A Parsimonious Hydrological Model for a Data Scarce Dryland Region

**Saket Pande · Hubert H. G. Savenije ·
Luis A. Bastidas · Ashvin K. Gosain**

Abstract Inapplicability of state of the art hydrological models due to scarce data motivates the need for a modeling approach that can be well constrained to available data and still model the dominant processes. Such an approach requires embedded model relationships to be simple and parsimonious in parameters for robust model selection. Simplicity in functional relationship is also important from water management point of view if these models are to be coupled with economic system models for meaningful policy assessment. We propose a similar approach, but rather than selecting (through calibration) processes from a set of processes predefined in terms of functionalities or modules, we model already known dominant processes in dryland areas (evaporation, Hortonian overland flows, transmission loses and subsurface flows) in a simple manner by explicitly programming them as constraints and obtain parameters by minimizing a performance based objective function. Such use of mathematical programming allows flexible model calibration and simulation in terms of available data and constraints. The model results of the approach are however not perfect given its infancy. Nonetheless its imperfections can guide us to further improvements, in particular with regards to model structure improvement.

Keywords Water management · Parsimonious modeling · Dryland areas · India · Mathematical programming

S. Pande (✉) · H. H. G. Savenije
Hydrology Division, Faculty of Civil Engineering and Geosciences,
Delft University of Technology, Delft, The Netherlands
e-mail: s.pande@tudelft.nl

L. A. Bastidas
Department of Civil and Environmental Engineering and Utah Water Research Laboratory,
Utah State University, Logan, UT, USA

A. K. Gosain
Department of Civil Engineering, Indian Institute of Technology, New Delhi, India

1 Introduction

Inapplicability of state of the art hydrological models due to scarce data motivates the need for a modeling approach that can be well constrained to available data and still model the dominant processes. Such an approach requires embedded model relationships to be simple and parsimonious in parameters, i.e. low in model complexity, for robust model selection (Pande et al. 2009). Simplicity in functional relationship is also important from water management point of view if these models are to be coupled with economic system models for meaningful policy assessment. Flexible modeling approaches (Fenicia et al. 2008; Clark et al. 2008; Savenije 2010) that propose to model processes relevant to application exist, yet they are far from being used in specific water management problems.

We propose a similar approach, but rather than selecting (through calibration) processes from a set of processes predefined in terms of functionalities or modules, we model already known dominant processes (Pilgrim et al. 1988) in dryland areas (evaporation, Hortonian overland flows, transmission loses and subsurface flows) in a simple manner by mathematical programming them as constraints. Use of mathematical programming allows flexible model calibration and simulation in terms of available data and constraints. Model state variables are (implicitly) solved and parameters are estimated simultaneously under a set of constraints imposed on maximum allowable storage and evaporation.

However the motivation of this work is not to present a flexible model but to propose a mathematical programming based parsimonious model. Mathematical programming based models have been prevalent in hydro-economic modeling literature (Yu et al. 2003; Harou et al. 2009; Maneta et al. 2009) with few applications in hydrological modeling (such as Sun et al. (2010) but without any spatial or temporal dynamics in estimating equations). This modeling approach enables a flexible structure as water and energy balance, hydrological connectivity and other constraints can be explicitly entered as constraints. The structure of mass balance constraints is flexible in the sense that such constraints can be programmed as a (sum) result of various subsuming mass balances (the latter representing different process conceptualizations). Energy balances can similarly be modeled. Note that we are not primarily advocating a flexible model structure since mathematical programming based approach also enables solution of hydrologic variables subject to these simultaneous constraints for arbitrary model parameters. By minimizing a model performance based objective function with respect to the parameters of the model and subject to the constraints; optimal parameter values with solution for hydrologic variables (such that constraints are met) are obtained. The approach is different from "traditional" flexible hydrologic models in explicit representation of constraints, including spatially coupled mass balances, which are solved to obtain state variables of the system. Parameters are then optimized based on ensuing model performance. We use MINOS5 DNLP (discrete nonlinear programming) solver (GAMS 2008) to solve coupled constraints and optimize parameters. Yet another advantage of using such an approach lies in its application to water resource management. Mathematical programming based model structure can be plugged in with ease into economic system models that model various commodity balances as constraints to decision making (and water is one such commodity).

Model structure can also be adapted to yield a robust model structure by explicitly making mass balances more or less complex based on model performance and constrained by data availability. The Langrange multipliers of various constraints can guide model structure selection. A robust model structure is such that model performance is robust under different climatic forcing instantiation from the underlying data generating processes. Explicit use of how binding various components of a model structure are (via Langrange multipliers) to model performance is not sufficiently considered in current modeling practice. A mathematical programming based modeling approach can enable such a unique model selection approach.

The motivation thus goes beyond presentation of a flexible model, it is motivated by the need to have an approach that not only enables a flexible model structure but also allows explicit representation of mass balance difference equations and other constraints, yielding a model structure that is more consistent with the structure of economic system models, as well as provides a platform for innovative new approaches to robust model selection (such as Sun et al. 2010).

Reliable hydrological models approximate the underlying set of processes as best as "data-possible". Models are conceptualizations (Savenije 2009) and need to be evaluated with observations before selecting the best available conceptualization. A model that conceptualizes the processes better (low "approximation" error) is possibly more complex (Cucker and Smale 2002). However, models selected from a set of complex models (in terms of number of parameters) are generally uncertain in prediction (high "estimation" error). Thus, there is an estimation/approximation error trade-off in model selection.

This estimation/approximation error trade-off calls for a balance between complexity and process representation. The modelling approach here attempts to strike such a balance, by modelling dominant processes using simple functional relationships, with explicit constraints and minimal set of parameters.

There exist several dryland hydrological models that have been applied at different time and spatial scales (Michaud and Sorooshian 1994; Beven 2002; Cudennec et al. 2007 and references within). See also Rodriguez-Iturbe and Porporato (2004), Dunkerley (2000), Eagleson (1982) for modeling water limited systems. The model presented here is similar in spirit to semi-distributed IHACRES (McIntyre and Al-Qurashi 2009). Subbasin representation allows spatial heterogeneities though at a cost of increased parameterization. Yet another unique feature of this approach is that it introduces parsimony in parameters by parameter regionalization (Vogel 2006).

We model the seasonal water balance at monthly time scale. We assume that inter-annual variation in monthly values of state variables and fluxes is negligible and that seasonal cycle repeats itself every year. In order to parameterize this conceptualization, we use annual mean monthly values of all time series data (GRACE, MERRA, and mean, minimum and maximum temperature) by averaging values over the years for each month. Yearly repetition of the seasonal cycle of storages is a cyclostationarity condition, which we explicitly incorporate by forcing the storage level in the 13th month to be storage level in the first month for all subbasins.

The model results are however not perfect given the infancy of the approach. Nonetheless its imperfections can guide us to further improvements, in particular with regards to model structure improvement. The remainder of the paper is

organized as follows. Section 2 describes the modeling approach. Section 3 describes the study area and the data set used. Section 4 presents the results. Section 5 discusses the results and concludes the paper.

