Anonymous Referee #2
Received and published: 18 July 2012

1. The paper of Ye et al. applies different models to simulate the regime curves of 197 catchments located in CONUS. The model development process follows a top-down approach, where model complexity is introduced in response to model failures. The objective of the paper is to associate different model structures to different catchments, and investigate and interpret the regional patterns that may arise.

We appreciate the reviewer's comments and suggestions. We have made our efforts to address these comments by cutting figures, shortening the manuscript, and adding the uncertainty estimation. We hope this will be adequate.

2. I am favourable to this work and I think that the research topic is interesting. However, I have the feeling that the work could be much better structured and refined, and that the Authors should spend some more time to reorganize their material. The paper is lengthy, 14 figures are too many for a scientific paper, and the presentation is at times chaotic. Reading this paper and browsing through the other 3 papers of the series, I recognize that the Authors have gone through an amazing body of work, which I fully respect. However, I had the feeling that the Authors did not want to discard anything of the work they had done, and that the process of synthesis and refinement has been overlooked.

We thank the reviewer for their comments and have removed figure 3 and figure 13a. Additionally, we have trimmed the manuscript (Sect. 2.2 and Sect. 3 have been merged) and hope the reviewer will find these changes satisfactory.

3. Regarding the methodology of the paper, the Authors have used regime curves for model calibration. Regime curves synthesize some aspect of the catchment response, at the expense of a loss of information. The question is why using data aggregates that are less informative than the data themselves. Indeed the Authors themselves state in section 4.6 that "a model focused on predicting the regime curves only cannot be expected to predict well the high and low flows". In my opinion, the Authors should have used the data series as they are for model calibration, and then evaluated the models independently on regime curves and flow duration curves.

We chose the regime curves in lieu of the raw data series for calibration based on our purpose for developing the model. The goal of this work is to explore runoff regime behavior, such as seasonal variation among 197 catchments across the continent. Instead of trying to predict the flow series by focusing on the detailed processes that define it, we are more interested in the holistic signatures of catchment response. We agree that calibrating on the data series would be the choice for most of the hydrological modeling exploring catchment characteristics as well as the flow response mechanisms in detail (which we term the "bottom-up" approach in this paper); this analysis aspires to crack the problem from the other side, using the top-down approach. It is an exercise in comparative hydrology, aiming for general understanding of first order impacts of different processes on flow generation mechanisms along climatic or other gradients. The

simulations of models with different combinations of processes were compared among all 197 catchments to present regional patterns of dominant processes. Since our motivation is first order effects, regime curves can provide sufficient information for this study. To keep it simple and robust, we use the regime curves over the long records of data series for calibration. We hope the reviewer finds this explanation clear and solid.

4. The mapping of model structure to catchments is unclear. How was this done? Which metrics have been used? I suspect the most complex model was fitting best all catchments. What made them prefer a lower complexity model for certain catchments?

The mapping of model structure was carried out in two directions: forward and backward, based on the AIC value, following the statistical model selection steps. For the forward selection, we started from a base model, determined which single process helped reduce the AIC (or, improve the model performance) most by adding one process at a time. The chosen process (which minimized the AIC) was regarded as the most dominant process. For the backward selection process, we started from the full model, removing processes one by one until the AIC value could not be reduced any further. The remaining processes were considered to represent the minimum complexity the model could endure.

We agree with the reviewer that the most complex model provided the best fit over all catchments, but indeed, as we can see from the results, not all of the four processes are necessary for all catchments. For example, the snow component was never invoked in the warm catchments and phenology displayed limited influence in southern catchments where temperature is always high. Therefore, we use the backward selection approach to eliminate unnecessary features from our simple model (minimize the complexity) . These remaining features were then used for the process class mapping.

5. Absence of validation. I think the Authors should split the data between calibration and validation, and see if results hold.

