International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Complexity regularized hydrological model selection

Saket Pande[1], Liselot Arkesteijn[1], Luis A. Bastidas[2]

[1]Department of water management, Delft University of Technology, Delft, Netherlands, [2]ENERCON Services Inc., Pittsburgh Office, Murrysville PA, USA
(s.pande@tudelft.nl)

Abstract: This paper uses a recently proposed measure of hydrological model complexity in a model selection exercise. It demonstrates that a robust hydrological model is selected by penalizing model complexity while maximizing a model performance measure. This especially holds when limited data is available. Here by a robust model, we mean a model that predicts a variable of interest conditioned on future input forcing better than a model that is fitted on limited data. We demonstrate this on a rainfall-runoff model structure, SAC-SMA, using MOPEX data set of Guadalupe river basin.

Keywords: Hydrologic complexity; robust model selection; complexity regularization.

1 INTRODUCTION

It is widely accepted that more complex models tend to overfit data especially when data, in terms of its information content, is scarce (Gupta and Sorooshian, 1983; Vapnik, 2002; Pande et al, 2009; Renard et al., 2010; Arkesteijn and Pande, 2013; Pande et al, 2012). As a result, predictions on future forcing data are often inaccurate. This is not to suggest that one should always select a model of minimum complexity. A model should be selected in a manner that trades off model complexity with its performance on a given sample of data and the tradeoff should be such that the selected model's performance on future unseen data is amongst the best. If a hydrological model can be selected in such a manner, it will be a choice that is robust in predicting future unseen data. Numerous studies have therefore focussed on considering model complexity in model selection problems (Schwarz, 1978; Jakeman and Hornberger, 1993; Young et al., 1996; Cavanaugh and Neath, 1999; Pande, 2005; Ye et al., 2008; Gelman et al., 2008; Schoups et al, 2008; Pande et al, 2009; Clement, 2011; Arkesteijn and Pande, 2013; Pande et al, 2012).

This paper utilizes a measure of complexity recently proposed by Arkesteijn and Pande (2013) and uses it to penalize hydrological model selection. The measure of complexity is based on bounding the behaviour of how a model simulation deviates from its expected value (estimated on many realizations of input forcing) at any time step. A rationale for such an approach is that a more complex model should have larger such deviation than a less complex model. This statement therefore provides a definition of what we mean by model complexity. Arkesteijn and Pande (2013) showed that such a deviation of any hydrological model for a given sample size can be quantified, thereby measuring its complexity. They also showed that the bounds can also be used to bind the rates of convergence of model performances, measured by Mean Absolute Errors, to its expected values. This statement also provides a framework within which model complexity, as per our definition, and sample size bind predictive uncertainty. As a result, an expression for complexity regularized robust model selection can be obtained such that model performance on a given data set trades off with its complexity and the size of available data.

In comparison, Bayesian approaches to complexity regularized model selection trade off the log likelihood value with parameter dimensionality or trade off the log likelihood value with the determinant of its Hessian at the parameter value that maximizes the likelihood (Ye et al, 2008; Marshall et al, 2005). The implicit measure of complexity in such approaches is the Hessian of the log-likelihood function at the optimum. However, unlike the approach presented here, it is application specific (in particular with respect to the variable of prediction interest) since the Hessian of the log-likelihood value at the optimum needs to be estimated. An approach that appears similar to the one

presented here is ROPE that is based on the concept of data depth functions (Bardossy and Singh, 2008). Both the approaches share the philosophy of interpreting robustness geometrically. However, the two approaches appear to diverge in the use of data of the variable of prediction interest. Also, ROPE is conditioned on the model structure that is being used. What this means to comparing 'robust' performances of models selected from two different model structures, in comparison to the approach presented here, remains to be explored.

Two hydrological model structures, SAC-SMA and SIXPAR, are considered. Their complexities are measured and compared. Then the complexity regularized model selection algorithm of Arkesteijn and Pande (2013) is used on SAC-SMA model structure for predicting runoff in the Guadalupe river basin, USA. The paper provides evidence that complexity regularized SAC-SMA models perform better on unseen data than when SAC-SMA models are selected without controlling for their complexity. This also provides evidence that complexity regularized model selection as presented here controls overfitting.

