or overparameterization of physically based models does not lead to an

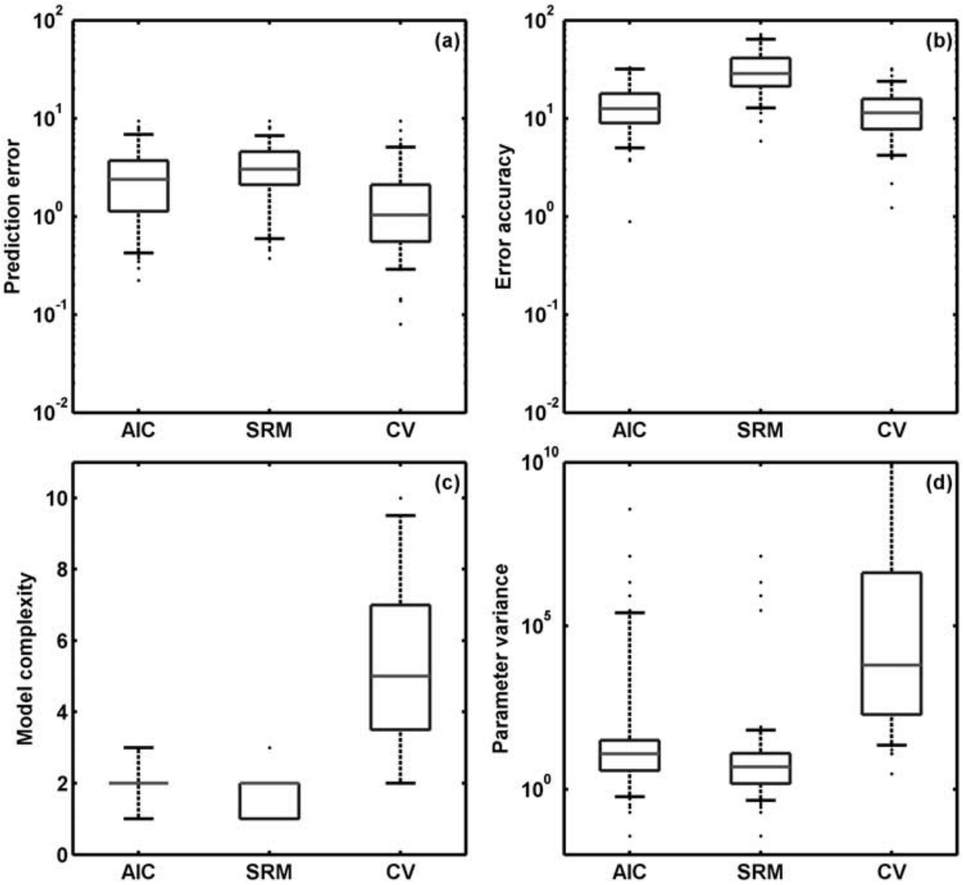

Figure 9. Model selection results for physical models: (a) true prediction error, (b) accuracy of estimated prediction error, (c) model complexity, and (d) parameter identifiability of optimal physically based models according to three model complexity control methods (n = 20, SNR = 1).

increase in prediction error, it causes poor parameter identifiability and large parameter equifinality. In addition, the flatness in prediction error with increasing model complexity constitutes in effect a form of equifinality in suitable model structures, i.e., a wide range of model structures perform similarly well, and one is not penalized for having a very complex model. Model complexity control provides a quantitative framework for resolving these equifinalities by identifying a parsimonious model with identifiable parameters. Alternatively, if for physical reasons model complexity is given, the amount of data needed for accurate parameter identification can be estimated.