We agree with the reviewer that separated datasets for calibration and validation to quantitatively evaluate the model performance is necessary for predictive models. However, our goal is not to deliver precise predictions of the streamflow time series, but rather, to gain a general understanding of first order impacts of different processes on flow generation mechanisms along the climatic or other gradient. For this reason, a qualitative validation, also called "scientific validation" (Biondi et al, 2012) suits the goal of our work better.

One goal of scientific validation is the assessment of model hypotheses: the identification of integral processes for which the model should account. This was proved in the model development section: we initially applied the base model to the nine selected catchments, assessed the model performance, and then added four processes one by one based on catchment characteristics to improve the model's predictions. This systematic model development procedure itself helps to validate the importance of each remaining process. The other goal of scientific validation is to "provide the proof of model adequacy to the

representation of real world". As a model could produce good results with a wide range of specific parameter values, it is important to consider the parameter set as a combined set (Freer et al., 1996). The Bayesian framework we used is able to find optimum parameter sets by giving greater weight to the better simulations. These parameter sets and predictions then can be chosen as more likely than others. In addition to the assessment of model hypotheses and parameters, a multi-criteria approach can also be used to verify model performance. In this work, we calibrate the parameters to optimize both the fast flow and slow flow simultaneously. This multi-objective check helps provide information regarding whether individual subsystems or processes are performed in the catchments. For example, some processes may not affect the total discharge, but could influence the quantities of observed fast flow (Fig. 7 and 8). This multi-objective calibration enables us to detect those improvements in model performance that negatively affect the global discharge but are beneficial for characterizing the fast flow component and detecting the main control processes.

6. No presentation of uncertainty estimates. As the Authors have used Bayesian methods, they could present uncertainty estimates of model parameters and predictions.

We agree with the reviewer that uncertainty analysis is necessary and helpful - we have conducted it as follows: given the best fit parameter set for each catchment, the minimum, mean, maximum and standard deviation values for each parameter present the distribution across catchments (these best fit sets). The upper and lower bounds are defined from the plot of likelihood and parameter values. For each catchment, along the MCMC sampling, there is a chain of likelihood values which are added up from the value of smallest parameter value, the upper and bottom bounds are then defined when the sum of the likelihood values just exceeds 5% and 95% of the total. The relative error is calculated as half of the range between the upper and lower bounds as a percentage of the parameter with the maximum likelihood value. Median relative error is the median level of the uncertainty among the catchments.

	S_{b1} (mm)	t_w (days)	α	S_e (mm)	t_u(days)	S_{b2} (mm)	t_c (days)
Minimum	0.001	0.013	0.000	0.037	1.548	4.184	0.073
Mean	0.069	0.189	0.274	49.756	187.987	326.358	1.538
Maximum	1.013	0.533	0.300	339.181	1301.191	879.561	9.659
SD	0.14	0.09	0.14	69.44	221.68	183.98	1.51
Median Rel. Error (%)	33.57	33.31	23.74	46.73	24.05	11.54	29.19

7. Introduction: I think the authors should state clearly that the focus of this paper is not the flow duration curve, but the regime curve. The first sentence of the introduction is misleading. The FDC and regime curves are 2 different signatures, and cannot be transformed one in another.

We are sorry about the confusion the first sentence caused. This paper, indeed, focuses on regime curves rather than flow duration curves. We have revised the beginning as follows in the subsequent paragraph. We hope these alterations clarify the confusion:

This paper is the second paper of a 4-part series (the others being Cheng et al., 2012; Coopersmith et al., 2012; and Yaeger et al., 2012) that attempt to understand the physical controls on regional patterns of variations within hydrological signatures of runoff variability. Instead of exploring the Flow Duration Curve (FDC, a key frequency-based signature of daily runoff variability) like the first paper, we will approach the issue from a different perspective, focusing on another compact signature of runoff variability, namely, the regime curve, which denotes the mean seasonal variation of within-year runoff variability.