2 METHODOLOGY

2.1 Model Structures and Data set

The two model structures that are considered are SAC-SMA and SIXPAR. SAC-SMA is a complex model structure with two layer reservoir architecture and a nonlinear percolation conceptualization. The two upper zone reservoirs represent a free water zone and a tension water zone, wherein the former controls the percolation to the lower zones while the tension water zone mainly controls the evaporation and feeds the free water zone. The percolation is a nonlinear complex function of demand from the lower reservoirs and available supply of water from the upper zone reservoirs. Both the upper and lower zones also control the flows. The SIXPAR model structure, which is a conceptual simplification of the SAC-SMA model with one upper and lower zone, excludes evaporation and the concept of tension water zones but retains the complex conceptualization of percolation. Additional details on the models can be found elsewhere (Burnash, 1995; Duan et al, 1992; Arkesteijn and Pande, 2013).

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

Table 1. Parameter ranges for SAC-SMA model structure.

Table 1 provide the ranges of the parameters used for SAC-SMA model structure. The parameters for SIXPAR model structure are obtained from SAC-SMA parameters following a routine called *TRANS*.

The role of *TRANS* transformation is to ensure that the parameter sets of SIXPAR are 'equivalent' to SAC-SMA so that difference in complexity between the two models is only due to differences in the two structures (Arkesteijn and Pande, 2013). The complexity algorithm, used in step 2 of Algorithm 1 is then applied.

TRANS (transformation routine):
The upper and lower zone storage capacity of SIXPAR is the sum of respective upper and lower zone storage capacities of SAC-SMA model structure.

UZ = UZTWM + UZFWM
BM = LZTWM + LZFSM + LZFPM

The upper and lower zone recession parameters are the geometric mean of upper zone recession parameters and the geometric mean of the lower zone recession parameters.

UK = UZK
BK = $\sqrt{\text{LZSK} * \text{LZPK}}$

The percolation parameters off both the model structures are kept the same.
z = zperc
x = rexp.

The data used is the Guadalupe River Basin in United States from the MOPEX data set (Duan et al, 2006). Daily potential evaporation, rainfall and streamflow data for a period of 1948-1970 is used. The models used are forced using max(P-PE,0) where P (mm/day) is the daily precipitation and PE (mm/day) is daily potential evaporation. Daily streamflow is then used to select complexity regularized or unregularized SAC-SMA models.

2.2 Algorithms

The following algorithm samples parameters for a model structure, estimates its complexity based on Algorithm 2 of Arkesteijn and Pande (2013) and uses it to select a model from the model structure such that its complexity is regularized.

Algorithm 1 (Complexity Regularized Model Selection):
1. Sample *P* parameter sets for SAC-SMA. A model corresponding to each such parameter is thus obtained.
2. For 10000 values of c between (10^{-3}, 10^{3}) on a logarithmic scale, calculate $T_1 = \xi_N + c\sqrt{F(h,N)}$ where ξ_N is mean absolute error estimated on a data set D of size N (a measure of model performance) and $F(h,N)$ is the measure of complexity obtained for a model using Algorithm 2 of Arkesteijn and Pande (2013).
3. For each c, determine the minimum of T_1 over the P different parameter sets (and hence models). The minimum of T_1 yields an optimal parameter set.
4. Calculate $T_2 = \xi_{N'}$ (mean absolute error) for each optimal set corresponding to each value of c obtained in step 3 on another data set D', independent of D, of length N'.
5. Minimize T_2 this time over different values of c and denote by θ_N^* and c_N^* the parameter set and c corresponding to the minimum thus obtained.
6. Calculate $T_2 = \xi_{i,N''}$ for parameter set θ_N^* on a third independent data set D" of length N".

In step 2 of the algorithm, the function of complexity has the form $F(h,N) = \beta_2 + \frac{\beta_1}{N} + \frac{\beta_0}{N}$ where *h* then represents the set $\{\beta_2, \beta_1, \beta_0\}$. Step 5 of the algorithm produces the parameters of a model that is selected by regularizing its complexity. For this, a complexity regularized risk function T_1 defined in step 2 is minimized. Since the tradeoff parameter c defines how mean absolute error trades off with the complexity measure, it is first estimated on an independent data set before the complexity regularized risk function T_1 is put to use. Step 6 of the algorithm validates the performance of complexity regularized SAC-SMA model that can then be compared with a SAC-SMA model that is selected without complexity regularization.