2 Methodology

Figure 1 shows model conceptualization. Consider the set of interconnected sub-basins in the study area, where a set of basins upstream to a subbasin i, indexed by ℓ, are defined by the set $U(i)$. The stream channels connect various subbasins and these "connections" are obtained by delineating DEM (digital elevation model) of a study area. Storage in each subbasin can be conceptualized as the sum of subsurface and surface water storage.

Actual evaporation for each store i (see Fig. 1), at monthly steps, is conceptualized as the sum of a fraction, F_{ce}, of residual rainfall $w_{i,t}$, a fraction, $F_{ce0} * \Theta_i$, of storage and irrigation applied (u_{it}). Thus, evaporation is limited by water stress where a fraction of rainfall contributes to infiltration (Rienecker et al. 2008). The evaporation fraction of soil moisture conceptualizes that some of the available water storage
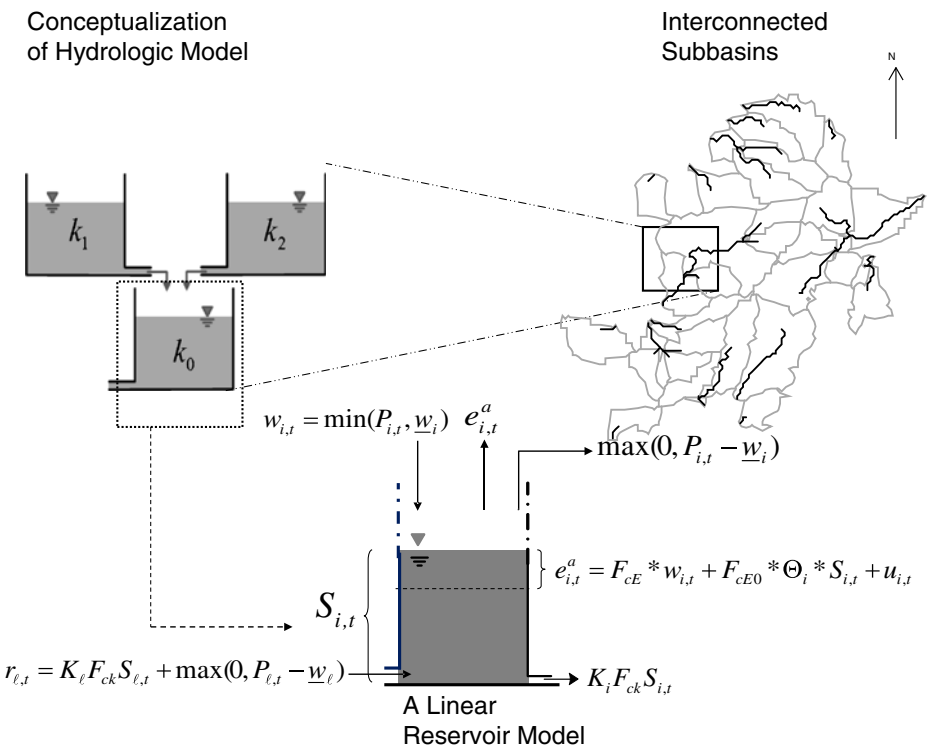

Fig. 1 Each of the interconnected subbasins is represented by a reservoir with a parameter set k_i, which is interconnected to reservoirs downstream. A reservoir models overland flow, subsurface flow, and evaporation while the storage $S_{i,t}$ is updated each time step based on mass balance

also contributes to soil evaporation. The fraction is assumed to be proportional to estimated field capacity (Θ_i) (Li and Sivapalan 2010).

Evaporation demand from land cover is calculated based on FAO guidelines (FAO 2003) wherein reference transpiration is calculated using Hargreaves equation (Equation 8 of Hargreaves and Allen 2003), corrected by crop coefficients, to obtain evaporation demand. Fifteen major crops (or crop types) are identified that are grown over the two growing seasons (rabi and kharif seasons). Proportions of cropland cover (Roy et al. 2003) are multiplied by district specific (administrative level crop area data, see Section 3 for agricultural census data) crop area proportions to obtain crop type specific land cover fractions. Remaining landcover types are identified (Roy et al. 2003) as grassland cover, shrubland cover and bareland cover. Growing period as well as landcover specific crop coefficients are obtained based on FAO guidelines and multiplied with reference transpiration calculated using Hargreaves equation to obtain landcover specific evaporation demand. Subbasin level aggregate monthly evaporation demand is obtained by summation of landcover specific evaporation demands weighted by respective landcover area fractions.

Subsurface flow is conceptualized as a linear function of water storage (Pulido-Velazquez et al. 2007; Clark et al. 2009). The storage-flow parameter, k_i, is a linear function of hydraulic conductivity K_i via parameter F_{ck} (Franchini et al. 1996). The overland flow at an aggregated monthly time scale, in the event when rainfall rate exceeds infiltration rate, is conceptualized by a threshold function $\max(0, \ P_{i,t} - \underline{w}_i)$. The threshold, $\underline{w}_i$, should be treated (with caution) as a function of both stochasticity of precipitation events in months when overland flow occurs as well as subbasin specific infiltration rate (that depend on soil properties).

Model equations are given by the following constraints:
(overland flow event)

$$w_{i,t} = \min\left(P_{i,t}, \ \underline{w}_i\right) \tag{1}$$

(actual evaporation for i^{th} subbasin at time t)

$$e_{i,t}^a = F_{ce}w_{i,t} + F_{ce0}\Theta_i S_{i,t} + u_{i,t} \tag{2}$$

(j land use specific evaporation demand at time t)

$$e_{i,t}^d = k_{i,j,t}^c E_{i,t}^0 \tag{3}$$

(i specific irrigation applied at time t)

$$u_{i,t} = \eta E_{i,t}^0 a_i^{irri} \tag{4}$$

(water balance for i^{th} subbasin at time t)

$$e_{i,t}^a - w_{i,t} = S_{i,t} - S_{i,t+1} + \sum_{\ell \in U(i)} r_{l,t} - r_{i,t}. \tag{5}$$

(constraint on actual evaporation)

$$e_{i,t}^a \leq \sum_j a_{ij} e_{j,t}^d \tag{6}$$

(subsurface flow equation from subbasin i to its downstream subbasins)

$$r_{i,t} = K_i F_{ck} S_{i,t} + \max\left(0, \ P_{i,t} - \underline{w}_i\right) \tag{7}$$

(T-period steady state constraint)

$$S_{i,T+1} = S_{i,1} \tag{8}$$

The parameters of the model are F_{ce}, F_{ce0}, F_{ck}, η and $\underline{w}_i$ for $i = 1, .., N$, where N is the number of subbasins. These parameters are introduced alongside other variables in Table 1.

Constraints (1–8) constitute a mathematical program, which can be solved for hydrologic variables $\{S_{i,t}, r_{i,t}, u_{i,t}, w_{i,t}, e^d_{j,t}, e^a_{i,t}\}$ given particular values of parameters $\{F_{ce}, F_{ce0}, F_{ck}, \eta$ and $\underline{w}_i\}$ and coefficients $\{K_i, k^c_{i,j,t}, a^{irri}_i, P_{i,t}, a_{ij}\}$ for $i = 1,..,N$. $j = 1,..,J$ and $t = 1,..,T$.

