Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models

Martyn P. Clark,[1] Andrew G. Slater,[2] David E. Rupp,[3] Ross A. Woods,[1] Jasper A. Vrugt,[4] Hoshin V. Gupta,[5] Thorsten Wagener,[6] and Lauren E. Hay[7]

Received 5 December 2007; revised 17 April 2008; accepted 12 May 2008; published 13 August 2008.

[1] The problems of identifying the most appropriate model structure for a given problem and quantifying the uncertainty in model structure remain outstanding research challenges for the discipline of hydrology. Progress on these problems requires understanding of the nature of differences between models. This paper presents a methodology to diagnose differences in hydrological model structures: the Framework for Understanding Structural Errors (FUSE). FUSE was used to construct 79 unique model structures by combining components of 4 existing hydrological models. These new models were used to simulate streamflow in two of the basins used in the Model Parameter Estimation Experiment (MOPEX): the Guadalupe River (Texas) and the French Broad River (North Carolina). Results show that the new models produced simulations of streamflow that were at least as good as the simulations produced by the models that participated in the MOPEX experiment. Our initial application of the FUSE method for the Guadalupe River exposed relationships between model structure and model performance, suggesting that the choice of model structure is just as important as the choice of model parameters. However, further work is needed to evaluate model simulations using multiple criteria to diagnose the relative importance of model structural differences in various climate regimes and to assess the amount of independent information in each of the models. This work will be crucial to both identifying the most appropriate model structure for a given problem and quantifying the uncertainty in model structure. To facilitate research on these problems, the FORTRAN-90 source code for FUSE is available upon request from the lead author.

Citation: Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. 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.

1. Introduction

[2] The accuracy of streamflow simulations in natural catchments will always be limited by simplified model representations of the real world as well as the availability and quality of hydrologic measurements. This is true for all models, regardless the amount of instrumentation within the basin. The outstanding research challenges are to identify the most appropriate model structure for a given problem and to quantify the predictive uncertainty in hydrologic model simulations.

[3] Progress on these problems requires understanding the nature of differences between models. Specific questions are as follows.

[4] 1. How do model structural differences influence simulations of model states and fluxes?

[5] 2. Is there a significant relationship between differences in model structure and model performance? How does this relationship vary regionally?

[6] 3. Why do some models perform better than others? Under what circumstances do models perform poorly?

[7] To address these questions, this paper introduces a computational framework to diagnose differences in hydrological model structures: the Framework for Understanding Structural Errors (FUSE). FUSE was used to construct 79 ''new'' hydrological models, each having a different structure. These new models were used to simulate streamflow in two of the basins used in the second and third workshops of the Model Parameter Estimation Experiment (MOPEX): the Guadalupe River (Texas) and the French Broad River (North Carolina). Model analyses involve assessment of overall model performance and

[1]NIWA, Christchurch, New Zealand.
[2]CIRES, University of Colorado, Boulder, Colorado, USA.
[3]DHI Water and Environment, Inc., Portland, Oregon, USA.
[4]Los Alamos National Laboratory, Los Alamos, New Mexico, USA.
[5]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.
[6]Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA.
[7]U.S. Geological Survey, Lakewood, Colorado, USA.

0043-1397/08/2007WR006735$09.00

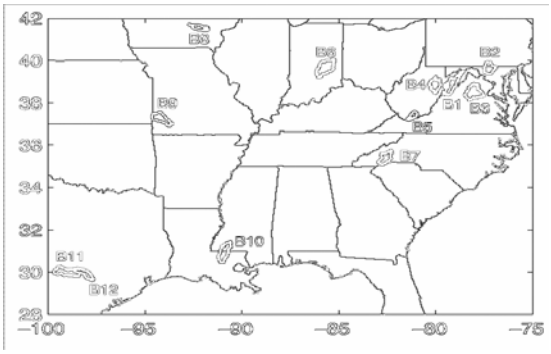

Figure 1. Location of the 12 MOPEX basins in the eastern United States. Basins marked B7 and B12 are the French Broad and Guadalupe rivers, respectively. See *Duan et al.* [2006] for more details.

diagnosis of model output during time periods when large model errors are observed.

2. The MOPEX Data Set

[8] The model simulations in this paper are produced using the MOPEX data set described by *Duan et al.* [2006]. This data set includes hydrometeorological and land surface characteristics data for twelve basins in the eastern United States (Figure 1). All models that participated in MOPEX were forced with daily estimates of basin-average precipitation and potential evapotranspiration. However, in this study the original MOPEX model forcings were replaced with daily estimates of (rain plus snowmelt) produced from simulations using the National Weather Service SNOW-17 model [*Anderson*, 1973] and daily estimates of adjusted potential evapotranspiration produced using the Sacramento model [*Burnash et al.*, 1973]. All models were evaluated using daily streamflow data obtained from the United States Geological Survey. Use of the same time series of rain plus melt and potential evapotranspiration as forcings for all models helps us maintain a control on the differences between models, thereby allowing us to concentrate on the impacts of model structural differences in the subsurface.

[9] Multimodel simulations are performed for two contrasting MOPEX basins, the French Broad River in North Carolina (the wettest of all twelve basins) and the Guadalupe River in Texas (the driest of all twelve basins). Figure 2 illustrates the varying controls of available energy and available water on the partitioning of precipitation between evaporation and runoff for each of the twelve MOPEX basins. When the annual available energy, expressed as potential evapotranspiration, is greater than the annual precipitation, the annual evaporation is limited by the annual supply of water. Conversely, when the available energy is less than the available precipitation, the annual evaporation is limited by the annual supply of energy [*Milly*, 1994; *Milly and Dunne*, 2002]. Evaporation in the French Broad River is constrained by the annual supply of energy, but in the Guadalupe River the annual supply of energy and water is approximately equal.

3. Modeling Philosophy

[10] A common device for understanding model structural differences is to run model intercomparison experiments. Recent examples in hydrology include the Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) [*Henderson-Sellers et al.*, 1993], the Distributed Model Intercomparison Project (DMIP) [*Reed et al.*, 2004], and the Model Parameter Estimation Experiment (MOPEX) [*Duan et al.*, 2006].

[11] These model intercomparison experiments have helped illuminate the wide range in model simulations that can arise when different models are forced with the same input data. However, the intercomparison experiments have been less successful in helping us understand the reasons for the intermodel differences. This is not surprising. Each individual model uses different parameterizations for different processes (e.g., surface runoff, percolation, base flow), and these different parameterizations interact in complex ways. It is therefore extremely difficult to link intermodel differences to the differences in specific process representations. Consequently, the understanding that is gained from model intercomparison experiments is largely limited to illuminating the different kinds of model behavior that can result from major differences in model structure [*Wetzel et al.*, 1996; *Koster and Milly*, 1997; *Slater et al.*, 2001].

[12] To improve our understanding of differences between models, the models in this study have been constructed in such a way that each model component can be evaluated in isolation. This is done via the following three steps.

[13] 1. Prescribe the type of model. In this paper the type of model is limited to lumped hydrological models run at a daily time step. In each model the vertical dimension

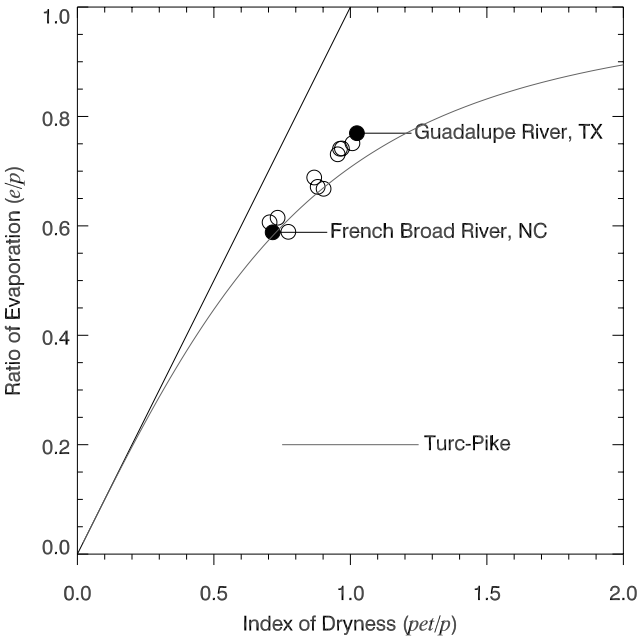

Figure 2. Relative controls of available energy (potential evapotranspiration (*pet*)) and available water (precipitation (*p*)) on the ratio of evaporation (*e*/*p*) in each of the twelve MOPEX basins. The two basins examined in this study, the French Broad and Guadalupe rivers, are depicted with filled circles. All basins have runoff ratios that are close to what is predicted by the Turc-Pike relationship, given by *Milly and Dunne* [2002] as $\frac{e}{p} = \left[1 + \left(\frac{pet}{p}\right)^{-v}\right]^{-1/v}$, where $v = 2$.

