Exploring the hydrological robustness of model-parameter values with alpha shapes
José-Luis Guerrero,¹ ² Ida K. Westerberg,¹ ³ ⁴ Sven Halldin,¹ Lars-Christer Lundin,¹ ⁵ and Chong-Yu Xu⁶

Received 5 June 2013; revised 11 September 2013; accepted 12 September 2013.

Estimation of parameter values in hydrological models has gradually moved from subjective, trial-and-error methods into objective estimation methods. Translation of nature’s complexity to bit operations is an uncertain process as a result of data errors, epistemic gaps, computational deficiencies, and other limitations, and relies on calibration to fit model output to observed data. The robustness of the calibrated parameter values to these types of uncertainties is therefore an important concern. In this study, we investigated how the hydrological robustness of the model-parameter values varied within the geometric structure of the behavioral (well-performing) parameter space with a depth function based on α shapes and an in-depth posterior performance analysis of the simulations in relation to the observed discharge uncertainty. The α shape depth is a nonconvex measure that may provide an accurate and tight delimitation of the geometric structure of the behavioral space for both unimodal and multimodal parameter-value distributions. WASMOD, a parsimonious rainfall-runoff model, was applied to six Honduran and one UK catchment, with differing data quality and hydrological characteristics. Model evaluation was done with two performance measures, the Nash-Sutcliffe efficiency and one based on flow-duration curves. Deep parameter vectors were in general found to be more hydrologically robust than shallow ones in the analyses we performed; model-performance values increased with depth, deviations to the observed data for the high-flow aspects of the hydrograph generally decreased with increasing depth, deep parameter vectors generally transferred in time with maintained high performance values, and the model had a low sensitivity to small changes in the parameter values. The tight delimitation of the behavioral space provided by the α shapes depth function showed a potential to improve the efficiency of calibration techniques that require further exploration. For computational reasons only a three-parameter model could be used, which limited the applicability of this depth measure and the conclusions drawn in this paper, especially concerning hydrological robustness at low flows.


1. Introduction

Hydrological models are important tools for water management but their representations of real-world hydrological systems are affected by significant uncertainties. Problems such as inadequate conceptualizations of physical processes, uncertainties in available data and commensurability errors between spatial and temporal scales of model and data hinder a derivation of model-parameter values directly from basin traits [Beven, 2002]. Most model applications require calibration, an adjustment of the model-parameter values so that observed values of at least one flux or state variable are adequately simulated. Historically, calibration techniques have gravitated toward finding the parameter-value combination (parameter vector) that gives the best result, according to one, or multiple, goodness-of-fit measures. Today, the uncertainties in the model calibration as a result of the time period used for calibration [Gupta and Sorooshian, 1985; Yapo et al., 1996; Wagener et al., 2003; Singh and Bårdossy, 2012], the selected objective function [Efstratiadis and Koutsoyiannis, 2010], and the nature and magnitude of data errors [Bårdossy and Singh, 2008; Renard et al., 2010; McMillan et al., 2011; Westerberg et al., 2011b], are increasingly recognized and taken into account. The robustness of the calibrated parameter values to these types of uncertainties is therefore an important concern.
[1] The term robust has multiple interpretations. Robustness can be linked to the model structure that should be as parsimonious as possible [Klemes, 1983; Dooge, 1986; Son and Sivapalan, 2007], to data that should have small errors and for which the error structure is either known or can be approximated [Beven and Westerberg, 2011], or to the calibration procedure [Bardossy and Singh, 2008; Peel and Blöschl, 2011]. For the latter, Bardossy and Singh [2008] (BS08 from now on) list the following criteria for robust parameter vectors: 1) they lead to good model performance, 2) they lead to hydrologically reasonable representations of the corresponding processes, 3) they are not sensitive to small changes in parameter values, and 4) they are transferable in time, leading to good model performance for other time periods and perhaps also other catchments. Model performance can in summary be expressed as the value of the performance measure used for model calibration and evaluation, and therefore depends on the choice of this measure [Freer et al., 2003; Krause et al., 2005]. Model performance can also be analyzed in depth to determine the ability of the simulations to represent the relevant hydrological processes in a posterior performance analysis against observational data, given their estimated uncertainties, e.g., McMillan et al. [2010] and Westerberg et al. [2011b]. In this paper we used a depth measure to analyze how hydrological robustness of calibrated model-parameter values varied within the geometric structure of the behavioral parameter space.