Let $\Delta \hat{S}_{i,t}$ and $\hat{e}^a_{i,t}$ represent observed monthly storage change and actual evaporation for subbasin i and month $t = 1, .., T$ (T is the number of months considered,

Table 1 Description of parameters (to estimate), variables, coefficients and indices used in the model

Symbol [units]	Description
Parameters	
F_{ce} [–]	Fraction of residual rainfall that directly evaporates
F_{ce0} [1/month]	Multiplier on Θ_i, fraction of storage that evaporates
F_{ck} [h/(in*month)]	Multiplier on K_i, fraction of storage contributing to slow flow
η [–]	Multiplier on $E^0_{i,t} a^{irri}_i$, fraction of maximum irrigation demand
$\underline{w}_i$ [mm/month]	Hortonian overland flow threshold parameter, rainfall above this threshold is conceptualized as overland flow contribution
Coefficients	
K_i [in/h]	Approximate hydraulic conductivity
$k^c_{i,j,t}$ [–]	Crop coefficients based on FAO guidelines for j^{th} landcover type in i^{th} subbasin and month t
a^{irri}_i [–]	Fraction of area irrigated in the ith subbasin
$P_{i,t}$ [mm/month]	Monthly rainfall for the ith subbasin in month t
a_{ij} [–]	Fraction of ith subbasin covered by j^{th} landcover type
Variables	
$S_{i,t}$ [mm]	Store levels in subbasin i and month t
$r_{i,t}$ [mm/month]	Slow flow out from i^{th} reservoir in month t
$E^0_{i,t}$ [mm/month]	Reference evaporation calculated using 1985 Hargreaves equation for i^{th} subbasin in month t
$u_{i,t}$ [mm/month]	Estimated irrigation demand for i^{th} subbasin in month t
$w_{i,t}$ [mm/month]	Residual rainfall in i^{th} subbasin and month t after subtracting Hortonian overland flow
$e^d_{j,t}$ [mm/month]	Evaporation demand of jth crop in month t.
$e^a_{i,t}$ [mm/month]	Actual total evaporation from i^{th} subbasin in month t
Indices	
$U(i)$	Set of subbasins upstream to the ith subbasin
i	Subbasin index, $\{1,...,N\}$
j	Landcover index, $\{1,...,J\}$
t	Month index, $\{1,...,T\}$
N	Number of subbasins, 34 in this study
J	Number of land cover types, 18 in this study (15 croptypes, 3 other land cover types)
T	Number of months, 12 in this study

which is 12 since annual cyclo-stationarity is assumed). Consider two residual definitions, corresponding to storage change and evaporation,

$$\Delta \hat{S}_{i,t} - \left(S_{i,t+1} - S_{i,t}\right) = \xi_{i,t} \tag{9}$$

$$\hat{e}_{i,t}^a - e_{i,t}^a = \zeta_{i,t}. \tag{10}$$

Then the solution of the following program yields parameters of the model such that equally weighted mean absolute error in model performance with respect to storage change and evaporation is minimized,

$\min_{F_{ce}, F_{ce0}, F_{ck}, \eta, \underline{w}_i, S_{i,t}} \left(0.5\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}|\xi_{i,t}| + 0.5\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}|\zeta_{i,t}|\right)/(NT)$ subject to constraints (1–10).

The mathematical program is computed and the model parameters are estimated using MINOS5 Discrete Nonlinear Programming (DNLP) solver in the GAMS environment (GAMS 2008). The relevant data sets used are explained in Section 3.2.

Model conceptualization for dryland areas presented here is similar to semi-distributed IHACRES (McIntyre and Al-Qurashi 2009) but at monthly scale. IHACRES models effective rainfall based on antecedent moisture conditions, rainfall and a certain loss parameter and routes it through a sub-basin using a linear reservoir (where in a semi-distributed version of IHACRES model, a basin is represented as a set of interconnected sub-basins). We conceptualize effective rainfall via a linear reservoir with a threshold (Eq. 1) and model evaporation flux explicitly (Eq. 5). Interception effect on effective rainfall is important but has been modeled implicitly with evaporation and transpiration of surface soil water as suggested by Pilgrim et al. (1988) (see Eq. 2). The model runs at monthly time steps in which flow volumes are estimated (Eq. 7). Significant transmission losses occur in overland flows during intense precipitation events, contributing to groundwater (Pilgrim et al. 1988; Sharma and Murthy 1998) (see Eqs. 1 and 7 where flow from upstream areas appear as $r_{\ell,t}$). Flow volumes in Eq. 7 are modeled at monthly scale without any need for flow routing as transmission loses and overland flows occur at much shorter time scales. Further, given scarcity of data and issues of non-identifiability (McIntyre and Al-Qurashi 2009), these two flows are conceptualized in a simple yet physically meaningful manner. An important element in modeling storage change is irrigation which is also included in a mass balance equation in Eq. 5 as $u_{i,t}$ (defined in Eqs. 2 and 4). Model conceptualization is however not complicated further, even though it underperforms in modeling storage changes, for reasons of non-identifiability (McIntyre and Al-Qurashi 2009).

3 Data and Study Area

3.1 Study Area

The study area, in Fig. 2, comprises the arid/semiarid states of Gujarat and Rajasthan in India, with an area of 538,346 km², minimum (maximum) average monthly precipitation and temperature of approx. 2 mm/month (202 mm/month) and 17°C (33°C) respectively. The elevation ranges from 0 to 1,350 m with average slope along
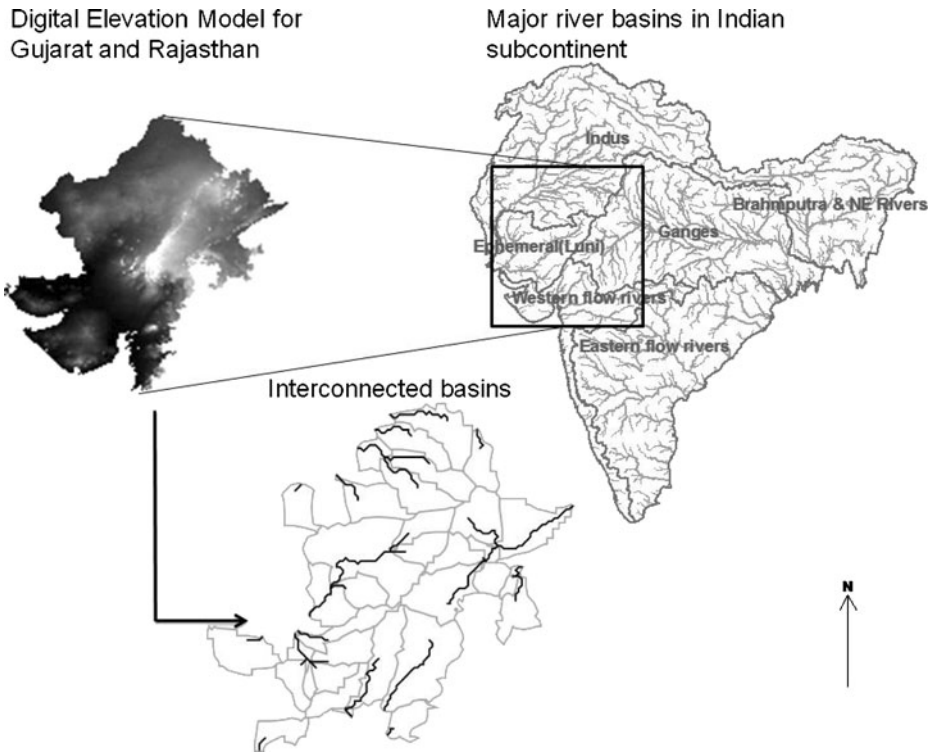
Fig. 2 Study area: Gujarat and Rajasthan in western India. Channels shown in *black* under interconnected basins

the drainage lines of 0.06° at a spatial resolution of 8 × 8 km. The soil is sandy (northwest) to loamy-clay (south-east).