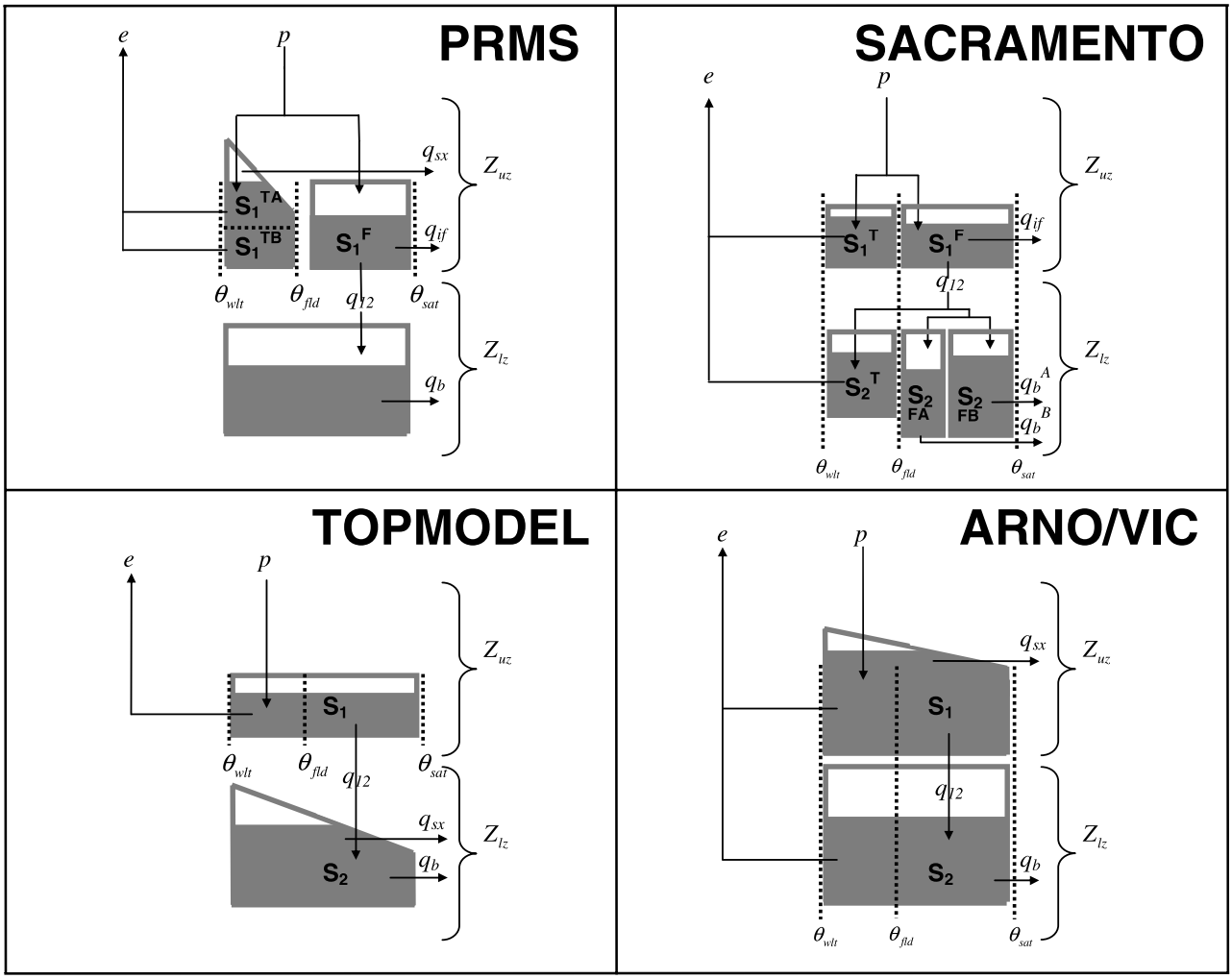

Figure 3. Simplified wiring diagrams for each of the four parent models (the state variables and fluxes are defined in Tables 1 and 2, respectively). Here Z_{uz} and Z_{lz} denote the depth of the upper and lower soil layers, and θ_{wlt}, θ_{fld}, and θ_{sat} denote the soil moisture at wilting point, field capacity, and saturation. Saturation-excess runoff (q_{sx}) is defined as the fraction of precipitation that falls on saturated areas of the basin and does not infiltrate into the soil; q_{sx} is shown as originating from the lower-zone storage in TOPMODEL because lower-zone storage in TOPMODEL controls the saturated area.

discretized into two zones: the unsaturated zone (above the water table) and the saturated zone (below the water table). For a review of other types of hydrological models see *Singh and Woolhiser* [2002] and *Kampf and Burges* [2007]. In this study we do not consider the models that use multiple soil layers to solve Richards' equation, as many of these models have an incomplete representation of interflow and base flow.

[14] 2. Define the major model-building decisions. These decisions include the architecture of the upper and lower soil layers, and the parameterizations for simulating evaporation, surface runoff, percolation of water between soil layers, interflow, and base flow.

[15] 3. Provide multiple options for each model building decision. Different modeling options were drawn from four parent models (Figure 3): the U.S. Geological Survey's Precipitation-Runoff Modeling System (PRMS) [*Leavesley et al.*, 1983, 1996], the NWS Sacramento model [*Burnash et al.*, 1973; *Burnash*, 1995; *Koren et al.*, 2004], TOPMODEL [*Beven and Kirkby*, 1979; *Ambroise et al.*, 1996; *Beven*, 1997; *Duan and Miller*, 1997; *Iorgulescu and Musy*, 1997]; and different versions of the Variable Infiltration Capacity (ARNO/VIC) model [*Zhao*, 1977; 1984; *Wood et al.*, 1992; *Liang et al.*, 1994] which borrows from the ARNO model [*Todini*, 1996]. The parent models include many and varied processes that interact with the subsurface (e.g., the vegetation submodels), but for this study we restrict attention to the subsurface in order to make the analysis manageable.

[16] In contrast to other studies that assess model complexity [e.g., *Desborough*, 1999; *Atkinson et al.*, 2002], this study diagnoses differences among model structures that are deemed (without additional information) to be equally plausible model structures, and we therefore have no a priori expectations of which models will perform better than others.

[17] The construction of models in this study is deliberately limited in scope in order to make the analysis manageable. We therefore consciously exclude performing

Table 1. State Variables

Variable	Description	Units
S_1	Total water content in the upper soil layer	mm
S_1^T	Tension water content in the upper soil layer	mm
S_1^{TA}	Primary tension water content in the upper soil layer	mm
S_1^{TB}	Secondary tension water content in the upper soil layer	mm
S_1^F	Free water content in the upper soil layer	mm
S_2	Total water content in the lower soil layer	mm
S_2^T	Tension water content in the lower soil layer	mm
S_2^{FA}	Free water content in the primary base flow reservoir	mm
S_2^{FB}	Free water content in the secondary base flow reservoir	mm

any surface energy balance calculations. We do not explicitly model interception and storage of water by the vegetation canopy or transpiration and evaporation of intercepted water. We also do not explicitly simulate the accumulation and ablation of the snowpack (models were forced with daily estimates of rain plus snowmelt produced from simulations using the National Weather Service SNOW-17 model). All models are run as a lumped model at a daily time step in which routing is calculated using a Gamma distribution. As noted previously, the parent models include many of the above processes.

[18] Despite these simplifications, the models are designed to provide a relatively complete representation of the major hydrologic fluxes in the subsurface. To illustrate this point, consider models that are excluded from this study. Many land surface models use multiple soil layers to solve Richards' equation (i.e., vertical water movement [*Boone and Wetzel*, 1996]), but use a relatively simple treatment of surface runoff and base flow. In many land surface models water simply dribbles out the bottom of a multilayer soil column [*Wetzel et al.*, 1996; *Boone and Wetzel*, 1996]. These models are well suited for their intended purpose: modeling energy and mass exchanges between the land and atmosphere, but are incomplete from a hydrologic perspective, and are thus not assessed as part of our study.

[19] Clearly, these design principles can be expected to influence the results that follow. By ensuring that all models represent the subsurface with a similar level of detail (or simplicity), and by ignoring other process components such as snow and vegetation, any intermodel differences are avoided that may occur because some models include specific processes and other models do not. For example, as part of PILPS 2(d) *Luo et al.* [2003] demonstrated large differences between models that do and do not include soil freezing processes. Nevertheless, we believe that the art of modeling is to ensure that all relevant hydrological processes are included and that appropriate computational weight be given to each process on the basis of its relative importance. Hence, the emphasis in FUSE is not in the intermodel differences that arise from "missing processes"; rather, in the intermodel differences that arise from different (but equally complete) plausible representations of the real world.

4. Model Formulation

[20] The major model building decisions are the architecture of the upper soil layer (the unsaturated zone), the architecture of the lower soil layer (the saturated zone), and the choice of parameterization for evaporation, vertical percolation of water between the two soil layers, interflow, base flow, and surface runoff. The following sections describe the construction of the 79 models in terms of their state equations and flux parameterizations. Tables 1 and 2 define model state variables and model fluxes, while Tables 3 and 4 define model parameters.