[2] Data depth is a concept in computational geometry, a discipline engendered by Shamos [1978], with links to statistical analysis since its inception. If parameter vectors in the behavioral space are considered samples from a multivariate distribution, then the deepest point (or region of the space) is an estimator of the distribution’s median [Aloupis, 2006].

[3] The first application of data depth to hydrology is found in the work of Chebana and Ouarda [2008], who use the Mahalanobis [1936] depth function to find the weights in a weighted linear least-squares regression for regional flood estimation in Canada. Shortly after came BS08, who build on the results of Bardossy [2007], and find that parameter vectors deep within the well-performing parameter space are robust in terms of sensitivity to small changes in parameter values and transferability to other time periods. BS08 also propose a calibration method (robust parameter estimation: ROPE) that iteratively refines the convex region encompassing the well-performing parameter values. BS08 use the half-space depth [Tukey, 1975], which is a convex depth measure and one of many depth functions in computational geometry [Liu et al., 1999; Zuo and Serfling, 2000; Aloupis, 2006; Lee, 2006].

[4] The half-space and the Mahalanobis depths are connected by the connection of a convex depth measure and one of many depth functions to identify outliers and to describe bivariate quantiles [Keefer et al., 2010]. The two main advantages of such a depth measure are that it does not require the assumption, implicit in most depth functions, that the underlying multivariate distribution is unimodal [Liu et al., 1999; Krasnoshekevich and Polishchuk, 2013], and that it might give a tighter delimitation of the behavioral parameter space, i.e., less space to be explored. An added advantage is that a depth function based on alpha shapes has potential to detect different modes in a multivariate distribution [Krasnoshekevich and Polishchuk, 2013].

![Figure 1. Illustration of Delaunay triangulation, convex hull and \(\alpha\) shape. Random points (black dots) generated between two circles of different radius (Annulus), and corresponding Delaunay triangulation (blue lines), convex hull (red line) and alpha shape (green lines).](image-url)
In this paper we used \( \alpha \) shapes to investigate how the hydrological robustness of calibrated parameter values varied within the geometric structure of the behavioral parameter space. We used an in-depth posterior performance analysis that accounted for observational uncertainties in the data to analyze how the hydrological robustness varied with depth for model simulations for six Honduran and one UK catchment with different data quality and hydrological characteristics. We also wanted to find out if the \( \alpha \) shape depth was a useful estimator of parameter depth and under which conditions deep parameters would be hydrologically robust. Our concrete objectives were to:

1. Investigate how the hydrological robustness of the behavioral parameter vectors changed within the behavioral parameter space.
2. Use \( \alpha \) shapes to detect multimodality in the behavioral parameter space.
3. Analyze the potential to reduce the computational cost of ROPE.

2. Background

2.1. Depth Functions

The geometric concept of depth deals with the ranking of data: ordering points in space from shallow to deep. Such classification requires a frame of reference, i.e., a direction along which depth increases, and defining (at least) one deepest point. In one dimension, calculating depth is straightforward, since there is a single axis along which data can be ordered. Computing depth in a multidimensional setting is much more complicated. In two or more dimensions the answer to the problem of ranking data is not unique [Barnett, 1976; Small, 1990; Gil et al., 1992; Aloupis, 2006]: there is no single axis along which values can be ordered. The selection of a direction of changing depth can be subjective and the a priori definition of a deepest point might be required. Depth functions are mathematical constructs used to calculate the depth of a point with respect to a multidimensional set, generally through a center-outward ordering. Different depth functions might give different results and yield differing centers and/or depth profiles [Hugg et al., 2006]. The two depth functions that have so far been used for parameter-value estimation in hydrology are the Mahalanobis depth and the half-space depth [Chebana and Ouarda, 2011a, BS08]. These two functions are based on some assumptions and have some properties that could potentially limit their scope:

1. Their use is based on the assumption that the deepest point, i.e., the median, is unique, which might be inadequate if the parameter-value distribution is multimodal. This might be an issue when, e.g., there exist multiple optima in the parameter-value distribution. Most depth functions implicitly assume unimodality in the underlying distribution [Zuo and Serfling, 2000].
2. The Mahalanobis and the half-space depths are based on convex sets (see also next section) for which the depth contours corresponding to a certain depth are convex [Dyckerhoff, 2004] and might not adequately describe the shape of a point set (Figures 1 and 2).
3. The Mahalanobis depth relies on strong assumptions about the normality of the underlying multivariate distribution.

2.2. Convexity and the Half-Space Depth

A convex set in Euclidean space has the property that a line between any two points of the set lays within the set. The convex hull of a set of points \( S \) is the smallest possible convex set containing \( S \) (Figure 1).

Convex sets are attractive in computational geometry because of the relative simplicity of operations on them, valid in any number of dimensions [Gruber and Willis, 1993]. Convex sets play a central role in the half-space depth. The half-space depth of a point \( x \) with respect to \( S \) will be the minimum number of points of the set on one side of any hyperplane passing through \( x \). The contours delimiting regions of equal depth are convex and the outermost is the convex hull of the set [Struyf and Rousseeuw, 1999].

A conceptually simpler notion of depth, on which the \( \alpha \) shape depth is based, is a successive peeling of points in the boundaries of the distribution, until the center or some limit is reached. In this study we used \( \alpha \) shapes [Edelsbrunner and Mücke, 1994] to define the successive boundaries and compute depth.

2.3. Alpha Shapes

From an exploratory point of view, the convex hull can be a poor descriptor of the shape of a set of points (Figures 1 and 2). Alpha shapes, on the other hand, are well...
Table 1. Characteristics of Catchments and Discharge Time Series

<table>
<thead>
<tr>
<th>Station</th>
<th>Period</th>
<th>Number of Measurements</th>
<th>Area (km²)</th>
<th>Elevation (m)</th>
<th>Mean Annual Runoff (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paso La Ceiba</td>
<td>5 May 1980 to 25 Aug 1997</td>
<td>Variable*</td>
<td>1760</td>
<td>675-2138</td>
<td>350</td>
</tr>
<tr>
<td>Río Del Hombre</td>
<td>5 May 1990 to 1 Oct 1998</td>
<td>470</td>
<td>273</td>
<td>931-2229</td>
<td>400</td>
</tr>
<tr>
<td>Brue</td>
<td>1 Jan 1995 to 30 Jun 1998</td>
<td>15 min¹</td>
<td>135</td>
<td>20-260</td>
<td>439</td>
</tr>
</tbody>
</table>

*Three measurements per day were available most of the time. The given period includes some intervals without data.
¹15 min discharge data were integrated to daily values of discharge.

3. Materials and Methods

The methods presented here were developed for six catchments within the Choluteca River basin, Honduras, and were tested on these as well as on one catchment in Great Britain. The six catchments in Honduras are located in the upper reaches of the Choluteca River basin and are characterized by shallow soils, with frequently occurring surface runoff. The Brue catchment in Great Britain has a different flow regime where runoff is dominated by subsurface processes.

3.1. The Choluteca River Basin

The Choluteca River basin in the south of Honduras has its outlet at the Gulf of Fonseca, in the Pacific Ocean and a small part of its 7500 km² encroaches on Nicaraguan territory. Six mountainous catchments in the upper part of the basin were used in this study with five of them situated within the larger Paso La Ceiba catchment (Table 1 and Figure 3).