3.2 Data

The Global 30 arc-second Elevation Data Set (GTOPO30), is used for the study area (*available from the U.S. Geological Survey*) to obtain subbasin delineation. Gravity Recovery and Climate Experiment (GRACE) data (RL04) from August 2002–August 2008, excluding June 2003 and January 2004, over land at 1° resolution is used for monthly changes in water storage, smoothed using Guassian averaging kernel with halfwidth of 300 km (Chambers 2006). GRACE terrestrial water storage change data has been extensively used in various hydrological studies (Tiwari et al. 2009; Sheffield et al. 2009; Gulden et al. 2007; Niu et al. 2007; Yeh et al. 2006; Syed et al. 2005) and recently has been used as a primary data source for model calibration (Lo et al. 2010; Sun et al. 2010; Werth et al. 2009). The Modern Era Retrospective-analysis for Research and Applications (MERRA), a NASA reanalysis using Goddard Earth Observing System Data Assimilation System Version 5 (GEOS-5), is used to obtain monthly average of surface evaporation fluxes (Bosilovich 2008). The data set used is at $1/2 \times 2/3°$ (lat x lon) resolution. CRU TS2.1 data set (Mitchell and Jones 2005) is used for yearly mean monthly total rainfall estimates

as well as mean, minimum and maximum temperature grid data from 1940 to 2000 at ½ degree resolution. Mean monthly temperature and monthly minimum and maximum temperature is used to calculate monthly reference evaporation using "1985 Hargreaves equation" (Eq. 8, Hargreaves and Allen 2003).

We use gridded monthly temperature range and average temperature and assume constant temperature values during a month. Monthly reference evaporation is obtained by multiplying daily reference evaporation calculated using monthly average, minimum and maximum gridded temperature data in Hargreaves equation, and multiplying it by 30.

FAO's digital soil texture map (FAO 2003) at 5 min resolution is used. It is used to obtain saturated hydraulic conductivity and field capacity estimates. European Union Joint Research Center's 30 arc seconds land cover map for year the 2000 over South East Asia (Roy et al. 2003) is also used. Agricultural Census of India for the year 2000–2001 (agcensus.nic.in/cendata/databasehome.aspx) provides data on irrigated and non irrigated areas by crops grown. These two data sets are used to obtain subbasin area fractions for 15 crop cover types and other non-crop land cover types as well as to estimate subbasin area fractions of irrigated areas.

All spatial data sets are then re-sampled on a 8×8 km raster using a nearest neighbor method. Finally, basin-level average values of various variables are considered for later analysis.

The study area is delineated using the DEM and the D8 algorithm of ILWIS hydro-processing toolbox (ITC 2009) to obtain a map of interconnected subbasins. Streams are first identified based on upstream contributing area threshold of 6,400 km^2 (for a pixel to be on a drainage network). A drainage network and stream ordering is then obtained based on indentified streams and their intersections. Finally subbasins are extracted for each stream in the network based on the drainage network and flow direction map of the DEM.

Figure 3 below shows maps of field capacity and approximate hydraulic conductivity estimated from soil texture map (FAO 2003). Maps of these variables and variables appearing in Section 4 are created using mapping software developed at SOW-VU (Stichting Onderzoek Wereldvoedsel-voorziening van de Vrije Universiteit) (van den Boom and Pande 2007).

Each mapping unit of FAO soil map (FAO 2003) provides textural classes for dominant and component (associated soils and inclusions) soils for 30 cm depth. A maximum of 8 soil unit types (one dominant soil type and 7 other associated types), if applicable, are described within each soil mapping unit. There are 9 texture-slope classes associated with each soil unit. The texture classes are coarse (sands, loamy sands and sandy loams with less than 18% clay and more than 65% sand), medium (sandy loams, loams, sandy clay loams, silt loams, silt, silty clay loams and clay loams with less than 35% clay and less than 65% sand; the sand fraction may be as high as 82% if a minimum of 18% clay is present) and fine (clay, silty clays, sandy clays, clay loams, with more than 35% clay).

Soil characteristics used in study are obtained by first decomposing each soil mapping unit into fractions of three texture classes (coarse, medium, fine) by aggregating the fractions of a texture class across various soil types (dominant and associated). A representative value of hydraulic conductivity and field capacity is then calculated using Rawls and Cosby soil parameter table (source: http://www.emc.emc.ncep.noaa.gov/mmb/nldas/LDAS/validation.php). Representative hydraulic conductivity for a texture class is obtained as geometric mean of hydraulic conductivities
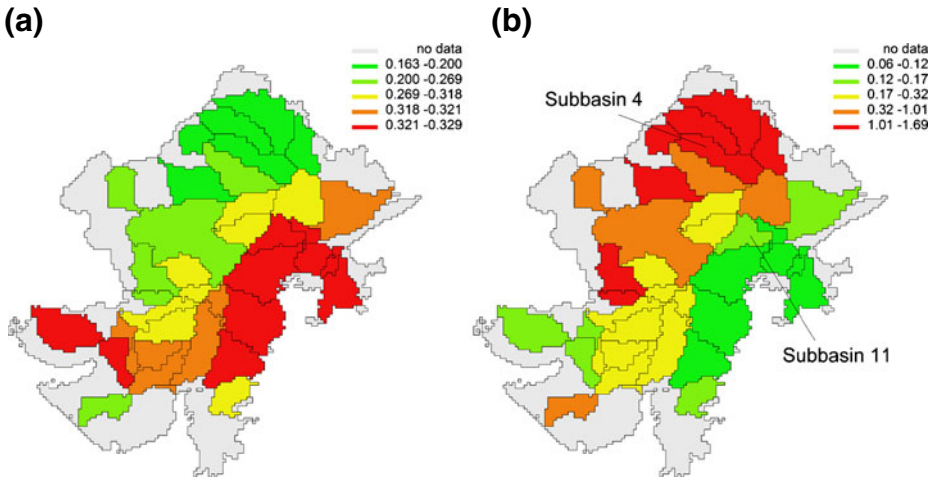
(a)

(b)

Fig. 3 **a** Estimated field capacity (–) and **b** hydraulic conductivity (in/h) from a soil texture map for the study area. Also shown are subbasins 4 and 11. These two subbasin are chosen as they lie across a hydrologic divide (Aravali hills) one flowing west (subbasin 4) while the other flowing east, both contributing to two different major basins in the Indian subcontinent. These two basins are representative of the performance of all other 32 subbasins and all have not been included for brevity reasons

of constituting texture types while field capacity is obtained as arithmetic mean of constituting field capacities. These representative soil parameter values are then resampled at 8 km × 8 km resolution. Finally basin level approximate conductivity and field capacity is obtained by calculating the geometric and arithmetic mean respectively.