4.1. State Equations for the Upper Layer

[21] The water content of the upper soil layer (S_1) can be defined by a single state variable (TOPMODEL, ARNO/VIC (equation (1a)), separate state variables for tension storage (below field capacity) and free storage (above field capacity) (Sacramento (equation (1b)), and further discretization of upper zone tension storage into two zones (PRMS (equation (1c)). The state equations are

$$\frac{dS_1}{dt} = (p - q_{sx}) - e_1 - q_{12} - q_{if} - q_{ufof} \tag{1a}$$

$$\begin{aligned} \frac{dS_1^T}{dt} &= (p - q_{sx}) - e_1 - q_{utof} \\ \frac{dS_1^F}{dt} &= q_{utof} - q_{12} - q_{if} - q_{ufof} \end{aligned} \tag{1b}$$

$$\begin{aligned} \frac{dS_1^{TA}}{dt} &= (p - q_{sx}) - e_1^A - q_{urof} \\ \frac{dS_1^{TB}}{dt} &= q_{urof} - e_1^B - q_{utof} \\ \frac{dS_1^F}{dt} &= q_{utof} - q_{12} - q_{if} - q_{ufof} \end{aligned} \tag{1c}$$

Table 2. Model Fluxes

Variable	Description	Units
p	Precipitation	mm d^{-1}
pet	Potential evapotranspiration	mm d^{-1}
e_1	Evaporation from the upper soil layer	mm d^{-1}
e_2	Evaporation from the lower soil layer	mm d^{-1}
e_1^A	Evaporation from the primary tension store	mm d^{-1}
e_1^B	Evaporation from the secondary tension store	mm d^{-1}
q_{sx}	Surface runoff	mm d^{-1}
q_{12}	Percolation of water from the upper to the lower layer	mm d^{-1}
q_{if}	Interflow	mm d^{-1}
q_b	Base flow	mm d^{-1}
q_b^A	Base flow from the primary reservoir	mm d^{-1}
q_b^B	Base flow from the secondary reservoir	mm d^{-1}
q_{urof}	Overflow of water from the primary tension store in the upper soil layer	mm d^{-1}
q_{utof}	Overflow of water from tension storage in the upper soil layer	mm d^{-1}
q_{ufof}	Overflow of water from free storage in the upper soil layer	mm d^{-1}
q_{stof}	Overflow of water from tension storage in the lower soil layer	mm d^{-1}
q_{sfof}	Overflow of water from free storage in the lower soil layer	mm d^{-1}
q_{sfofa}	Overflow of water from primary base flow storage in the lower soil layer	mm d^{-1}
q_{sfofb}	Overflow of water from secondary base flow storage in the lower soil layer	mm d^{-1}

Table 3. Adjustable Model Parameters

Parameter	Description	Units	Lower Limit	Upper Limit
$S_{1,\max}$	Maximum storage in the upper layer	mm	50.000	5000.000
$S_{2,\max}$	Maximum storage in the lower layer	mm	100.000	10000.000
ϕ_{tens}	Fraction total storage as tension storage	-	0.050	0.950
ϕ_{rchr}	Fraction of tension storage in primary zone (upper layer)	-	0.050	0.950
ϕ_{base}	Fraction of free storage in primary reservoir (lower layer)	-	0.050	0.950
r_1	Fraction of roots in the upper layer	-	0.050	0.950
k_u	Percolation rate	mm day^{-1}	0.010	1000.000
c	Percolation exponent	-	1.000	20.000
α	Percolation multiplier for the lower layer	-	1.000	250.000
ψ	Percolation exponent for the lower layer	-	1.000	5.000
κ	Fraction of percolation to tension storage in the lower layer	-	0.050	0.950
k_i	Interflow rate	mm day^{-1}	0.010	1000.000
k_s	Base flow rate	mm day^{-1}	0.001	10000.000
n	Base flow exponent	-	1.000	10.000
v	Base flow depletion rate for single reservoir	d^{-1}	0.001	0.250
v_A	Base flow depletion rate for primary reservoir	d^{-1}	0.001	0.250
v_B	Base flow depletion rate for secondary reservoir	d^{-1}	0.001	0.250
$A_{c,\max}$	Maximum saturated area (fraction)	-	0.050	0.950
b	ARNO/VIC "b" exponent	-	0.001	3.000
λ	Mean of the log-transformed topographic index distribution	m	5.000	10.000
χ	Shape parameter defining the topographic index distribution	-	2.000	5.000
μ_τ	Time delay in runoff	days	0.010	5.000

where state variables and fluxes are defined in Tables 1 and 2, respectively. In equations (1b) and (1c) precipitation is added to free storage when the tension storage is at capacity (as represented by q_{utof}; see section 4.8). In the formulations that follow, the variables S_1^T, S_1^F, and S_1 are required. These variables are not always tracked as model states, but can be estimated from equation (1a) as $S_1^T = \min(S_1, S_{1,\max}^T)$ and $S_1^F = \max(0, S_1 - S_{1,\max}^T)$; from equation (1b) as $S_1 = S_1^T + S_1^F$; and from equation (1c) as $S_1^T = S_1^{TA} + S_1^{TB}$ and $S_1 = S_1^{TA} + S_1^{TB} + S_1^F$.

4.2. State Equations for the Lower Layer

[22] The lower soil layer can be defined by a single state variable with no evaporation (TOPMODEL and PRMS (equation (2a)), a single state variable with evaporation (ARNO/VIC (equation (2b)), or a tension reservoir combined with two parallel tanks (Sacramento (equation (2c)). The state equations are

$$\frac{dS_2}{dt} = q_{12} - q_b \tag{2a}$$

$$\frac{dS_2}{dt} = q_{12} - e_2 - q_b - q_{sfof} \tag{2b}$$

$$\begin{aligned} \frac{dS_2^T}{dt} &= \kappa q_{12} - e_2 - q_{stof} \\ \frac{dS_2^{FA}}{dt} &= \frac{(1-\kappa)q_{12}}{2} + \frac{q_{stof}}{2} - q_b^A - q_{sfofa} \\ \frac{dS_2^{FB}}{dt} &= \frac{(1-\kappa)q_{12}}{2} + \frac{q_{stof}}{2} - q_b^B - q_{sfofb} \end{aligned} \tag{2c}$$

where again state variables and fluxes are defined in Tables 1 and 2, respectively. Tension storage in equation (2b) (used to compute evaporation) is $S_2^T = \min(S_2, S_{2,\max}^T)$, and total lower-zone storage in equation (2c) is $S_2 = S_2^T + S_2^{FA} + S_2^{FB}$.

4.3. Evaporation

[23] When evaporation is modeled in both soil layers, evaporation parameterizations can be broadly classified into "sequential" and "root weighting" schemes. In the sequential method, the potential evaporative demand (*pet*) is first satisfied by evaporation from the upper soil layer, and any residual evaporative demand is satisfied by evaporation from the lower soil layer:

$$e_1 = pet \frac{\min\left(S_1^T, S_{1,\max}^T\right)}{S_{1,\max}^T} \tag{3a}$$

$$e_2 = (pet - e_1) \frac{\min\left(S_2^T, S_{2,\max}^T\right)}{S_{2,\max}^T} \tag{3b}$$

[24] In the root-weighting method, evaporation is computed on the basis of the relative root fractions in each of the soil layers [e.g., *Desborough*, 1997]:

$$e_1 = pet\ r_1 \frac{\min\left(S_1^T, S_{1,\max}^T\right)}{S_{1,\max}^T} \tag{3c}$$

$$e_2 = pet\ r_2 \frac{\min\left(S_2^T, S_{2,\max}^T\right)}{S_{2,\max}^T} \tag{3d}$$

where r_1 and r_2 are the relative root fractions in the upper and lower layer ($r_1 + r_2 = 1$). Note from equations (3a)–(3d) that the root weighting method will produce higher evaporation from the lower soil layer when the soil is at field capacity (assuming $r_2 > 0$).

[25] The peculiarities in model architecture require slightly different applications of the evaporation parameterizations.

Table 4. Derived Model Parameters

Parameter	Description	Units	Equation
$S^T_{1,\max}$	Maximum tension storage in the upper layer	mm	$S^T_{1,\overline{\max}}\phi_{tens}S_{1,\max}$
$S^T_{2,\max}$	Maximum tension storage in the lower layer	mm	$S^T_{2,\overline{\max}}\phi_{tens}S_{2,\max}$
$S^F_{1,\max}$	Maximum free storage in the upper layer	mm	$S_{1,\max}{}^F = (1 - \phi_{tens})S_{1,\max}$
$S^F_{2,\max}$	Maximum free storage in the lower layer	mm	$S^F_{2,\overline{\max}}(1 - \phi_{tens})S_{2,\max}$
$S^{TA}_{1,\max}$	Maximum storage in the primary tension reservoir	mm	$S^{TA}_{1,\overline{\max}}\ \phi_{rchr}S^T_{1,\max}$
$S^{TB}_{1,\max}$	Maximum storage in the secondary tension reservoir	mm	$S^{TB}_{1,\overline{\max}}\ (1 - \phi_{rchr})S^T_{1,\max}$
$S^{FA}_{2,\max}$	Maximum storage in the primary base flow reservoir	mm	$S^{FA}_{2,\overline{\max}}\ \phi_{base}S_{2,\max}{}^F$
$S^{FB}_{2,\max}$	Maximum storage in the secondary base flow reservoir	mm	$S^{FB}_{2,\overline{\max}}\ (1 - \phi_{base})S^F_{2,\max}$
r_2	Root fraction in the lower soil layer	-	$r_2 = 1 - r_1$
λ_n	Mean of the power-transformed topographic index	m	equation (8)

In the PRMS architecture where there are two tension reservoirs in the upper layer (refer to the state equation (1c)), the fluxes e^A_1 and e^B_1 are computed using either (3a) and (3b) or (3c) and (3d), with corresponding substitutions of the state variables for each of the tension stores (in this case evaporation is not computed from the lower soil layer). Moreover, (3a) is used in the TOPMODEL architecture where evaporation is only computed from the upper soil layer (refer to state equation (2a)).