Two large dams within the Paso La Ceiba catchment supply water to Tegucigalpa, the Honduran capital, but only the flow at Paso La Ceiba is affected by them. The characteristics of the Choluteca River basin and the Paso La Ceiba catchment are described by Westerberg et al. [2010, 2011a] and Guerrero et al. [2012]. The precipitation regime over the area is complex and highly influenced by the meridional migration of the Inter-Tropical Convergence Zone (ITCZ) together with the configuration of the region’s high mountain range, which contributes to a splitting of the local climate into a dry and a rainy season during which around 80% of the precipitation falls. The precipitation during the rainy season, which starts around April–May and ends around November–December, is modulated by a relative low known as “El Veranillo,” or the midsummer drought, in July–August. The “Veranillo” is associated with a strengthening of the trade winds, which in turn is related to variations in the position of the ITCZ over the eastern Pacific [Magaña et al., 1999; Hastenrath, 2002]. Furthermore, the regime is affected by interannual climate variations related to El Niño Southern Oscillation [Giannini et al., 2001].

Shallow soils with limited water retention capacity, steep topography and the intensity of the rainfall make for a fast rainfall-runoff response in this area. Even at the outlet of the Paso La Ceiba, the largest catchment, the delay between precipitation in the headwater areas and peak flow at the outlet is less than 24 h [Westerberg et al., 2011b].

The discharge data were available at varying time periods (Table 1). The precipitation and potential evaporation were interpolated for each station so that there was 1 year of data for model warm-up.
3.1.1. Rainfall

Precipitation data from a network of 29 rain gages in or nearby the Paso La Ceiba catchment were spatially interpolated using the inverse-distance weighting method. The lack of correlation in daily precipitation at different stations prevented the use of more complex methods [Westerberg et al., 2010]. Precipitation was interpolated to a 9° grid, coincident to the Shuttle Radar Topography Mission (SRTM) [Farr et al., 2007] grid, and then averaged to each catchment.

Precipitation readings are normally taken at 07:00. Since most of the rain falls during the latter half of the day and the delay time between precipitation and runoff is less than 24 h, the registration of the precipitation was moved to the day of the actual measurement to agree with the daily time step in the model [Westerberg et al., 2011b].

3.1.2. Discharge

Time series of discharge and upper and lower bounds for the estimated discharge uncertainty, calculated with a time-variable fuzzy rating-curve method according to Westerberg et al. [2011a], were used for the Paso La Ceiba catchment. For the remaining stations, a large set of momentary discharge measurements taken to determine the stage-discharge rating were used to calibrate the model (Table 1). The uncertainty in these discharge measurements was assumed to be of ±50%, similar to the range quantified for daily average discharge at Paso La Ceiba [Westerberg et al., 2011a].

3.1.3. Evaporation

Data from the Toncontín station (Tegucigalpa’s international airport), were used to calculate a baseline daily time series of potential evaporation using the Penman-Monteith [Penman, 1948; Allen et al., 1998] formula. The Toncontíin evaporation series were spatially extrapolated over the study area by means of the long-term averages of temperature from nine other stations in the basin, which were linearly regressed against elevation. Temperature changed with elevation at a rate of −0.0055°C/m with a correlation coefficient of 0.79. The regression was used to extrapolate daily temperature from Toncontín to the SRTM [Farr et al., 2007] elevation grid. This temperature field provided a basis for the spatial extrapolation together with the Thornthwaite [1948] equation:

\[ ET_P = C \left( \frac{10^T}{I} \right)^a \]

This equation relates monthly potential evaporation \( ET_P \) (mm) to a constant \( C \) generally equal to 16, to monthly average temperature \( T \) (°C), to a temperature index \( I \), and to an exponent \( a \) that is a function of \( I \). The Thornthwaite equation is based on a month with 30 days and a day of 12 h, and was adjusted using the actual month length and the average daylight derived with the formula recommended by Forsythe et al. [1995].