The geometric mean (by area) of hydraulic conductivity classified by soil texture (Cosby et al. 1984; Rawls et al. 1982) data corresponding to each subbasin is used as its approximate saturated hydraulic conductivity while arithmetic mean (by area) of field capacity classified by soil texture is used for subbasin level field capacity. The former estimation is equivalent to calculating effective hydraulic conductivity assuming lognormally distributed (isotropic) conductivity field with correlation length significantly smaller than the spatial scale of the subbasins (Hoeksema and Kitanidas 1984; Seong 2002) while the latter estimation equivalently assumes spatial variation that is less malignant and secondary (Seong 2002).

4 Results

The model described in Section 2 is calibrated (parameters F_{ce}, F_{ce0}, F_{ck}, η and $\underline{w}_i$ are estimated) by minimizing equally weighted mean absolute errors (MAE). Note that we only model the seasonal cycle here, though extension to modeling monthly storages and fluxes over years is straightforward. One MAE measures the mean of the absolute deviation of modeled storage change to observed storages changes, averaged over 12 months (T = 12) and over all the subbasins. The second MAE similarly measures the mean of the absolute deviation between modeled and actual

evaporation (from reanalyzed MERRA evaporation data). At the obtained minima, equally weighted MAE value is 37.62 mm.

Figure 4 shows the modeled performance of storage changes and evaporation for two subbasins 4 and 11 (as shown in Fig. 3). These two subbasins are chosen as they lie across a hydrologic divide (Aravali hills) one flowing west (subbasin 4) while the other flowing east, both contributing to two different major basins in the Indian subcontinent. These two basins are representative of the performance of all other 32 subbasins and all have not been included for brevity reasons.

Figure 4 displays estimated parameters F_{ce}, F_{ce0}, F_{ck} and η (= 0.1948, 0.0811, 0.0008, 0.1466), while Fig. 5a shows a map of $\Sigma_{t=1}^{12}\max(0, P_{i,t} - \underline{w}_i)$, annual total overland flow. Equally weighted mean absolute error value of 37.62 mm is obtained. Modeled storage changes underpredict the falling part of observed storage change in the beginning of the year (dry months). It also predicts the rising part of observed storage changes with a lag. However, the modeled values lie within the uncertainty bounds expressed by historic data for years 2002 to 2008 for most of the months. Modeled evaporation underpredicts peak values of MERRA evaporation data while
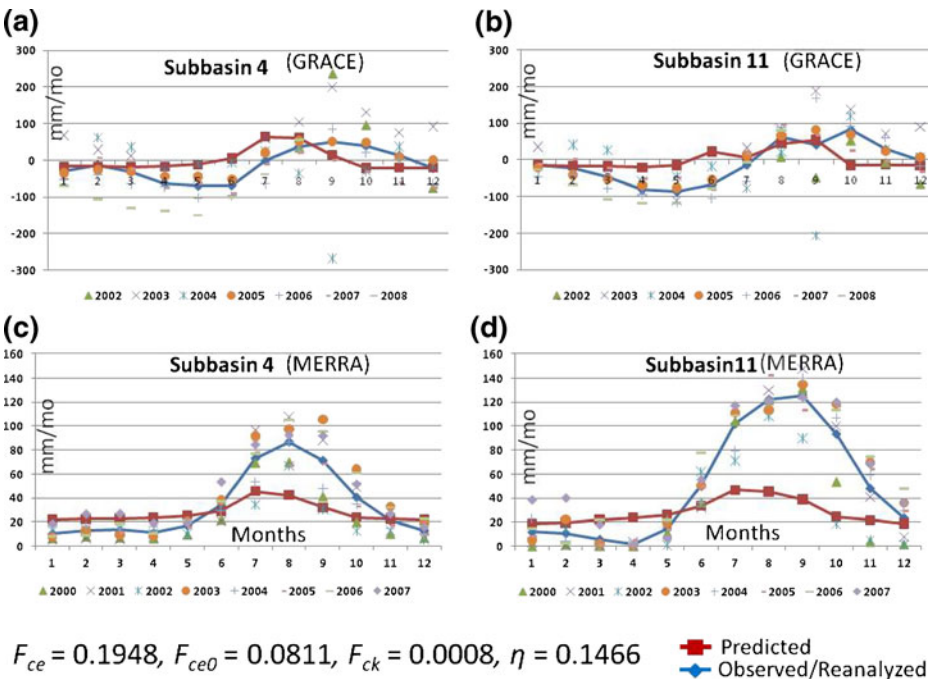

Fig. 4 Performance of modeled storage change and evaporation. The equally weighted mean absolute error value is 37.62 mm. (**a**) and (**b**) (for GRACE) show data values for years 2002–2008 by *symbols* indicated each year. Its mean monthly values are displayed by "Observed/Reanalyzed" values (*solid line* with superimposed marker). Data is similarly displayed for MERRA in (**c**) and (**d**) for years 2000–2007 along with mean monthly values indicated by "Observed/Reanalyzed" values (*solid line with a superimposed symbol*). Note that only annual cycle is modeled, which is displayed as "Predicted" (*solid line with a superimposed symbol*). For calibration of a model given via constraints (*1–10*), we use annual mean monthly time series of GRACE and MERRA2D
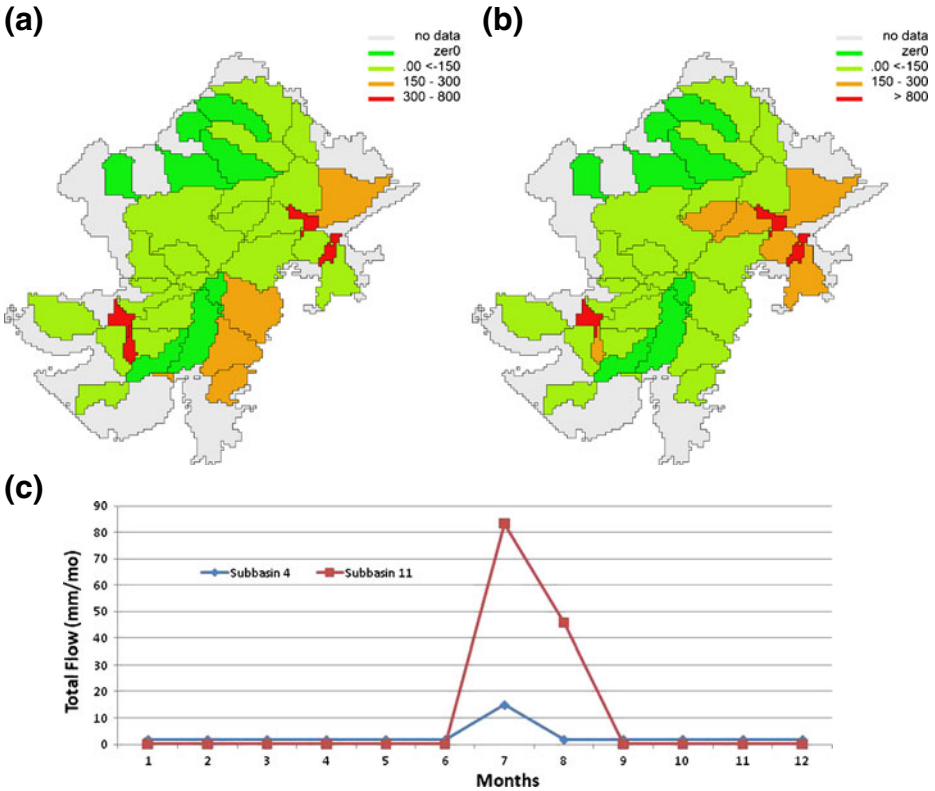
(a)