8. Methodology: the Authors distinguish between "satisfactory" and "non-satisfactory" models based on an acceptance threshold, which is MSE=0.53. Catchments with performance < .53 where left out of the analysis. Clearly this threshold is quite important in determining the outcomes of this study. How was it determined? Similarly, in motivating model improvements, the Authors refer to "non satisfactory" model performance. How was this assessed? Where similar threshold adopted? More generally, I think these types of "absolute" thresholds are quite dangerous, because model performance can be affected by many aspects other than model structure, such as data uncertainty. The Authors should think of a better way of motivating model improvement. Perhaps the performance relative to the most complex model could be a better alternative.

We agree with the reviewer that it is sudden to give 0.53 without an explanation; we have now described how we arrived at it as follows. We hope the reviewer find it adequate.

The initial screening of the model's simulations suggested that even the complete model was insufficient in certain catchments, for example, those in the Midwest, where human impacts cannot be ignored. In some catchments, the flow regime curves were bimodal while the model can only capture one of the flow peaks. As a simple model, we would not expect that it could accommodate the anthropogenic activities; therefore, we need to eliminate these catchments where the model performs poorly. To ensure that the model captures the dynamics as well as the volume of the flow, we use MSE as our criterion. The decomposition of the MSE (or Nash-Sutcliffe efficiency) shows that the MSE consists of three components: mean, variance and correlation coefficient (Gupta et al., 2009). However, as the error is scaled by the standard deviation, it could cause trouble in comparisons among catchments. To avoid this, we standardized the flow before the MSE calculation. We selected the 90% of the catchments with lowest MSE in fast flow, slow flow and total flow separately and then obtained the intersection of these three sets to determine those catchments that had the lowest MSE in fast flow, slow flow and total flow simulation. These catchments were then considered as satisfactory catchments. The 0.53 value was then turned out as all these satisfactory catchments had MSE less than 0.53.

9. Model description: note that equations 5, 8 and, line 7 of page 7044 are dimensionally wrong (they equal storages to fluxes).

We appreciate the reviewer's correction, and have fixed it:
Equation 5: all the fluxes and storages are normalized by the area, the units are L/T and L respectively; we think this one is correct.

Equation 8: $Q_{2f} = \frac{S_2 - S_{b2}}{\Delta t}$

Line 7 of page 7044: $Q_{1f} = (S_1 - S_{b1})/\Delta t$

10. Model description: please mention the numerical methods used to solve model equations (e.g. explicit Euler?)

Yes, explicit Euler is the method we used. Thank you for helping us clarify this, we have now added it to the methodology section in the revised manuscript.

11. Parameter calibration: I did not fully understand why the Bayesian approach was used if no uncertainty estimates of model parameters and model predictions are shown. The purpose of MCMC method is the evaluation of uncertainties, not calibration, for which much more efficient methods can be used. To my understanding, only optimal values of model parameters and predictions were used.

We agree with the reviewer on the purpose of MCMC method. The Bayesian framework provides easy estimation of the parameter uncertainty, as well as the MCMC method. As we explained in comment 6, we have now included the uncertainty estimation in the manuscript. The minimum, mean, maximum and standard deviation values present the distribution across catchments. The upper and lower bounds are defined from the plot of likelihood and parameter values. For each catchment, the likelihood values of each MCMC simulation are added up from the smallest parameter value, the upper and bottom bounds are defined when the sum of the likelihoods values just exceeds 5% and 95% of the total. The relative error is calculated as the half range between the upper and lower bounds as a percentage of the parameter with maximum likelihood value. Median relative error is the median level of the uncertainty among the sites.

	S_{b1} (mm)	t_w (days)	α	S_e (mm)	t_u(days)	S_{b2} (mm)	t_c (days)
Minimum	0.001	0.013	0.000	0.037	1.548	4.184	0.073
Mean	0.069	0.189	0.274	49.756	187.987	326.358	1.538
Maximum	1.013	0.533	0.300	339.181	1301.191	879.561	9.659
SD	0.14	0.09	0.14	69.44	221.68	183.98	1.51
Median Rel. Error (%)	33.57	33.31	23.74	46.73	24.05	11.54	29.19

12. Equation 14: the Authors should specify what N denotes. If N(z|mean,var) is the pdf of a Gaussian deviate z, the equation should be corrected accordingly.