3 RESULTS

The Algorithm1 is implemented for *P* = 500 parameters of SAC-SMA sampled from the parameter ranges specified in Table 1. Three different lengths of D are considered, N = 1/6, 1/3, and ½ year, to simulate small sample sizes available for model selection. The lengths of D' and D" are N' = N" = 5 years. All are independent data sets. In order to demonstrate that complexity algorithm of Arkesteijn

and Pande (2013) detects a more complex model structure from a competing set of structures, we use a *TRANS* transformation on SIXPAR model structure that is a conceptual simplification of SAC-SMA. The *TRANS* routine has been described in section 2.1.

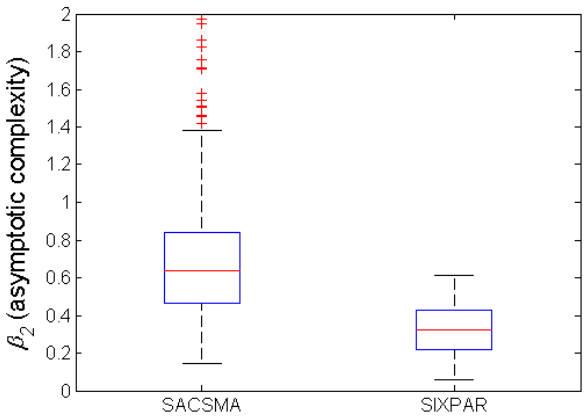

Figure 1. Asymptotic complexities (value of F(h,N) for N sufficiently large or β_2) of SACSMA and SIXPAR. 500 parameters from the ranges specified in Table 1 are used to estimate SACSMA complexity. *TRANS* routine is then used to sample corresponding parameters for SIXPAR and its complexity is then estimated.

Figure 1 demonstrates that when the effect of hydrological parameter magnitudes (Pande et al, 2012; Arkesteijn and Pande, 2013) is controlled by using the *TRANS* routine, the complexity algorithm of Arkesteijn and Pande (2013) finds that SAC-SMA is 'structurally' more complex than SIXPAR model structure.

We now provide evidence of how complexity regularized model selection yields robust model selection that avoids overfitting especially on small sample sizes. In order to do so, we first measure the 'robustness' of regularized and unregularized model selection, where the selection is done out of *P* sampled parameters sets (see step 1 of Algorithm 1). For example, an unregularized SAC-SMA model is the one out of *P* (=500) sampled models that minimizes the Mean Absolute Error (MAE) on data set D of size N.

The robustness of a model selection method is measured by comparing the performance of regularized (or unregularized) model with the performance of the remaining *P*-1 models on independent data sets. The role of independent data sets is to mimic future unseen data. A less robust model selection is expected to yield a model that performs more often poorer than remaining *P*-1 models on independent data sets. A comparison of the robustness of complexity regularized versus non-regularized SAC-SMA model selection then follows.

Figure 2B shows the distribution of the difference in MAE between SAC-SMA model corresponding to $\tilde{\theta}_N$ and models corresponding to the remaining 499 parameter sets. Negative values imply that MAE of unregularized SAC-SMA model is smaller than another SAC-SMA (from amongst the remaining 499 parameters). We observe signs of overfitting at small values of *N* (1/6 and 1/3 year) even amongst the few competing models (only *P* = 500 models corresponding to the sampled parameters) that are considered as can be seen from the mass of the box plots lying to the right side of 0. Nonetheless N = ½ year appears to be sufficiently large for SAC-SMA to model the Guadalupe river basin since the $\tilde{\theta}_N$ model performs better than the remaining 499 models.