(b)

(c)

Fig. 5 a Annual total Hortonian flow (mm), **b** ratio of annual total Hortonian flow to the modeled baseflow, **c** modeled monthly total flow for subbasins 4 and 11

it slightly overpredicts evaporation during the dry season in the beginning months of a year.

Figure 5b maps the ratio $\Sigma_{t=1}^{12} \max(0, P_{i,t} - \underline{w}_i)$ (annual total overland flow) to $\Sigma_{t=1}^{12} K_i F_{ck} S_{i,t}$ (annual total base flow). It shows (when comparing it with Fig. 5a) that there are some areas where baseflow is negligible, leading to ratios in Fig. 5b larger than corresponding annual total overland flows in those areas. Figure 5c shows monthly time series for the two subbasins (4 and 11). Subbasin 11 has considerably larger overland flows when compared with subbasin 4, while baseflow remains negligible for both. The two selected subbasins are different in their soil properties with subbasin 4 being faster (higher saturated hydraulic conductivity) than subbasin 11.

Figure 6a maps the ratio $f_i = \underline{w}_i / \max_t \{P_{it}\}$, referred to as infiltration index, to display an approximate measure of infiltration capacity. Note that the threshold $\underline{w}_i$ can be both a function of monthly rainfall stochasticity as well as infiltration capacity and normalizing it by maximum monthly rainfall attempts to remove rainfall effect.

Figure 6b shows the variation of f_i with approximate hydraulic conductivity K_i. The plot shows that f_i, an approximation for infiltration capacity, increases with hydraulic conductivity. In order to test the significance of this increasing relationship,
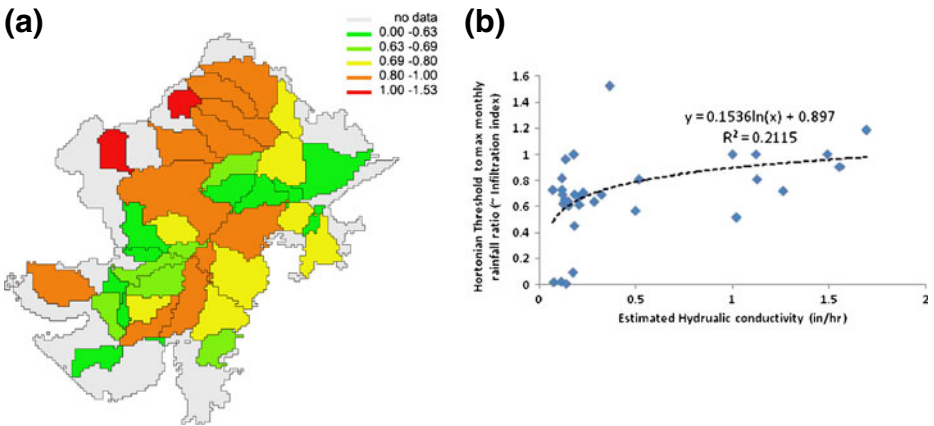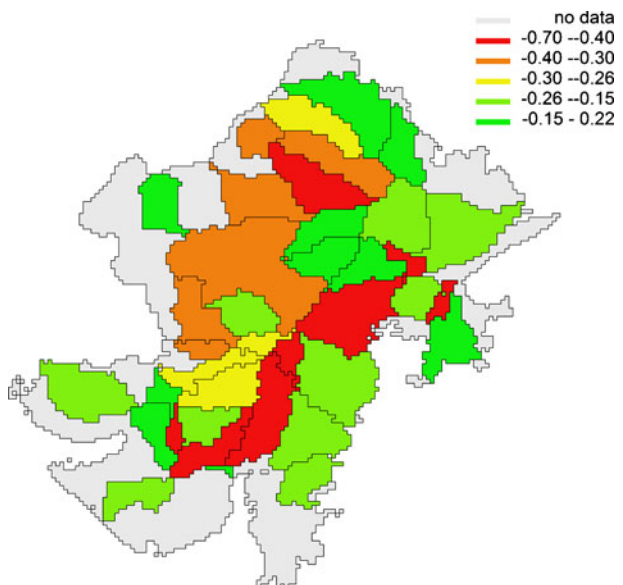
Fig. 6 a ratio, f_i (also called infiltration index), of modeled Hortonian threshold $\underline{w}_i$ to maximum monthly rainfall $max_t\{P_{it}\}$, **b** variation of f_i with hydraulic conductivity K_i

consider a linear regression with f_i as the dependent variable and K_i as the independent variable. Two tests are considered, T-test for the slope (null hypothesis: slope of the linear regression is zero) and F-test (null hypothesis: f_i increasing with estimated hydraulic conductivity, K_i, is by chance). T-statistic (slope of linear regression is 0.273 with standard error of 0.105) of estimated slope is 2.59 with critical t-statistic value for a significance level of 0.05 being 2.03 (degrees of freedom is 32). F-statistic has a value of 6.73 which is significant at 0.014 level (degrees of freedom is 32). Both these tests allow rejection of the null hypotheses at significance level of 0.05, thereby indicating a significant increasing relationship between the infiltration index and estimated hydraulic conductivity.

Fig. 7 Correlation between storage change residuals (predicted–modeled) and monthly rainfall

Finally, Fig. 7 maps the correlation between storage change residuals (modeled–predicted) and monthly rainfall. The correlation is mostly negative and significantly different from 0, indicating towards bias in model selection and incomplete model concept.

5 Discussion and Conclusion

The parameter estimates are robust even though only 12 monthly time steps of data for each subbasin were used. This is so because the model conceptualization is simple, though well constrained and parsimonious. It is further supported by observed (statistically significant) increasing relationship between infiltration index (outcome of parameter estimation) and conductivity (shown in Fig. 6b). This increasing relationship was not used in parameter estimation itself. This therefore acts as a validation step since a quantity conceptualizing infiltration capacity should theoretically increase with conductivity. Further, maximum modeled evaporation though appears to be biased (Fig. 4c,d), is similar to figures estimated by another study in the same area and based on a more detailed hydrologic model (see modeled evaporation for Luni river basin at http://gisserver.civil.iitd.ac.in/natcom/). This suggests that estimated parameters are physically meaningful, implying that the selected model is not physically unrealistic.