We appreciate the reviewer's correction; N here represents pdf of a normal distribution.

13. Page 7050. It is not necessary to explain how the MCMC works.

We agree with the reviewer that MCMC is a widely used algorithm, not much description needed. For the sake of integrity, we still include a brief explanation about the MCMC we used, but have made our efforts to trim it, hope the reviewer found it satisfactory:

We then employ the Metropolis algorithm (Metropolis et al., 1953; Kuczera and Parent, 1998) adapted from Harman et al. (2011) to sample the parameter space towards constructing the posterior distribution. The algorithm, a Markov Chain Monte Carlo (MCMC) technique, is able to sample the parameters efficiently in the vicinity of the maximum likelihood. Starting with an optimum based on previous model development, we calculate the maximum likelihood value for each randomly selected set of parameters (θ_{i+1}) near the current parameter (θ_i). The new parameter set is accepted if it has a larger likelihood value ($L(X|\theta_{i+1})>L(X|\theta_i)$), i.e., it helps predict the runoff regime better than the previous set, and then a new search starts from a new set (θ_{i+1}). However, there is the possibility that this set can lead to another local optimum. To reach the globally optimal parameter set, we accept the inadequate parameter set if the ratio of the likelihood values $L(X|\theta_{i+1})/L(X|\theta_i)$ is larger than a uniform random value between zero and one. We run this algorithm to search the next available parameter set that improves upon the largest likelihood and save the 500 samples in a chain. This algorithm is run twice to generate1000 samples in total for each site. The parameter set with largest likelihood was selected as optimal for the full model.

14. Progression of model development. Considering the results presented by the Authors, I found that the only model modification that made a significant difference was the inclusion of snowmelt. The inclusion of other processes did not provide a considerable improvement. Is this a correct interpretation, and if yes, how can these model improvements be justified?

The reason that the snowmelt component improvement is more significant is that 1) snowmelt is only dominant influence in the runoff generation mechanism in the Idaho mountainous catchment; 2) snowmelt could transform both the timing and magnitude from rainfall to runoff. Therefore, we cannot get anywhere close to the observation if we do not include the snowmelt component.

Nonetheless, this significant improvement is not common; it cannot be seen for other processes, i.e. in the Georgia catchment. The reason is 1) in this Georgia catchment, interception, subsurface-influenced fast flow and phenology are all important for runoff generation and as a result, none of them are the only dominant process like snowmelt in Idaho; 2) in reality, the influence of these three processes in timing and magnitude of runoff regime curves is not as significant as the snowmelt, they are mostly adjustments to the estimated runoff; 3) in Georgia, the runoff regime curves follow the trends of precipitation regime curves, and the regime curves simulated by the base model already captures the trend and the timing, but misses the magnitude of peak flow. Thus, we don't expect to observe drastic improvements with the inclusion of these three components like the snowmelt component.

Yet, this does not mean they are unimportant, as can be seen in Fig.7-9, where the inclusion of these features helps improve the prediction of fast flow or slow flow. Moreover, with the combination of all three processes, the simulated regime curves improved considerably and approach the observed regime curves in terms of both timing and magnitude (Fig. 9).

15. Regional distribution of model parameters. Can the Authors show uncertainty estimates of model parameters? Where model parameters reasonably constrained through their calibration on regime curves?

We have updated Tab. 1 to present the minimum and maximum values, standard deviations (SD) and the relative errors (median rel. error (%)) of the parameters. The upper and lower bounds are defined from the plot of likelihood and parameter values. For each catchment, the likelihood values of each MCMC simulation are added up from the smallest parameter value, the upper and bottom bounds are defined when the sum of the likelihoods values just exceeds 5% and 95% of the total. The relative error is defined as the half range between the upper and lower bounds as a percentage of the parameter with maximum likelihood value. Median relative error is the median level among the sites.