4.4. Percolation

[26] Richards' equation is commonly viewed as the physically correct method to model vertical water movement [e.g., *Boone and Wetzel*, 1996]. However, large-scale application of Richards' equation is based on the assumptions that the soil is spatially homogeneous and that functional relations can be specified that relate moisture content, capillary potential, and hydraulic conductivity of the soil [*Henderson-Sellers et al.*, 1993; *Beven*, 2002]. In this study three conceptual models are used to parameterize percolation of water from the upper to lower soil layer:

[27] Percolation is parameterized as

$$q_{12} = k_u\left(\frac{S_1}{S_{1,\max}}\right)^c \tag{4a}$$

$$q_{12} = k_u\left(\frac{S^F_1}{S^F_{1,\max}}\right)^c \tag{4b}$$

$$q_{12} = q_0 d_{lz}\left(\frac{S^F_1}{S^F_{1,\max}}\right) \tag{4c}$$

where in equation (4c) q_0 is the base flow at saturation (computed using equation (6), described below), and d_{lz} is the lower-zone percolation demand

$$d_{lz} = 1 + \alpha\left(\frac{S_2}{S_{2,\max}}\right)^{\psi} \tag{4d}$$

[28] Note that each parameterization has two parameters. Equation (4a) (used in VIC) is equivalent to the gravity drainage term in Richards' equation, and often has a large exponent c to limit drainage below field capacity. In contrast, equation (4b) (used in PRMS) does not allow drainage below field capacity and the exponent is often close to unity. Nonlinearities in percolation in the Sacramento parameterization in equation (4c) [*Burnash et al.*, 1973] are controlled by lower-zone storage; percolation will be fastest when the lower zone is dry.

4.5. Interflow

[29] In this study we use a simple parameterization of interflow

$$q_{if} = 0 \tag{5a}$$

$$q_{if} = k_i\left(\frac{S^F_1}{S^F_{1,\max}}\right) \tag{5b}$$

where the option of zero interflow is allowed because interflow is not parameterized explicitly in TOPMODEL and ARNO/VIC.

4.6. Base Flow

[30] The parameterizations for base flow in this study are

$$q_b = vS_2 \tag{6a}$$

$$q_b = v_A S^{FA}_2 + v_B S^{FB}_2 \tag{6b}$$

$$q_b = k_s\left(\frac{S_2}{S_{2,\max}}\right)^n \tag{6c}$$

$$q_b = \frac{k_s m}{\lambda^n_n}\left(\frac{S_2}{m\,n}\right)^n \tag{6d}$$

which define a single linear reservoir (PRMS, equation (6a)), two parallel linear reservoirs (Sacramento, equation (6b)), a nonlinear storage function used to mimic the parameterization in ARNO/VIC (equation (6c)), and the TOPMODEL power law parameterization (equation (6d)). In the TOPMODEL case, the storage capacity of the lower zone is $mn = S_{2,\max}$, and therefore the subsurface depth scaling parameter $m = S_{2,\max}/n$. In equation (6d) the parameter λ_n is the mean of the power-transformed topographic index (defined in equation (8) below).

[31] Implementing the TOPMODEL parameterization requires a distribution of topographic index values for each river basin [*Beven and Kirkby*, 1979]. While it is possible to derive such distributions from digital terrain data, in this

study the topographic index distribution was defined using a three-parameter Gamma distribution. Following *Sivapalan et al.* [1987],

$$f(\zeta) = \frac{1}{\chi\Gamma(\phi)}\left(\frac{\zeta-\mu}{\chi}\right)^{\phi-1}\exp\left(-\frac{\zeta-\mu}{\chi}\right) \tag{7}$$

[32] The variable $\zeta = \ln(a/\tan\beta)$ has mean $\lambda = \chi\phi + \mu$ and variance $\chi^2\phi$, where $\phi = (\lambda - \mu)/\chi$. The mean ($\lambda$) and shape parameters ($\chi$) are kept as adjustable parameters, but the offset is set to $\mu = 3$. The offset $\mu = 3$ is consistent with published probability distributions of the log-transformed topographic index (as shown by *Beven* [1997]).

[33] Equation (7) defines the topographic index in log space (mean value = λ), so it is necessary to transform the topographic index to be consistent with the power law transmissivity profile used in equation (6d). The mean of the power-transformed topographic index is

$$\lambda_n = \int_0^\infty [\exp(\zeta)]^{1/n} f(\zeta)\, d\zeta \tag{8}$$

where λ_n can be computed as a preprocessing step. Note that the quantity $(k_s m)/\lambda_n^n$ in equation (6d) is temporally constant, so equation (6d) is functionally very similar to equation (6c).

[34] The base flow parameterizations are intimately tied to the lower-zone architecture. The correspondence is as follows: The single linear reservoir (equation (6a)) is used in conjunction with a single reservoir of infinite size (state equation (2a)), two parallel reservoirs (equation (6b)) are used in conjunction with the two parallel reservoirs described by state equation (2c), the nonlinear storage function (equation (6c)) is used in conjunction with a single reservoir of fixed size (state equation (2b)), and the TOPMODEL power law parameterization (equation (6d)) is used in conjunction with a single reservoir of infinite size (state equation (2a)). Other combinations of base flow parameterizations and lower-layer architecture are technically possible, but do not add any additional information. For example, using (6c) in conjunction with (2a) is equivalent to using (6d) with (2a); it is possible to modify the parameters in (6a) to account for fixed storage, but this modification changes the form of the base flow parameterization and is really just a special case of equation (6c) where the exponent is one. For these reasons the selection of base flow parameterizations and lower-zone architecture is a single modeling decision (see section 5).

4.7. Surface Runoff

[35] In this study surface runoff is only generated using the saturation-excess mechanism, where rain falls on saturated areas of the basin. Saturated area, A_c, is computed as

$$A_c = \frac{S_1^T}{S_{1,\max}^T} A_{c,\max} \tag{9a}$$

$$A_c = 1 - \left(1 - \frac{S_1}{S_{1,\max}}\right)^b \tag{9b}$$

$$A_c = \int_{\zeta^{crit}}^\infty f(\zeta)\, d\zeta \tag{9c}$$

where the parameterizations in equations (9a)–(9c) are based loosely on the methods used in PRMS, ARNO/VIC, and TOPMODEL, respectively, and ζ^{crit} is defined below.

[36] The topographic index distribution is defined using the Gamma distribution in equation (7). The critical (power transformed) topographic index value for saturation is [*Rupp and Woods*, 2008]

$$\zeta_n^{crit} = \lambda_n\left(\frac{S_2}{S_{2,\max}}\right)^{-1} \tag{10a}$$

which is transformed to log space by

$$\zeta^{crit} = \ln\left[\left(\zeta_n^{crit}\right)^n\right] \tag{10b}$$

so the integral in equation (9c) can be solved efficiently using the incomplete Gamma function.

[37] The modeled saturation-excess runoff is then simply

$$q_{sx} = A_c p \tag{11}$$

4.8. Bucket Overflow

[38] Additional fluxes of water occur when one of the storages reach capacity (Table 1). In the upper soil layer, the bucket overflow from the primary tension store (q_{urof}) represents precipitation into the second tension store (equation (1c)); the bucket overflows from tension storage (q_{utof}) represent precipitation into free storage (equations (1b) and (1c)); and the bucket overflow from free storage (q_{ufof}) represents additional surface runoff (equations (1a)–(1c)). In the lower soil layer, the bucket overflow from tension storage (q_{stof}) represents additional vertical drainage (q_{12}) into free storage (equation (2c)), and the bucket overflow from free storage represents additional base flow (equations (2b) and (2c)).