The processes conceptualized are known to be the dominant ones in dryland areas. This along with simple model conceptualization allows low "estimation" as well as "approximation" error. Error in prediction can theoretically be decomposed into error in estimating parameters of a model and error in how well that model is able to approximate the underlying processes. The former is called "estimation" error and the latter "approximation" error (Cucker and Smale 2002). Lower estimation error also implies lower data needs for precise parameter estimation (Cucker and Smale 2002; Norkin and Keyzer 2009). This statement can be further substantiated by a more rigorous analysis, though not presented in this study.

Evidence of approximation error is evident from Fig. 7. It shows the correlation between monthly rainfall and the residuals between modeled and predicted storage changes. This is indicative of model bias (Doherty and Welter 2010) and has parallel in literature on endogeneity bias in parameter estimation (Heckman 2009). Consider an arbitrary linear model wherein the error term ε is in part a function the independent variable x.

$$y = \beta x + \varepsilon$$

The parameter estimate $\hat{\beta}$ can be shown to be biased due to correlation between x and ε. This is elicited in Fig. 7 but for a hydrologic model where x is w_{it}, y is monthly storage change. Such a bias may be ascribed to nonlinearity of w_{it} in the underlying data generating process (governing differential equation) or a missing model component that is also a function of w_{it} (that appears in the error term when using a linear hydrological model).

Figure 4 also demonstrates inadequacy of the presented model. It suggests the model has an inability to mimic the negative amplitude of observed storage change as well as the timing of its peak. This can be due to insufficient representation of evaporation due to irrigation. Irrigation in the study is generally extracted from

groundwater sources. Irrigation can be conceptually extracted from the modeled (subbasin specific) storages. If irrigation is underrepresented, modeled storage during a period dominated by irrigation will be higher than when irrigation is not underrepresented in the model. Thus storage levels will be lower with higher evaporation due to better representation of irrigation.

Further model improvements may therefore be possible with finer resolution information on irrigation arrangements in the study area. However Fig. 7 suggests that the process conceptualization missing in the current model is negatively correlated with monthly rainfall. Since evaporation appears with a sign opposite to rainfall in mass balance, missing evaporation due to irrigation implies that irrigation is positively correlated with rainfall. This may serve as a contradiction since more irrigation is expected in dry or low rainfall conditions and less irrigation is during high rainfall months.

The negative correlation in Fig. 7 may suggest that the wet season conceptualization of total evaporation is incomplete. In arid/semi-arid areas, transition from water limited regime to energy limited regime can only occur during rainy season. However the system reverts back to water limited regime by the end of the rainy season. During the transition from water limited to energy limited regime, actual evaporation also increases towards its upper bound (evaporation demand). Thus the transition from water limited to energy limited regime in arid/semiarid areas is associated with two phenomena: increase in rainfall and increase in evaporation, which by definition implies positive correlation. Since in arid/semiarid regions, the period of energy limited state can be small in comparison with the period of transition from water limited to energy limited regime and vice versa, positive correlation between rainfall and evaporation can appear during the rainy season. Meanwhile evaporation during the dry season (in months January to June) may remain water limited in spite of irrigation leading to possibly weak negative correlation contribution to the overall rainfall-evaporation correlation.

Yet another missing concept can be outflow due to a missing store (that memorizes past rainfall events leading to positive correlation between rainfall and its store levels), thus leading to negative correlation between residuals and rainfall. Other possible estimation biases such as fraction of monthly rainfall directly evaporated and irrigation water use can be due to spatially uniform representation of these parameters.

There are also arguments for low reliability of reanalysis (evaporation) data in presence of irrigation, even though time to peak and the shape of monthly evaporation curve is useful information to estimate model parameters. Further parameter estimation is done in conjunction with observed storage change data. GRACE storage change data also has its limitations due to the spatial resolution of available data as well as due to applied smoothing (product used is 300 km half width smoothed) and that the GRACE product used in this study was optimized for ocean studies (Chambers 2006).

Finally Fig. 7 shows another limitation of the model. It shows that baseflow are nearly uniform and nonzero (though close to zero) throughout the year with overland flow occurring in high rainfall months of July and August. However flows are zero for driest part of the year in dryland areas (Wittenberg and Sivapalan 1999) suggesting a missing threshold conceptualization to model switch to zero flow condition under low store conditions (which in turn conceptualizes ground water level falling below the

stream bed in driest periods). This is not modeled due to non-identifiability of this threshold given that another threshold conceptualization (for Hortonian overland flow) already exists. Low variation in baseflow with rainfall events may be due to linear reservoir representation that keeps process conceptualization parsimonious.

The current study has not yet delved into various aspects of uncertainty in parameter estimation and model selection. A more detailed modeling exercise can be envisaged that incorporates sensitivity analysis, and multi-weight analysis. However this is relegated to a later study as the motivation for the current study is to present a constrained parsimonious modeling approach through mathematical programming.

A mathematical programming based approach to model data scarce dryland area was presented. Such an approach allows flexible introduction of constraints on water balance models. Spatial heterogeneity was introduced by representing a basin as a set of interconnected set of subbasins. One evident advantage of using mathematical programming is that it allowed implicit solution to state variables (store levels). The model results presented suggest that model conceptualization is incomplete though its estimation is robust. Assessment of correlation of residuals with monthly rainfall suggested that current model conceptualization needs improvement. In conclusion, mathematically programming dominant processes holds promise for parsimonious modeling in data scarce dryland areas.

References

Beven KJ (2002) Runoff generation in semi-arid areas. In: Bull LJ, Kirby MJ (eds) Dryland rivers: hydrology and geomorphology of semi-arid channels. Wiley, England, p 393

Bosilovich M (2008) NASA's modern era retrospective-analysis for research and applications: integrating Earth observations. Earthzine. Available at www.earthzine.org/2008/09/26/nasas-modern-era-retrospective-analysis/

Chambers DP (2006) Evaluation of new GRACE time-variable gravity data over the ocean. Geophys Res Lett 33(17):LI7603

Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE (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

Clark MP, Rupp DE, Woods RA, Tromp-van Meerveld HJ, Peters NE, Freer JE (2009) Consistency between hydrological models and field observations: linking processes at the hillslope scale to hydrological responses at the watershed scale. Hydrol Process 23:311–319

Cosby BJ, Hornberger GM, Clapp RB, Ginn TR (1984) A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resour Res 20(6): 682–690

Cucker F, Smale S (2002) On the mathematical foundations of learning. Bull Am Math Soc 39(1): 1–49

Cudennec C, Leduc C, Koutsoyiannis D (2007) Dryland hydrology in Mediterranean regions—a review. Hydrol Sci J 52(6):1077–1087

Doherty J, Welter D (2010) A short exploration of structural noise. Water Resour Res 46:W05525. doi:10.1029/2009WR008377