		S_{b1} (mm)	t_w (days)	α	S_e (mm)	t_u(days)	S_{b2} (mm)	t_c (days)
East	**Mean**	0.065	0.218	0.306	36.846	120.260	281.858	1.469
	SD	0.158	0.078	0.128	49.540	64.644	163.704	1.268
	Median Rel. Error (%)	31.47	30.35	13.87	42.65	21.10	9.49	24.26
Center	**Mean**	0.068	0.140	0.221	78.007	323.567	350.640	1.763
	SD	0.098	0.084	0.147	101.615	282.408	160.895	2.049
	Median Rel. Error (%)	11.32	17.50	20.93	32.46	16.78	9.34	14.39
West	**Mean**	0.062	0.159	0.225	56.099	189.287	394.281	1.447
	SD	0.094	0.100	0.132	81.326	351.256	262.644	1.826
	Median Rel. Error (%)	29.19	23.86	20.34	51.94	29.30	7.71	27.28

16. Figures 1 and 2. It strikes that the pattern of PET is always so smooth. How was this calculated?

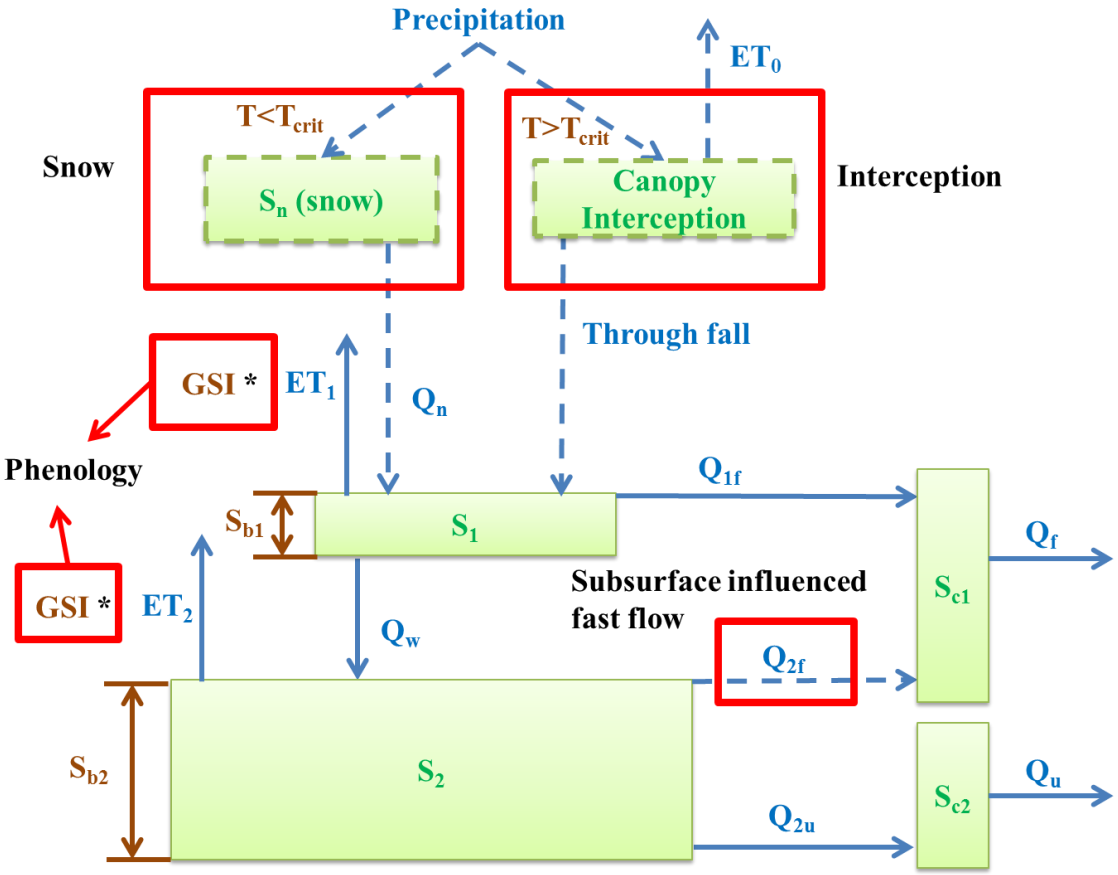
The PET is given from MOPEX website; it was calculated based on NOAA Pan Evaporation Atlas (NOAA, 1982). They calculated the PET by using Penman (1948)'s