[34] A challenge remains in formulating physically based models with flexible model structures that span a range of

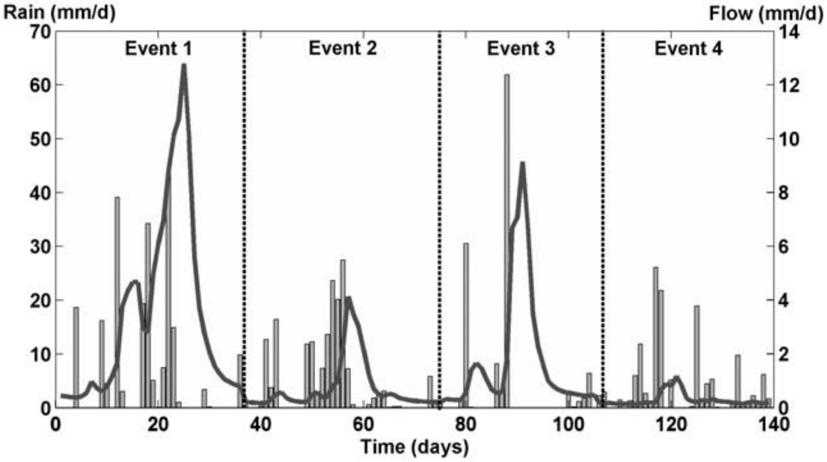

Figure 10. Daily observed rainfall and resulting streamflow for four nonconsecutive "events" in Leaf River basin.

Table 3. Model Selection Results for Leaf River Case Study[a]

Model Parameters	R_e	n	h	p	r_{SRM}	r_{AIC}	R_p SRM	R_p AIC	*PVar*[b]
Constant in time	1.49	139	3	0.02	1.64	1.04	2.45	1.56	1
Seasonally variable	1.41	139	6	0.04	1.91	1.09	2.70	1.54	32
Variable by event	1.22	139	12	0.09	2.43	1.19	2.97	1.45	70

[a] Model complexity h equals number of model parameters. There are three basic parameters: (1) maximum storage capacity in the basin, (2) a parameter describing spatial variability of storage capacity, and (3) a linear reservoir coefficient related to residence time of effective rainfall in the basin. Number of model parameters is doubled or tripled by considering time-variance of these parameters by season (×2) or by event (×4).

[b] Variable *PVar* indicates total parameter variance of each model, normalized relative to variance in the model with time-constant parameters.

model complexities [*Clark et al.*, 2008]. Most obviously this can be done by a downward approach, where additional detail and complexity is added in a stepwise manner [*Sivapalan et al.*, 2003]. Alternatively, one can follow an upward approach and start from a complex model, perhaps on the basis of small-scale physics, combined with some form of regularization [*Tonkin and Doherty*, 2005]. Regularization allows one to introduce additional constraints representing a priori physical knowledge about the system. Model complexity control can be used in this context to identify an optimal value for the regularization parameter, thereby determining optimal weight to be given to a priori physical knowledge.

4.2. Effects of Data Quantity and Quality

[35] Our results indicate that model complexity control is able to account for effects of data quantity and quality, identifying models that have a level of complexity that matches available information. The influence of these effects can clearly be seen in estimated prediction error in (4), which forms the basis for model selection. First, data quality is accounted for in empirical error R_e, which quantifies how well a model can fit data - this fitting ability clearly depends on the level of noise in the data. Second, data quantity is explicitly included in the penalization factor, as seen in (5) and (6), which penalizes complex models on the basis of small sample sizes. Often in hydrology, data are either accurate but sparse, as with point measurements, or ample but less accurate, as with remote sensing. In each case, data quality and quantity will have an effect on the level of model complexity that can be supported.

[36] It should be emphasized that increased sample size will mainly be beneficial if additional data contains independent or orthogonal information about the underlying system behavior. This is reflected in the need to provide the model complexity control methods with n independent data samples (see Table 1). Even though data independence can often not be guaranteed in hydrology, our results for Leaf River basin using rainfall-streamflow time series suggest that model complexity control can still yield useful results. Nevertheless, there is a need to develop and apply methods, such as value-of-information analysis [e.g., *Yokota and Thompson*, 2004], that allow identification of independent data with high information content. Such an analysis in combination with model complexity control could result in a very powerful tool for integrated data and model identification.