Dunkerley D (2000) Measuring interception loss and canopy storage in dryland vegetation: a brief review and evaluation of available research strategies. Hydrol Process 14:669–678

Eagleson PS (1982) Ecological optimality in water-limited natural soil-vegetation systems. 1. Theory hypothesis. Water Resour Res 18(2):325–340

FAO (2003) The digital soil map of the W, version 2.3. Land and water development division. FAO, Rome

Fenicia F, McDonnell JJ, Savenije HHG (2008) Learning from model improvement: on the contribution of complementary data to process understanding. Water Resour Res 44:W06419. doi:10.1029/2007WR006386

Franchini M, Wendling J, Obled C, Todini E (1996) Physical interpretation and sensitivity analysis of the TOPMODEL. J Hydrol 175:293–338

GAMS (2008) GAMS-A user's guide: tutorial by Richard E Rosenthal. GAMS Development Corporation, Washington DC, p 273

Gulden LE, Rosero E, Yang Z-L, Rodell M, Jackson CS, Niu G-Y, Yeh PJ-F, Famiglietti J (2007) Improving land-surface model hydrology: is an explicit aquifer model better than a deeper soil profile? Geophys Res Lett 34:L09402. doi:10.1029/2007GL029804

Hargreaves HG, Allen RG (2003) History and evaluation of Hargreaves evapotranspiration equation. J Irrig Drain Eng January/February 2003:53–63

Harou JJ, Pulido-Velazquez M, Rosenberg DE, Medellin-Azuara J, Lund JR, Howitt RE (2009) Hydro-economic models: concepts, design, applications, and future prospects. J Hydrol 375: 627–643

Heckman JJ (2009) The scientific model of causality. Sociol Method 35(1):1–97

Hoeksema RJ, Kitanidas PK (1984) An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling. Water Resour Res 20(7):1003–1020

ITC (2009) ILWIS 2.1 application guide. Available online at http://www2.itc.nl/ilwis/documentation/version_2/aguide.asp. Accessed on 08 December 2009, p 352

Li H, Sivapalan M (2010) Functional approach to exploring climatic and landscape controls on runoff generation. 1. Behavorial constraints on climate, soil and topography. Mimeo

Lo M-H, Famiglietti JS, Yeh PJF, Syed TH (2010) Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data. Water Resour Res 46:W0551710. doi:1029/2009WR007855

Maneta MP, Torres MO, Wallender WW, Vosti S, Howitt R, Rodrigues L, Bassoi LH, Panday S (2009) A spatially distributed hydroeconomic model to assess the effects of drought on land use, farm profits, and agricultural employment. Water Resour Res 45:W11412. doi:10.1029/2008WR007534

McIntyre N, Al-Qurashi A (2009) Performance of ten rainfall–runoff models applied to an arid catchment in Oman. Environ Model Softw 24:726–738

Michaud J, Sorooshian S (1994) Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. Water Resour Res 3(3):593–605

Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int J Climatol 25:693–712. doi:10.1002/joc.1181

Niu G-Y, Yang Z-L, Dickinson RE, Gulden LE, Su H (2007) Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. J Geophys Res 112:D07103. doi:10.1029/2006JD007522

Norkin VI, Keyzer MA (2009) Asymptotic efficiency of kernel support vector machines (SVM). Cybern Syst Anal 45(4):575–588

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

Pilgrim DH, Chapman TG, Doran DG (1988) Problems of rainfall–runoff modelling in arid and semiarid regions. Hydrol Sci J 33(4):379–400

Pulido-Velazquez D, Sahuquillo A, Andreu J, Pulido-Velazquez M (2007) An efficient conceptual model to simulate surface water body–aquifer interaction in conjunctive use management models. Water Resour Res 43:W07407. doi:10.1029/2006WR005064

Rawls WJ, Brakensiek DL, Saxton KE (1982) Estimation of soil water properties. Trans ASAE 25(5):1316–1320, 1328

Rienecker MM and co-authors (2008) The GEOS-5 data assimilation system—documentation of versions 5.0.1 and 5.1.0, vol 27. NASA GSFC Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2007-104606, p 92

Rodriguez-Iturbe I, Porporato A (2004) Ecohydrology of water-controlled ecosystems: soil moisture and plant dynamics. Cambridge University Press, NY, p 442

Roy PS, Agrawal S, Joshi P, Shukla Y (2003) The land cover map for southern Asia for the year 2000. GLC2000 database, European Commision Joint Research Centre, 2003. http://www-gem. jrc.it/glc2000

Savenije HHG (2009) The art of hydrology. Hydrol Earth Syst Sci 13:157–161

Savenije HHG (2010) HESS opinions "Topography driven conceptual modelling (FLEX-Topo)". Hydrol Earth Syst Sci 14:2681–2692. doi:10.5194/hess-14-2681-2010

Seong K (2002) A study on the effective hydraulic conductivity of an anisotropic porous medium. KSME Int J 16(7):959–965

Sharma KD, Murthy JSR (1998) A practical approach to rainfall–runoff modeling in arid zone drainage basins. Hydrol Sci J 43(3):331–348

Sheffield J, Ferguson CR, Troy TJ, Wood EF, McCabe MF (2009) Closing the terrestrial water budget from satellite remote sensing. Geophys Res Lett 36:L07403. doi:10.1029/2009GL037338

Sun AY, Green R, Rodell M, Swenson S (2010) Inferring aquifer storage parameters using satellite and in situ measurements: estimation under uncertainty. Geophys Res Lett 37:L10401. doi:10.1029/2010GL043231

Syed TH, Famiglietti JS, Chen J, Rodell M, Seneviratne SI, Viterbo P, Wilson CR (2005) Total basin discharge for the Amazon and Mississippi River basins from GRACE and a land–atmosphere water balance. Geophys Res Lett 32:L24404. doi:10.1029/2005GL024851

Tiwari VM, Wahr J, Swenson S (2009) Dwindling groundwater resources in northern India, from satellite gravity observations. Geophys Res Lett 36:L18401. doi:10.1029/2009GL039401

van den Boom B, Pande S (2007) A user manual for the SAS facility to plot maps. Working Paper 07-03. Amsterdam, SOW-VU, p 38

Vogel RM (2006) Regional calibration of watershed models. In: Singh V, Frevert D (eds) Watershed models. Taylor & Francis, pp 47–69

Werth S, Güntner A, Petrovic S, Schmidt R (2009) Integration of GRACE mass variations into a global hydrological model. Earth Planet Sci Lett 277:166–173

Wittenberg H, Sivapalan M (1999) Watershed groundwater balance estimation using streamflow recession analysis and baseflow separation. J Hydrol 219:20–33

Yeh PJ-F, Swenson SC, Famiglietti JS, Rodell M (2006) Remote sensing of groundwater storage changes in Illinois using the Gravity Recovery and Climate Experiment (GRACE). Water Resour Res 42:W12203. doi:10.1029/2006WR005374

Yu B, Tisdell J, Podger G, Salbe L (2003) A hydrologic and economic model for water trading and reallocation using linear programming techniques. Paper presented at MODSIM 2003 international congress on modelling and simulation, Townsville, Queensland, Australia