[39] Following *Kavetski and Kuczera* [2007], logistic functions are used to smooth the thresholds associated with the fixed capacity of model storages:

$$q_{urof} = (p - q_{sx})\,\Phi\left(S_1^{TA}, S_{1,\max}^{TA}, \omega\right) \tag{12a}$$

$$q_{utof} = \begin{cases} (p - q_{sx})\ \Phi\left(S_1^T, S_{1,\max}^T, \omega\right), & \text{(equation (1b))} \\ q_{urof}\ \Phi\left(S_1^{TB}, S_{1,\max}^{TB}, \omega\right), & \text{(equation (1c)).} \end{cases} \tag{12b}$$

$$q_{ufof} = \begin{cases} (p - q_{sx})\ \Phi\left(S_1, S_{1,\max}, \omega\right), & \text{(equation (1a))} \\ q_{utof}\ \Phi\left(S_1^F, S_{1,\max}^F, \omega\right), & \text{(equations (1b) and (1c))} \end{cases} \tag{12c}$$

$$q_{stof} = \kappa\, q_{12}\ \Phi\left(S_2^T, S_{2,\max}^T, \omega\right) \tag{12d}$$

$$q_{sfof} = q_{12}\ \Phi\left(S_2, S_{2,\max}, \omega\right) \tag{12e}$$

$$q_{sfofa} = \left[\frac{(1-\kappa)q_{12}}{2} + \frac{q_{stof}}{2}\right]\Phi\left(S_2^{FA}, S_{2,\max}^{FA}, \omega\right) \tag{12f}$$

$$q_{sfofb} = \left[\frac{(1-\kappa)q_{12}}{2} + \frac{q_{stof}}{2}\right]\Phi\left(S_2^{FB}, S_{2,\max}^{FB}, \omega\right) \tag{12g}$$

In equations (12a)–(12g) $\Phi(S,\ S_{\max},\ \omega)$ is the logistic function, and is computed as

$$\Phi(S, S_{\max}, \omega) = \frac{1}{1 + \exp\left(\dfrac{S - S_{\max} - \omega\varepsilon}{\omega}\right)} \tag{12h}$$

where ω is the degree of smoothing ($\omega = \rho S_{\max}$) and $\varepsilon = 5$ is a multiplier that ensures storage is always less than capacity. The ρ parameter is used to specify the degree of smoothing as a fraction of total storage in each state variable. In this study we use $\rho = 0.01$ for all model state variables.

4.9. Routing

[40] The time delay in runoff is modeled using a two-parameter Gamma distribution [*Press et al.*, 1992]

$$P(a, x) = \frac{\gamma(a, x)}{\Gamma(a)} \tag{13a}$$

Here $\gamma(.)$ is the incomplete Gamma function, a is the shape of the Gamma distribution, and

$$x = \tau\frac{a}{\mu_\tau} \tag{13b}$$

where τ is the time delay (in days), and μ_τ is the mean of the Gamma distribution (also in days). While the mean of the time delay histogram, μ_τ, is held as an adjustable parameter, the shape of the time delay histogram is fixed ($a = 3$). Equation (13a) is used to compute the fraction of runoff in the current time step which is discharged in each future time step.

5. Multimodel Configuration

[41] Models in FUSE are constructed by combining many of the different architecture and flux parameterizations given in section 4. The scope for producing a comprehensive multimodel set is limited by the ability of different components to work together in a seamless fashion as well as managing the vast computing power required as the dimension of the problem increases. To make this a feasible task several key decisions were made that reduced the complexity of the problem. These are given below.

[42] 1. Choose between three possible upper layer architectures (equations (1a)–(1c)).

[43] 2. Choose between three possible lower-layer architectures (equations (2a)–(2c)) and their associated base flow parameterizations (equations (6a)–(6d)).

[44] 3. Choose between three possible percolation parameterizations (equations (4a)–(4c)).

[45] 4. Choose between three possible parameterizations to compute saturated areas and surface runoff (equations (9a)–(9c)).

[46] 5. Only the sequential method for evaporation was applied (equations (3a)–(3b)).

[47] 6. Interflow was not computed in any of the models.

[48] 7. Models were run using a daily time step, and hence infiltration-excess runoff was not computed by any of the models.

[49] 8. A gamma distribution was used to route runoff to the basin outlet in all cases (equation (13a)).

[50] This resulted in 108 possible models (3 * 4 * 3 * 3) – recall that the base flow parameterizations (equations (6a)–(6d)) are tied to the choice of lower-zone architecture (equations (2a)–(2c)), and together provide 4 modeling options (not 4 * 3). Some model combinations were deemed unsuitable (for example, equation (4a) computes percolation of water between the two soil layers on the basis of total water storage in the upper soil layer, but this parameterization is only really appropriate when the upper soil layer is defined by a single state variable (equation (1a)). All cases where model components were incompatible were removed, resulting in 79 models for the subsequent analyses.

[51] To simplify programming, the multiple modeling options were configured using a modular structure. Each flux equation (i.e., evaporation, percolation, interflow, base flow, and surface runoff) was formulated as a function of the model state. A separate subroutine was used for each flux parameterization, and FORTRAN-90 ''case'' statements were used to distinguish between the different modeling options. The modeled fluxes were then used in model state equations to compute the time derivative of model states, where again FORTRAN-90 case statements were used to distinguish between the different model architectures. FUSE differs from other modular hydrological models [e.g., *Leavesley et al.*, 1996; 2002] because it modularizes individual flux equations, rather than linking existing submodels. Imposing a modular structure at the level of individual flux equations greatly simplifies adding new modeling options; the only real constraint is the computing resources required to run the large number (>1000) of possible model structures.

[52] The computational scheme for FUSE is detailed in Appendix A. Briefly, model equations are solved using an implicit scheme with adaptive substeps. As pointed out by *Kavetski et al.* [2003], controlled numerical solutions such as the implicit scheme used in FUSE are much more accurate than the fixed-step ''operator-splitting'' explicit Euler method that is commonly used in hydrological models. Moreover, it is straightforward to use the implicit scheme to solve equations from multiple model combinations (for example, there is no need to define the order of flux calculations as in the operator-splitting approach).

6. Results

6.1. Model Performance

[53] The 79 models assessed in this study are all equally plausible model structures, and given that there is no further information on their performance, we assume that all

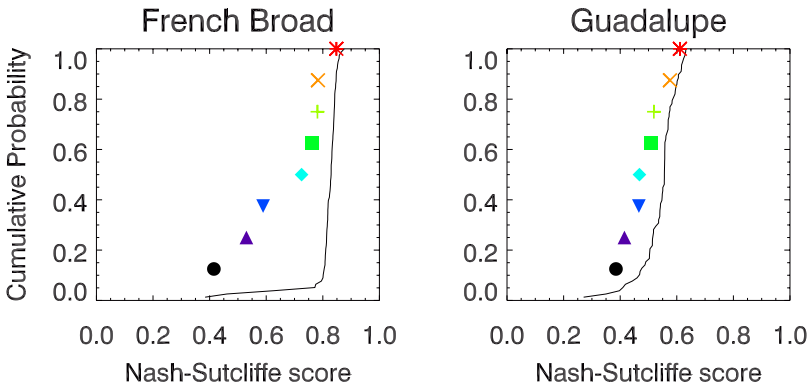

Figure 4. Cumulative probability distribution of Nash-Sutcliffe scores in the two study basins. Models participating in MOPEX [*Duan et al.*, 2006] are shown with symbols and the bold line shows results from the 79 models constructed for this study.

models perform equally well when provided with an optimal parameter set. To test this assumption each model was calibrated separately over the period 1980–1990 using the *Duan et al.* [1992] shuffled complex evolution (SCE) method (data from 1979 was used for model spin-up). Specifically, SCE was used to identify the parameter set in each individual model with the lowest root mean squared error (RMSE) as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_i^{sim} - q_i^{obs}\right)^2} \tag{14}$$

where q_i^{sim} and q_i^{obs} are the simulated and observed streamflow for the ith day in the calibration period, and n is the number of days in the calibration period ($n = 4018$ days). SCE calibration runs were limited to a maximum of 10,000 function evaluations.

[54] To enable comparisons with published MOPEX results [*Duan et al.*, 2006], model performance is assessed using the Nash-Sutcliffe score,

$$NS = 1 - \frac{\sum_{i=1}^{n}\left(q_i^{sim} - q_i^{obs}\right)^2}{\sum_{i=1}^{n}\left(q_i^{obs} - \overline{q^{obs}}\right)^2} \tag{15}$$

which is simply the ratio of the sum of squared errors to the variance in observed streamflow. The only difference between the Nash-Sutcliffe score and the RMSE is that the Nash-Sutcliffe score is scaled by the variance in observed streamflow (the denominator in equation (15)). Use of the RMSE and Nash-Sutcliffe scores emphasizes errors in high flow.

[55] Figure 4 compares the performance of the 79 models developed with FUSE against the performance of the eight models that participated in the MOPEX experiment. Results show that the new models perform just as well as the models that participated in MOPEX. The slightly "better" performance in Figure 4 most likely occurs because our calibration strategy (minimizing RMSE) mimics the MOPEX assessment of model performance (Nash-Sutcliffe); no attempt was made to ensure that the optimal parameter sets in each model minimizes other objective functions related to low flow and runoff ratios. The various groups that participated in MOPEX all used different calibration strategies, some groups relied on a priori model parameters and other groups examined multiple objective functions [*Duan et al.*, 2006].

[56] The most salient feature of Figure 4 is that model performance is generally worse in the Guadalupe than in the French Broad. There can be many reasons for poor model performance in the Guadalupe, such as missing processes (no mechanism for infiltration-excess runoff, no vegetation submodel), bad input data, or simply that runoff in the Guadalupe is sensitive to the spatial patterns of precipitation that are not resolved by lumped models. FUSE can easily be extended to include additional processes, additional spatial detail, and the additional temporal resolution required to resolve the infiltration-excess runoff mechanism.