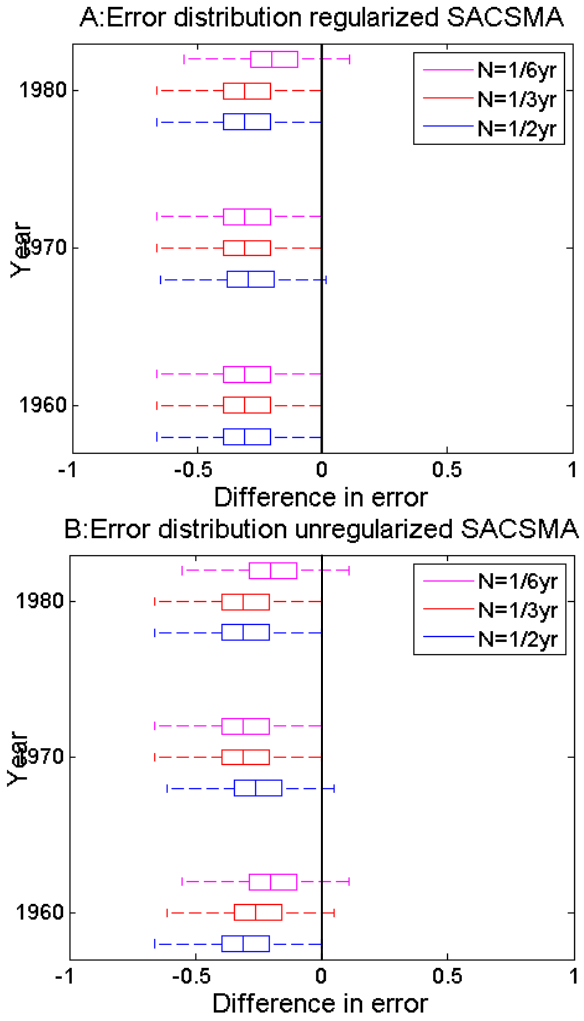

Figure 2. Distribution of the difference between the performance (measured by Mean Absolute Error) of SACSMA models corresponding to A) complexity regularized optimal parameter set θ_N^* and all 499 other parameter sets and B) unregularized optimal parameter set $\tilde{\theta}_N$ and all 499 other parameter sets on 5 year test sets D" starting from years specified along the y–axis.

Meanwhile Figure 2A shows the performance of complexity regularized SAC-SMA model, represented by θ_N^*, in terms of its difference from the remaining 499 models. The same P = 500 sampled parameters are used as in Figure 5B. In comparison with Figure 5B, regularized SAC-SMA model is a more robust choice (from amongst other competing 499 models) for predicting daily streamflow on 1960-1964 dataset when the model is selected on small sample sizes of 1/6 and 1/3 year. This can be seen from less mass of the boxplots lying to the right of 0 when compared with the performance of unregularized SAC-SMA performance for 1960-1964. The performance of complexity regularized SAC-SMA, when selected at sample sizes 1/6 and 1/3 year, is almost always smaller than other 499 models when evaluated on all 3 independent 5 year data sets.

However the performance of regularized SAC-SMA is not significantly different from the performance of unregularized SAC-SMA if all 3 independent data sets are considered. It appears that unregularized SAC-SMA model selection selects a model that is almost as close in its representation of the underlying processes as the model selected by complexity regularized model selection. One possible reason may be that SAC-SMA has a model structure that is well constrained and not prone to overfitting. Yet another reason may be that we only sampled P = 500 parameter sets. A more appropriate parameter set may be revealed if an exhaustive parameter sampling is undertaken.

4 CONCLUSIONS

A complexity regularized model selection approach was presented. Complexity of SAC-SMA model structure was quantified based on Arkesteijn and Pande (2013) and used to demonstrate that SAC-SMA is (indeed) structurally more complex than SIXPAR when the effect of the magnitude of parameters on model complexity is controlled for. Evidence was also provided that complexity regularized selection avoids overfitting of models to small data sizes. This demonstrates the utility of complexity regularized model selection, which can be extended to transferability of models in space, not just in time as was demonstrated here.

The difference in the performance of regularized versus unregularized SAC-SMA was limited possibly either due to the nature of SAC-SMA model structure (the structure is perhaps well constrained) or due to limited number of sampled parameters (P=500). Limited parameters were sampled due to the computational burden of computing complexity. A future study is envisaged where the computational burden of computing complexity will be minimized so that the number of parameter samples can be increased. Further in depth investigations into the concept of model complexity presented here are also envisaged.

REFERENCES

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

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., 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

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

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.

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

Marshall, L., D. Nott, and A. Sharma (2005). Hydrological model selection: A Bayesian alternative, Water Resour. Res., 41, W10422, doi:10.1029/2004WR003719.

Pande, S. (2005). "Generalized Local Learning in Water Resources Management." Unpublished Ph.D. dissertation, Department of Civil and Environmental Engineering, Utah State University.

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.

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

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.

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

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

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

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.