4.3. Comparison of Model Complexity Control Methods

[37] Comparing results of three model complexity control methods, we can conclude the following. First, SRM consistently selects the least complex models; that is, it is the most conservative method since it is based on worst case upper bounds on prediction error for small samples, as opposed to AIC which assumes infinitely large sample size and therefore tends to select more complex models. Being conservative intuitively makes sense when data quantity or quality is low, since it attaches less importance to few and noisy data. Our results reflect this notion in that SRM yields models that are better identifiable, thereby avoiding the equifinality problem of complex overparameterized models.

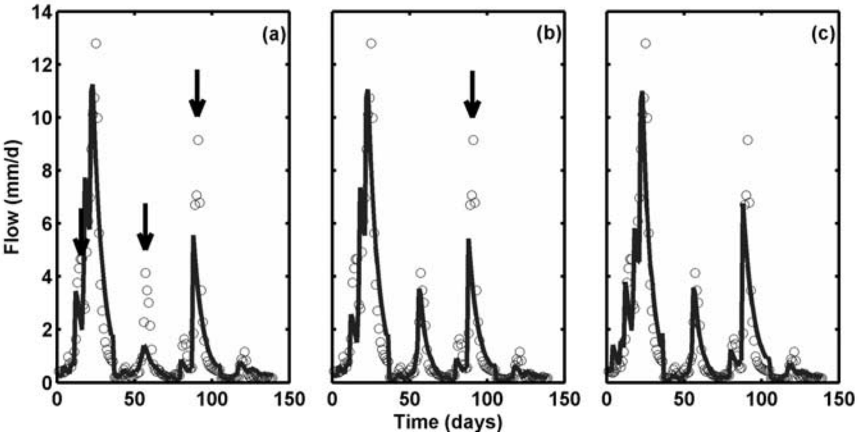

Figure 11. Simulated and observed hydrographs for Leaf River basin using three levels of model complexity: (a) parameters constant in time, (b) seasonally variable parameters, and (c) parameters variable by event. Vertical arrows highlight discrepancies.

Parsimonious models with unique parameters have somewhat larger prediction errors, but only marginally so. Our numerical experiments confirm results from statistical learning theory (SLT) on which SRM is based. SLT theoretically predicts the onset of equifinality by theorem A6.3 of *Vapnik* [1998]. The relevance of this theoretical insight is that it allows us to choose in advance how complex a model should be to avoid equifinality, given available data. On the other hand, if physics dictates model complexity, we can predict the number of data points that need to be collected to arrive at unique parameter values. The predictability of the onset of equifinality as a function of model complexity and data availability is an important advantage of SRM.

[38] Second, results show that CV is more sensitive to sampling variation, whereas analytical methods (AIC and SRM) are more robust. Analytical methods are also preferable to CV in terms of computational demand. A potential drawback of the analytical methods is that they require an estimation of model complexity. For a model linear in its parameters such an estimate can be based on the number of model parameters. However, in the case of nonlinear models estimating model complexity is less straightforward and more empirical. For example, in SRM model complexity is measured by the model's Vapnik-Chernovenkis (VC) dimension, a measure of the model's "wiggliness" or flexibility to fit data points. In our case studies we estimated the VC dimension of the nonlinear physically based models using the number of model parameters. A more rigorous approach would require the VC dimension to be estimated empirically [*Shao et al.*, 2000]. This is still an open research field.

[39] Finally, it should be mentioned that there are also methods that automatically identify the best model complexity control method, i.e., a metaselection method, as presented by *De Luna and Skouras* [2003]. The authors demonstrated their approach for estimating order of time series models, focusing on a comparison between AIC and BIC performance.

5. Conclusions

[40] Model complexity control provides a systematic approach to balance complexity of a hydrologic model with quantity and quality of available information. There are several advantages associated with such an approach to hydrologic modeling, one being the reduction of parameter equifinality, in the sense that the approach leads to models with identifiable parameters on the basis of available data. Performance of three model complexity control methods was compared using a simple flow problem for which the exact physics are known, and a more comprehensive rainfall-runoff example. This allowed a systematic evaluation of the effects of various assumptions that underlie these methods. We reach the following conclusions in this paper:

[41] 1. Smoothness of physically based hydrologic parameterizations (e.g., storage flux laws) effectively bounds hydrologic prediction uncertainty, even when models become overparameterized and parameter values cannot be identified from available data.