[57] An interesting result from Figure 4 is that the differences in the performance of 79 models developed for this study are much smaller in the French Broad basin (the wetter basin) than in the Guadalupe (the drier basin). In the French Broad the majority of models have a Nash-Sutcliffe score close to 0.8, whereas Nash-Sutcliffe scores in the Guadalupe range from ~0.4 to ~0.65. The miniscule differences in model performance in the French Broad imply there is enough flexibility in the different model structures to enable a good fit to measured streamflow, whereas the large differences in model performance in the Guadalupe imply that some model structures may be better suited to that basin than others. Put differently, the choice of model structure in the Guadalupe is just as important as the choice of model parameters. Section 6.2 seeks to identify reasons for the large differences in model performance in the Guadalupe.

[58] Note that in contrast to the results from the 79 models, there are large differences in performance of the MOPEX models in the French Broad (Figure 4). The differences between the performance of the MOPEX models can be attributed to both differences in the calibration strategy as well as differences in model structure. In this study an identical calibration strategy was used for each of the 79 models, so, unlike the MOPEX experiment, differences in model performance in this study can be attributed solely to differences in model structure.

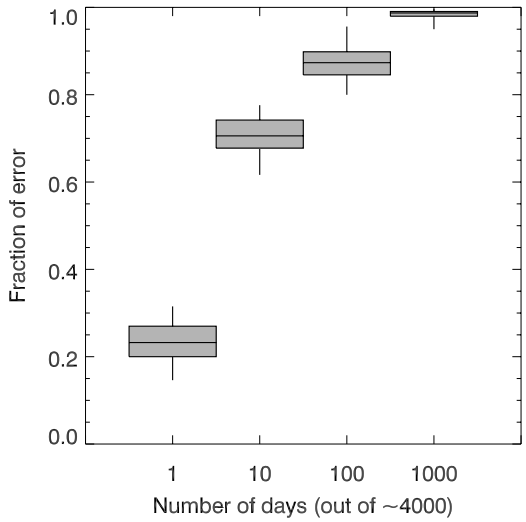

Figure 5. The fraction of total error that is attributable to a specific number of days in the time series in the Guadalupe River basin (see text for computational details). The box-and-whisker icon shows the distribution of fractional error from all 79 models (minimum, lower quartile, median, upper quartile, and maximum).

6.2. Relationship Between Model Performance and Model Structure

[59] In this section we seek to identify reasons for the differences in model performance in the Guadalupe River basin. Attention is focused solely on the Guadalupe because significant differences in model performance are not discernable in the French Broad. The reasons for differences in model performance will undoubtedly depend on the dominant runoff generation mechanisms in the basin where the model is applied, which depends on both the climate regime and basin characteristics. Results for the Guadalupe are presented as an example of how the choice of model structure can impact model performance.

[60] A first step for diagnosing differences between models is to examine periods where models perform poorly. The squared errors on each day have a discrete contribution to the root mean squared error (equation (14)). The fraction of total error attributable to a specific number of days in the time series can then be identified simply by ranking the time series of squared errors in each model from highest to lowest. These results are shown for the Guadalupe River basin in Figure 5, and show that approximately 70% of the total error is caused by errors on only 10 days in the calibration period. Close inspection of these periods may help understand reasons for the difference in model performance.

[61] Figure 6 illustrates time series of precipitation and runoff for the Guadalupe River basin, along with color coding that identifies time periods when each of the 79 models have absolute errors larger than 1 mm d^{-1}. As expected, errors are largest during large precipitation events, periods generally less than 1–2 days in duration when the models either overestimate or underestimate the observed streamflow response to precipitation. Differences in model performance could therefore potentially be explained by differences in surface runoff generation mechanisms.

[62] Figure 7 illustrates the mean and standard deviation of the time series of saturated area, as computed by each of the 79 models. The models with the highest skill (purple-blue colors) are those with a relatively low mean and high standard deviation in saturated areas. In these models saturated area is controlled by lower-zone storage (equation (9c), the squares in Figure 7). Inspection of the time series of saturated area from each of the models shows that the models that use equation (9c) have low-frequency variability in saturated areas (not shown). Figure 7 thus demonstrates there is some relationship between model performance and model structure; however, there are no data to test if the saturated area dynamics predicted by equation (9c) are realistic for the Guadalupe River basin (as was done by *Guntner et al.* [1999]).

7. Summary and Discussion

[63] This paper introduces a computational framework to diagnose differences in hydrological model structures: the Framework for Understanding Structural Errors (FUSE). FUSE was used to construct 79 "new" hydrological models, each having a different structure. These new models were used to simulate streamflow in two of the basins used in the second and third workshops of the Model Parameter Estimation Experiment (MOPEX): the Guadalupe River (Texas) and the French Broad River (North Carolina). Model analyses involve assessment of model performance and diagnosis of model output during time periods with large errors.

[64] Results show that the performance of the 79 models, as evaluated using the Nash-Sutcliffe score, is as least as good as the performance of the eight models that participated in the MOPEX experiment. However, the Nash-Sutcliffe score emphasizes errors in high flow, and by itself is a weak metric for model evaluation [*Schaefli and Gupta*, 2007]. Further work is needed to evaluate model performance with respect to multiple criteria, including assessment of model performance during low-flow periods [*Boyle et al.*, 2000], assessment of model performance in the frequency domain [*Parada et al.*, 2003], and assessment of model performance with respect to "diagnostic signatures" that are extracted from the data to explain different hydrological processes in the basin [*Gupta et al.*, 2008; *Yilmaz et al.*, 2008]. This research will help identify extensions to the model framework that are necessary to simulate dominant hydrological processes in basins where the model is applied.

[65] Differences in model performance are much smaller in the French Broad River basin (the wetter basin) than in the Guadalupe River basin (the drier basin). The nondiscernable differences in model performance in the French Broad imply there is enough flexibility in the different model structures to enable a good fit to measured streamflow; that is, model parameters can compensate for model structural differences. In contrast to the French Broad, the large differences in model performance in the Guadalupe imply that some model structures may be better suited to that basin than others. However, it is unlikely that a single model structure provides the best streamflow simulation for multiple basins in different climate regimes [e.g., *van Werkhoven et al.*, 2008], so future work is necessary to

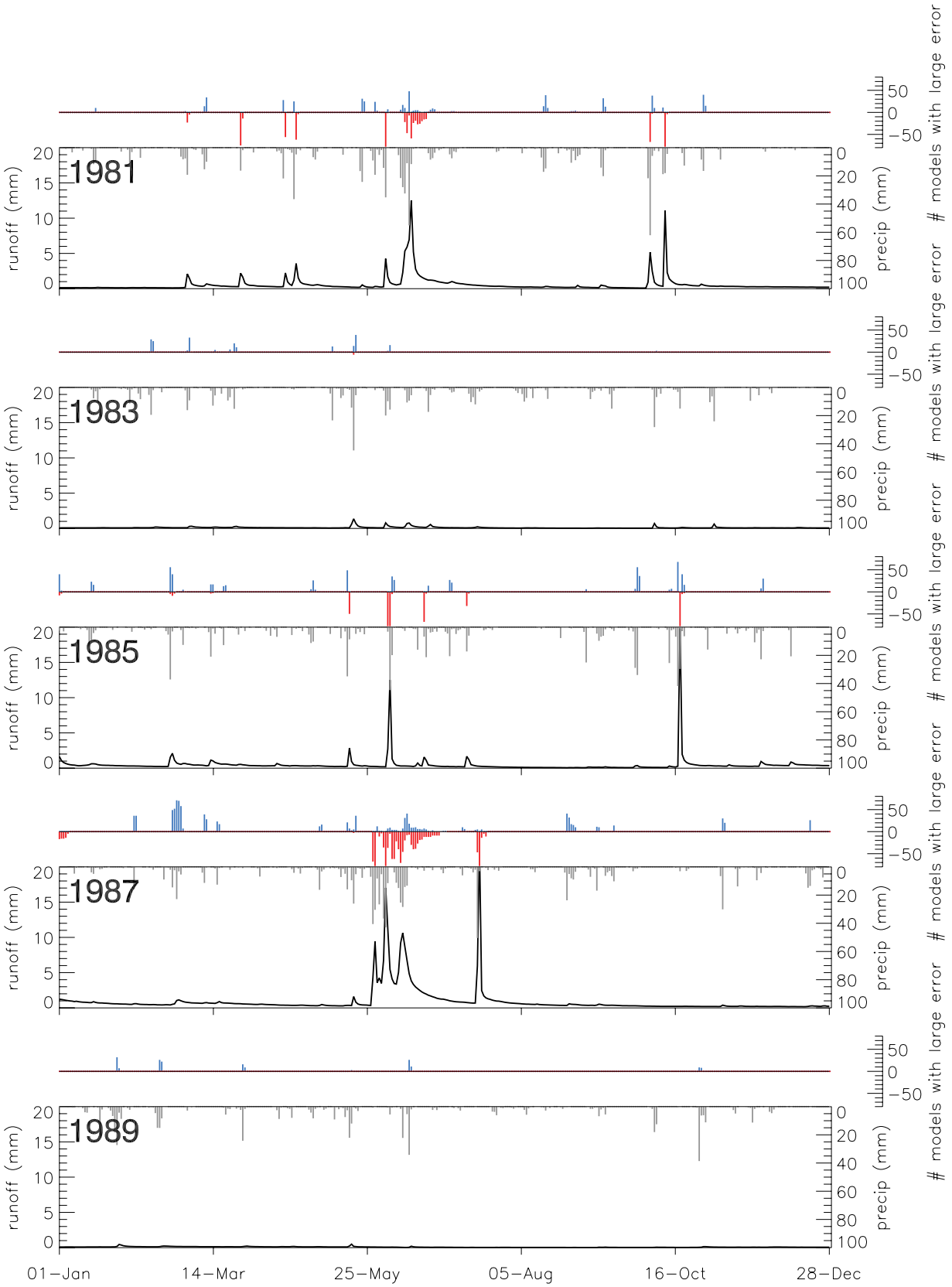

Figure 6. Time series of observed precipitation and observed runoff for every second year in the 11-year calibration period (1980–1990) for the Guadalupe River basin. The colored bar charts above each plot denote the number of models (out of 79) with large errors in each time step: red bars denote models that underestimate runoff (residual < -1 mm d^{-1}), and blue bars denote models that overestimate runoff (residual > 1 mm d^{-1}).