method, the solar radiation required in the calculation was estimated from percent sunshine (Hamon et al, 1954).

17. Most figures could be improved. Figure 3 and 4: maybe use some more specific software for making these figures, represent reservoirs as reservoirs, and have different colour codes for model parameters, states and fluxes?

Thank you for the suggestion about these two figures. As Fig. 4 contains all the information in Fig.3, to shorten the manuscript, we cut Fig. 3 and re-sketched Fig. 4, hoping the reviewer finds this one appropriate:

Reservoirs are represented in solid green boxes; green is used for states, blue for fluxes and brown for model parameters. Red boxes show the four added processes and dashed lines denote the fluxes from these added processes.



18. Figures 7 and 8: model improvements are not apparent. Maybe use different metrics, and also show model improvements in other catchments. A different way of summarizing results is probably necessary.

Although the improvement in the total discharge is not apparent, the improvement in fast flow is obvious. As mentioned in the manuscript, we performed a multi-objective calibration procedure. We not only aimed to predict the total flow, but also to predict the

separation between fast flow and slow flow. It could be possible that a given process influences the fast flow or slow flow generation mechanism, but due to the small contribution of fast flow to the total flow, this influence could be overwhelmed when we view the total flow. Since our stated goal is not to predict the flow perfectly, but to detect the dominant processes, we should consider all the processes that could impact either the fast flow or the slow flow.

19. Figures 11 and 12: the attribution of model structures to different catchments is not clear to me. It seems a bit speculative if not properly justified.

The attribution of model structures to different catchments was based on the AIC value, following the statistical model selection steps in two directions: forward selection and backward selection.

For the forward selection, we started from a base model, compared the AIC value for the base model with the AIC values for the base model plus one of the four processes, and recorded which process helped reduce the AIC (or, improve model performance) most. That process was then regarded as the most dominant process. This is shown in Fig. 11. For the backward selection process, we started from the full model, removed processes one by one until the AIC value could not be further reduced. The remaining processes were considered as the minimum acceptable complexity for the model.

The reason we did this is that not all of the four processes are necessary for all the catchments, certain process may be dominant in some catchments but may never occur in others.. For example, the snowmelt process was apparently dominant in those mountainous catchments, but was never invoked in the warm catchments; and phenology, which is indispensable in the northeastern cold catchments, could have limit influence in southern catchments where temperature is always high. The backward selection approach was therefore used to eliminate unnecessary features from our simple model (minimize the complexity) to reveal and concentrate on the most necessary processes in those catchments.

References

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montanari, A.: Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice, Phys. Chem. Earth , 42-44, 70-76, doi:10.1016/j.pce.2011.07.037, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, F.: Decomposition of the mean squared error and NSE performance criteria: Implications from improving hydrological modeling, J. Hydrol., 377, 80-91, 2009.

Freer, J., Beven, K., and Ambroise, B.: Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, Water Resour. Res., 32(7), 2161-2173, 1996.

Hamon, R. W., Weiss, L. L., and Wilson, W. T.: Insolation as an Empirical Function of Daily Sunshine Duration, Monthly Weather Rev., 82 (6), 141-146, 1954.

NOAA Technical report NWS 33: Evaporation atlas for the contiguous 48 United States, Washington, D.C, 1982.

Penman, H. L.: Natural Evaporation from Open Water, Bare Soil and Grass, Proceedings of the Royal Society of London, Ser. A, Vol. 193, No. 1032, PP120-145 , April 1948.