[42] 2. Model complexity control methods present a formal and objective way of identifying the simplest model supported by the data (Occam's razor), as a function of both quantity and quality of available data. Small and noisy data sets support less complex models than large and accurate data sets. Hence, model complexity control quantifies an intuitive link between model complexity and data availability.

[43] 3. A comparison of various model complexity control methods shows that structural risk minimization (SRM) is preferable to other methods, as it consistently identifies parsimonious models with unique parameters, even in cases where the assumption of data independence is violated. Methods developed for application with large data sets, such as AIC, tend to select models that are too complex, and which suffer from parameter equifinality. Cross validation is computationally intensive and sensitive to sampling variation.

[44] On the basis of these preliminary results from synthetic and practical case studies, we conclude that model complexity control holds promise as a systematic, quantitative, and statistically sound methodology for dealing with issues of parameter and model equifinality in hydrology. Future work should focus on applying analytical model complexity control methods to complex hydrologic case studies. A rigorous application of SRM in this context will require an accurate estimation of the VC dimension as a complexity measure of hydrologic models. A promising research direction also lies in combining model complexity control with value-of-information tools that identify orthogonal data with large information content.

[45] **Acknowledgments.** We would like to thank the associate editor, Alberto Montanari, and three anonymous reviewers for their critical and constructive comments.

References

Akaike, H. (1970), Statistical predictor identification, *Ann. Inst. Stat. Math.*, *22*, 203–217, doi:10.1007/BF02506337.

Beven, K. (1993), Prophecy, reality and uncertainty in distributed hydrological modeling, *Adv. Water Resour.*, *16*, 41–51, doi:10.1016/0309-1708(93)90028-E.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36.

Cherkassky, V. (2002), Model complexity control and statistical learning theory, *Nat. Comput.*, *1*(1), 109–133, doi:10.1023/A:1015007927558.

Cherkassky, V., and F. Mulier (2007), *Learning from Data: Concepts, Theory, and Methods*, 2nd ed., 538 pp., John Wiley, New York.

Cherkassky, V., X. Shao, F. Mulier, and V. Vapnik (1999), Model complexity control for regression using VC generalization bounds, *IEEE Trans. Neural Networks*, *10*(5), 1075–1089, doi:10.1109/72.788648.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

De Luna, X., and K. Skouras (2003), Choosing a model selection strategy, *Scand. J. Stat.*, *30*, 113–128, doi:10.1111/1467-9469.00321.

De Ridder, F., R. Pintelon, J. Schoukens, and D. P. Gillikin (2005), Modified AIC and MDL model selection criteria for short data records, *IEEE Trans. Instrum. Meas.*, *54*, 144–150, doi:10.1109/TIM.2004.838132.

De Wit, M. J. M., and E. J. Pebesma (2001), Nutrient fluxes at the river basin scale. II: The balance between data availability and model complexity, *Hydrol. Processes*, *15*, 761–775, doi:10.1002/hyp.176.

Dibike, B. Y., S. Velickov, D. Solomatine, and M. B. Abbott (2001), Model induction with support vector machines: Introduction and applications, *J. Comput. Civ. Eng.*, *15*, 208–216, doi:10.1061/(ASCE)0887-3801(2001)15:3(208).

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*, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008), Understanding catchment behavior through stepwise model concept improvement, *Water Resour. Res.*, *44*, W01402, doi:10.1029/2006WR005563.

Freeze, R. A., and R. L. Harlan (1969), Blueprint for a physically-based digitally simulated, hydrologic response model, *J. Hydrol.*, *9*, 237–258, doi:10.1016/0022-1694(69)90020-1.

Gregory, J. M., T. M. L. Wigley, and P. D. Jones (1992), Determining and interpreting the order of a two-state Markov chain: Application to models of daily precipitation, *Water Resour. Res.*, *28*, 1443–1446, doi:10.1029/92WR00477.