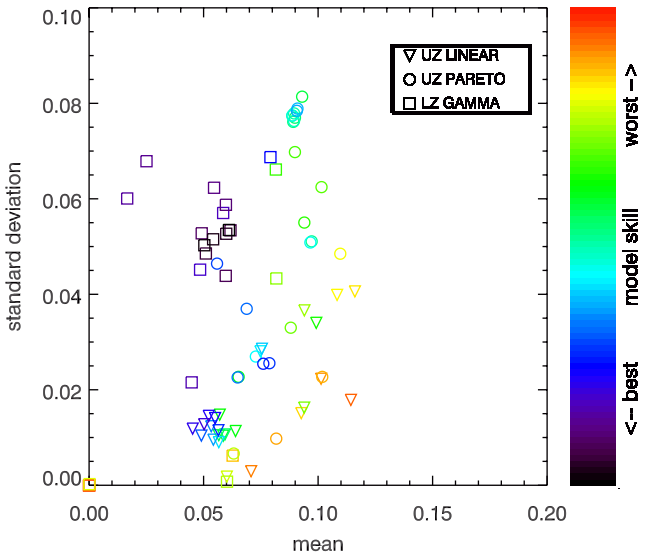

Figure 7. Mean and standard deviation of the saturated area time series for the Guadalupe River basin, as simulated by each of the 79 models. The color coding defines model skill, and the symbols identify the parameterization that was used to simulate saturated areas: UZ LINEAR refers to equation (9a) (PRMS); UZ PARETO refers to equation (9b) (ARNO-VIC); and LZ GAMMA refers to equation (9c) (TOPMODEL). The labels PARETO and GAMMA refer to the parametric distributions used in equations (9b) and (9c), respectively.

diagnose relations between model performance and structure across a diverse range of river basins.

[66] The differences in model performance in the Guadalupe River basin (for high-flow events) are related at least in part to the choice of parameterization for saturated areas. The models that have the highest Nash-Sutcliffe score use surface runoff parameterizations in which saturated areas are controlled by water storage in the lower soil layer. Variability in saturated areas in these parameterizations is much more damped than in other parameterizations. However, there are no data to test if the modeled saturated area dynamics are realistic for the Guadalupe River basin. In this context it is worth noting that no models have mechanisms for generating infiltration-excess runoff, and the saturated-area parameterizations may compensate for this model weakness. Future work in well-instrumented basins is also necessary to evaluate the capabilities of different models to simulate hydrological states and fluxes at internal points in the river basin (i.e., are we getting the right answers for the right reasons?).

[67] Diagnosis of model errors in the Guadalupe River basin shows that the multimodel response to large precipitation events is quite consistent in that most models either overestimate or underestimate measured streamflow. At first glance this suggests there is actually limited potential to use outputs from our multimodel configuration as an estimate of model uncertainty. Ideally, simulations of streamflow from different model structures will bracket the observed streamflow [e.g., *Butts et al.*, 2004; *Georgakakos et al.*, 2004; *Vrugt and Robinson*, 2007]; when this does not occur (as in this study) it can be viewed as being indicative of a lack of independent information in different models. However, the consistency in model errors may also arise because errors in model inputs affect different models in similar ways. Moreover, the ensemble may not bracket the measurements because all models have similar weaknesses (e.g., no mechanisms for generating infiltration-excess runoff, no vegetation submodel, no representation of the spatial variability in precipitation). Future work is also necessary to both separate errors in model inputs from errors in model structure [e.g., *Clark and Slater*, 2006; *Kavetski et al.*, 2006a, 2006b; J. A. Vrugt et al., Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov Chain Monte Carlo simulation, submitted to *Water Resources Research*, 2007], and rigorously quantify the independence between different models [e.g., *Abramowitz and Gupta*, 2008].

[68] The problems of identifying the most appropriate model structure for a given problem and quantifying the uncertainty in model structure remain outstanding research challenges for the discipline of hydrology. We did not seek to fully resolve these problems in this paper. We have presented the differences between models using consistent notation, and introduced FUSE so as to evaluate model differences in a controlled way. Our initial application of FUSE exposed a relationship between model performance and model structure for the drier of the two basins examined, suggesting that the choice of model structure is just as important as the choice of model parameters. New understanding of model structural differences can be obtained by addressing the following questions.

[69] 1. What are the differences between model structures when they are evaluated using multiple criteria? Do some model structures reproduce different parts of the hydrograph (or different internal states) better than others? Is the trade-off between objective functions different in different model structures?

[70] 2. Are relations between model performance and structure consistent between different river basins? Is it possible to identify model structures that are best suited to different climate regimes?

[71] 3. Do model simulations from different model structures mimic reality (are we getting the right answers for the right reasons)? To what extent can data from well-instrumented basins be used to identify model structures that produce credible simulations of hydrological states and fluxes?

[72] 4. How much independent information is in different model structures? Can model structures be designed to maximize the information content in (and value of) the multimodel ensemble?

[73] These questions pose staunch challenges to the hydrologic modeling community, but progress on these matters will surely aid our predictive abilities. To facilitate research on these problems, the FORTRAN-90 source code for FUSE is available upon request from the lead author.

Appendix A: Computational Scheme

[74] Model equations are solved using an implicit scheme with adaptive substeps. The implicit scheme

$$\mathbf{S}_{n+1} = \mathbf{S}_n + \frac{d\mathbf{S}'_{n+1}}{dt}\Delta t \qquad \text{(A1)}$$

requires iteratively computing model fluxes for the vector of states $\mathbf{S}'_{n+1}$ until $\mathbf{S}_{n+1} = \mathbf{S}'_{n+1}$ (within some error tolerance). Equation (A1) is solved using the Newton-Raphson method combined with line searches, as detailed by *Press et al.* [1992]. New state vectors in the Newton-Raphson method are based on the Jacobian matrix defining partial derivatives of the function $\mathbf{F} = |\mathbf{S}_{n+1} - \mathbf{S}'_{n+1}|$ with respect to model states $\mathbf{S}$

$$J_{ij} = \frac{\partial F_i}{\partial S_j} \tag{A2}$$

where all derivatives in J are computed numerically. Step size is controlled by comparing the implicit solution using two time steps with the implicit solution in a single time step. A new step size is introduced according to

$$\Delta t' = s\,\Delta t \sqrt{\frac{\varepsilon_{crit}}{D_{\max}}} \tag{A3}$$

where s is a safety factor ($s = 0.9$), ε_{crit} is the critical threshold for reducing/increasing the step size ($\varepsilon_{crit} = 0.001$), and $D_{\max}$ is the maximum absolute difference in fractional model states between the one-step and two-step solutions. The use of the implicit solution requires many function evaluations to compute the Jacobian matrix, especially in models with many state variables, but it is straightforward to use this method to solve equations from multiple model combinations.

[75] **Acknowledgments.** We are grateful to Yun Duan for information on the MOPEX experiment, to George Leavesley for information on the intricacies of the PRMS model, and to Hilary McMillan for comments on an earlier draft of this manuscript. We are also indebted to three anonymous referees for their insightful comments. This research was funded by the New Zealand Foundation for Research Science and Technology (contract C01X0401), the National Aeronautic and Space Administration (contract NNG06GH10G), and the National Oceanic and Atmospheric Administration (contract NA06OAR4310065).

References

Abramowitz, G., and H. Gupta (2008), Toward a model space and model independence metric, *Geophys. Res. Lett.*, *35*, L05705, doi:10.1029/2007GL032834.

Ambroise, B., K. Beven, and J. Freer (1996), Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity, *Water Resour. Res.*, *32*, 2135–2145, doi:10.1029/95WR03716.

Anderson, E. A. (1973), National Weather Service River Forecast System—Snow accumulation and ablation model, *Tech. Memo. NWS HYDRO-17*, NOAA, Silver Spring, Md., Nov.

Atkinson, S. E., R. A. Woods, and M. Sivapalan (2002), Climate and landscape controls on water balance model complexity over changing timescales, *Water Resour. Res.*, *38*(12), 1314, doi:10.1029/2002WR001487.