Honjo, Y., and N. Kashiwagi (1999), Matching objective and subjective information in groundwater inverse analysis by Akaike's Bayesian information criterion, *Water Resour. Res.*, *35*, 435–447, doi:10.1029/98WR02365.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*, 2637–2649, doi:10.1029/93WR00877.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resour. Res.*, *43*, W03411, doi:10.1029/2006WR005195.

Khalil, A. F., M. N. Almasri, M. McKee, and J. J. Kaluarachchi (2005), Applicability of statistical learning algorithms in groundwater quality modeling, *Water Resour. Res.*, *41*, W05010, doi:10.1029/2004WR003608.

Klemeš, V. (1986), Dilettantism in hydrology: Transition or destiny?, *Water Resour. Res.*, *22*, 177S–188S, doi:10.1029/WR022i09Sp0177S.

Knotters, M., and J. G. De Gooijer (1999), TARSO modeling of water table depths, *Water Resour. Res.*, *35*, 695–705, doi:10.1029/1998WR900049.

Loague, L., and J. E. VanderKwaak (2004), Physics-based hydrologic response simulation: Platinum bridge, 1958 Edsel, or useful tool, *Hydrol. Processes*, *18*, 2949–2956, doi:10.1002/hyp.5737.

Mutua, F. M. (1994), The use of the Akaike information criterion in the identification of an optimum flood frequency model, *Hydrol. Sci. J.*, *39*, 235–244.

Perrin, C., C. Michel, and V. Andreassian (2001), Does a large number of parameters enhance model performance: Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*, 275–301, doi:10.1016/S0022-1694(00)00393-0.

Savenije, H. H. G. (2001), Equifinality, a blessing in disguise?, *Hydrol. Processes*, *15*, 2835–2838, doi:10.1002/hyp.494.

Schwartz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464, doi:10.1214/aos/1176344136.

Shao, X., V. Cherkassky, and W. Li (2000), Measuring the VC-dimension using optimized experimental design, *Neural Comput.*, *12*, 1969–1986, doi:10.1162/089976600300015222.

Sivapalan, M., G. Bloschl, L. Zhang, and R. Vertessy (2003), Downward approach to hydrological prediction, *Hydrol. Processes*, *17*, 2101–2111, doi:10.1002/hyp.1425.

Stomph, T. J., N. de Ridder, and N. van de Giesen (2002), A flow-meter for low discharges from laboratory flumes, *Trans. ASAE*, *45*, 345–349.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems to identify model parameters and structure, *Hydrol. Sci. J.*, *44*, 779–798.

van der Perk, M. (1997), Effect of model structure on the accuracy and uncertainty of results from water quality models, *Hydrol. Processes*, *11*, 227–239, doi:10.1002/(SICI)1099-1085(19970315)11:3<227::AID-HYP440>3.0.CO;2-#.

Vapnik, V. (1998), *Statistical Learning Theory: Inference From Small Samples*, John Wiley, New York.

Wagener, T., M. J. Lees, and H. S. Wheater (2001), A toolkit for the development and application of parsimonious hydrological models, in *Mathematical Models of Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, pp. 91–140, Water Resour. Publ., Highlands Ranch, Colo.

Wagener, T., M. Sivapalan, P. A. Troch, and R. Woods (2007), Catchment classification and hydrologic similarity, *Geogr. Compass*, *1*, 901–931, doi:10.1111/j.1749-8198.2007.00039.x.

Yokota, F., and K. Thompson (2004), Value of information literature analysis: A review of applications in health risk assessment, *Med. Decision Making*, *24*, 287–298, doi:10.1177/0272989X04263157.

Young, P. C. (2003), Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale, *Hydrol. Processes*, *17*, 2195–2217, doi:10.1002/hyp.1328.

H. H. G. Savenije, G. Schoups, and N. C. van de Giesen, Department of Water Management, Delft University of Technology, Stevinweg 1, P.O. Box 5048, NL-2600 GA Delft, Netherlands. (g.h.w.schoups@tudelft.nl)