Beven, K. (1997), TOPMODEL: A critique, *Hydrol. Processes*, *11*, 1069–1085, doi:10.1002/(SICI)1099-1085(199707)11:9<1069::AID-HYP545>3.0.CO;2-O.

Beven, K. (2002), Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, *Hydrol. Processes*, *16*, 189–206, doi:10.1002/hyp.343.

Beven, K. J., and M. J. Kirkby (1979), A physically based, variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*, 43–69.

Boone, A., and P. J. Wetzel (1996), Issues related to low resolution modeling of soil moisture: Experience with the PLACE model, *Global Planet. Change*, *13*, 161–181, doi:10.1016/0921-8181(95)00044-5.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*, 3663–3674, doi:10.1029/2000WR900207.

Burnash, R. J. C. (1995), The NWS River Forecast System—Catchment modeling, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 311–366, Water Resour. Publ., Littleton, Colo.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, technical report, U.S. Natl. Weather Serv., Sacramento, Calif.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266, doi:10.1016/j.jhydrol.2004.03.042.

Clark, M. P., and A. G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeorol.*, *7*, 3–22, doi:10.1175/JHM474.1.

Desborough, C. E. (1997), The impact of root weighting on the response of transpiration to moisture stress in land surface schemes, *Mon. Weather Rev.*, *125*, 1920–1930, doi:10.1175/1520-0493(1997)125<1920:TIORWO>2.0.CO;2.

Desborough, C. E. (1999), Surface energy balance complexity in GCM land surface models, *Clim. Dyn.*, *15*, 389–403, doi:10.1007/s003820050289.

Duan, J. F., and N. L. Miller (1997), A generalized power function for the subsurface transmissivity profile in TOPMODEL, *Water Resour. Res.*, *33*, 2559–2562, doi:10.1029/97WR02186.

Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*, 1015–1031, doi:10.1029/91WR02985.

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.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*, 222–241, doi:10.1016/j.jhydrol.2004.03.037.

Guntner, A., S. Uhlenbrook, J. Seibert, and C. Leibundgut (1999), Multicriterial validation of TOPMODEL in a mountainous catchment, *Hydrol. Processes*, *13*, 1603–1620, doi:10.1002/(SICI)1099-1085(19990815)13:11<1603::AID-HYP830>3.0.CO;2-K.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, in press.

Henderson-Sellers, A., Z. L. Yang, and R. E. Dickinson (1993), The Project for Intercomparison of Land-surface Schemes (PILPS), *Bull. Am. Meteorol. Soc.*, *74*, 1335–1349, doi:10.1175/1520-0477(1993)074<1335:TPFIOL>2.0.CO;2.

Iorgulescu, I., and A. Musy (1997), Generalization of TOPMODEL for a power law transmissivity profile, *Hydrol. Processes*, *11*, 1353–1355, doi:10.1002/(SICI)1099-1085(199707)11:9<1353::AID-HYP585>3.0.CO;2-U.

Kampf, S. K., and S. J. Burges (2007), A framework for classifying and comparing distributed hillslope and catchment hydrologic models, *Water Resour. Res.*, *43*, W05423, doi:10.1029/2006WR005370.

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.

Kavetski, D., G. Kuczera, and S. W. Franks (2003), Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, *Water Resour. Res.*, *39*(9), 1246, doi:10.1029/2003WR002122.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.

Koren, V., S. Reed, M. Smith, Z. Zhang, and D.-J. Seo (2004), Hydrology Laboratory Research Modeling System (HL-RMS) of the US National Weather Service, *J. Hydrol.*, *291*, 297–318, doi:10.1016/j.jhydrol.2003.12.039.

Koster, R. D., and P. C. D. Milly (1997), The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models, *J. Clim.*, *10*, 1578–1591, doi:10.1175/1520-0442(1997)010<1578:TIBTAR>2.0.CO;2.

Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon (1983), Precipitation-runoff modeling system: User's manual, *U.S. Geol. Surv. Water Invest. Rep.*, *83-4238*, 207 pp.

Leavesley, G. H., P. J. Restrepo, S. L. Markstrom, M. Dixon, and L. G. Stannard (1996), The modular modeling system—MMS: User's manual, *U.S. Geol. Surv. Open File Rep.*, *96-151*, 142 pp.

Leavesley, G. H., S. L. Markstrom, P. J. Restrepo, and R. J. Viger (2002), A modular approach for addressing model design, scale, and parameter

estimation issues in distributed hydrological modeling, *Hydrol. Processes*, *16*, 173–187, doi:10.1002/hyp.344.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994), A simple hydrologically based model of land surface water and energy fluxes for general-circulation models, *J. Geophys. Res.*, *99*(D7), 14,415–14,428, doi:10.1029/94JD00483.

Luo, L. F., et al. (2003), Effects of frozen soil on soil temperature, spring infiltration, and runoff: Results from the PILPS 2(d) experiment at Valdai, Russia, *J. Hydrometeorol.*, *4*, 334–351, doi:10.1175/1525-7541(2003)4<334:EOFSOS>2.0.CO;2.

Milly, P. C. D. (1994), Climate, soil-water storage, and the average annual water balance, *Water Resour. Res.*, *30*, 2143–2156, doi:10.1029/94WR00586.

Milly, P. C. D., and K. A. Dunne (2002), Macroscale water fluxes: 2. Water and energy supply control of their interannual variability, *Water Resour. Res.*, *38*(10), 1206, doi:10.1029/2001WR000760.

Parada, L. M., J. P. Fram, and X. Liang (2003), Multi-resolution calibration methodology for hydrologic models: Application to a sub-humid catchment, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 197–211, AGU, Washington, D. C.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1992), *Numerical Recipes: The Art of Scientific Computing*, 2nd ed., Cambridge Univ. Press, Cambridge, U. K.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, and D.-J. Seo (2004), Overall distributed model intercomparison project results, *J. Hydrol.*, *298*, 27–60, doi:10.1016/j.jhydrol.2004.03.031.

Rupp, D. E., and R. A. Woods (2008), Increased flexibility in base flow modelling using a power law transmissivity profile, *Hydrol. Processes*, *22*, 2667–2671, doi:10.1002/hyp.6863.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*, 2075–2080, doi:10.1002/hyp.6825.

Singh, V. P., and D. A. Woolhiser (2002), Mathematical modeling of watershed hydrology, *J. Hydrol. Eng.*, *7*, 270–292, doi:10.1061/(ASCE)1084-0699(2002)7:4(270).

Sivapalan, M., K. Beven, and E. F. Wood (1987), On hydrologic similarity: 2. A scaled model of storm runoff production, *Water Resour. Res.*, *23*, 2266–2278, doi:10.1029/WR023i012p02266.

Slater, A. G., et al. (2001), The representation of snow in land surface schemes: Results from PILPS 2(d), *J. Hydrometeorol.*, *2*, 7–25, doi:10.1175/1525-7541(2001)002<0007:TROSIL>2.0.CO;2.

Todini, E. (1996), The ARNO rainfall-runoff model, *J. Hydrol.*, *175*, 339–382, doi:10.1016/S0022-1694(96)80016-3.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, *Water Resour. Res.*, *44*, W01429, doi:10.1029/2007WR006271.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Wetzel, P. J., X. Liang, P. Irannejad, A. Boone, J. Noilhan, Y. P. Shao, C. Skelly, Y. K. Xue, and Z.-L. Yang (1996), Modeling vadose zone liquid water fluxes: Infiltration, runoff, drainage, interflow, *Global Planet. Change*, *13*, 57–71, doi:10.1016/0921-8181(95)00037-2.

Wood, E. F., D. P. Lettenmaier, and V. G. Zartarian (1992), A land-surface hydrology parameterization with subgrid variability for general-circulation models, *J. Geophys. Res.*, *97*(D3), 2717–2728.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), Toward improved distributed modeling of watersheds: A process-based diagnostic approach to model evaluation, *Water Resour. Res.*, doi:10.1029/2007WR006716, in press.

Zhao, R. J. (1977), *Flood Forecasting Method for Humid Regions of China*, East China Coll. of Hydraul. Eng., Nanjing, China.

Zhao, R. J. (1984), *Watershed Hydrological Modelling*, Water Resour. and Electr. Power Press, Beijing.

M. P. Clark and R. A. Woods, NIWA, P.O. Box 8602, Riccarton, Christchurch, New Zealand. (mp.clark@niwa.co.nz)

H. V. Gupta, Department of Hydrology and Water Resources, University of Arizona, P.O. Box 210011, 1133 East North Campus, Room 318, Tucson, AZ 85721, USA.

L. E. Hay, U.S. Geological Survey, Lakewood, CO 80225, USA.

D. E. Rupp, DHI Water and Environment, Inc., 319 SW Washington Street, Suite 614, Portland, OR 97204, USA.

A. G. Slater, CIRES, University of Colorado, Campus Box 449, Boulder, CO 80309-0449, USA.

J. A. Vrugt, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

T. Wagener, Department of Civil and Environmental Engineering, Pennsylvania State University, 226B Sackett Building, University Park, PA 16802, USA.