A separate constant \( C \) was used for every month, ranging from 19 to 26. The Thornthwaite equation was calibrated to equate the Penman values at Toncontín. The monthly \( C \) values were assumed constant over the area and used to calculate the Thornthwaite potential evaporation for all extrapolated temperatures. The baseline evaporation was multiplied by the ratio between the Thornthwaite and Penman yearly averages at each grid point to obtain daily estimates that were then averaged to each catchment.

3.2. Brue Catchment

The Brue catchment is located in southwest England and has an area of around 135 km² (Figure 3). The land use is dominated by grasslands, underneath which clayey soils are superimposed on alternating bands of permeable and impermeable rocks. The topography is characterized by low hills of up to 300 meters above mean sea level. An extensive precipitation data set was available from the HYREX (HYdrological Radar EXperiment) [Wood et al., 2000; Moore et al., 2000] project, with 49 rain gages that were previously used to calculate the areal precipitation over the catchment with a nearest-neighbor technique by Younger et al. [2009]. This areal precipitation was used as a uniform input to the hydrological model. The average areal precipitation over the catchment was 770 mm per year for the period 1 January 1995 to 31 December 1997. Based on the results of Calder et al. [1983], potential evaporation was calculated using a seasonal sine curve. Discharge and the related uncertainty for the Lovington gaging station are presented by Westerberg et al. [2011b] using the same
method as for the Paso La Ceiba catchment, but in this case with a constant rating curve.

3.3. Hydrological Model and Simulations

[37] WASMOD [Xu, 2002] is a parsimonious, lumped, conceptual, rainfall-runoff model based on the variable-source-area concept (Table 2). Originally a monthly water-balance model [Xu et al., 1996], it has evolved and has been applied at a global scale [Widen-Nilsson et al., 2009], at a daily time step [Westerberg et al., 2011b], and in regions of widely varying climate conditions [Xu et al., 2006; Li et al., 2011; Kizza et al., 2011]. We used the same version of the model as Westerberg et al. [2011b] who present good results for the Paso La Ceiba catchment with simulations that have high reliability and good precision for five different aspects of the hydrograph (peak flows, base flows, rising limbs, falling limbs and troughs).

[38] A first round of 10⁶ Monte-Carlo (MC) simulations was performed for all stations using all available discharge data for calibration, except when performing split-sample testing. The empirical cumulative distribution functions of behavioral parameter values were compared and the parameter $A_\text{et}$ was considered fixed to 0.7 in a second round of 10⁶ simulations that were the basis for the geometrical analyses since the software only allowed analysis in three dimensions or less. In order to assess the impact of fixing the $A_\text{et}$ parameter on streamflow simulations, 10⁶ random MC simulations with a free, randomly varying, $A_\text{et}$ parameter were also performed, but not used for depth calculations.

3.4. Performance Measures

[39] The Nash-Sutcliffe efficiency (NSE) [Nash and Sutcliffe, 1970] is one of the most widely used performance measures in hydrology, despite being much debated [see Westerberg et al., 2011b, and references therein], and was included as a baseline in our model evaluations using subjectively a behavioral threshold of 0.6.

[40] Model performance was also evaluated using flow-duration curves (FDC) within the extended Generalised Likelihood Uncertainty Estimation (GLUE) limits-of-acceptability framework, according to the procedure described by Westerberg et al. [2011b]. This allows model-rejection criteria to be based on the estimated uncertainty in the discharge data, enables simultaneous calibration to the whole range of flows, and is less sensitive to some types of epistemic data errors. However, in some circumstances the calibration provides little constraint for timing errors and a posterior performance analysis should therefore be performed to analyze if the FDC provides sufficient constraints for the particular model and data used. In calibration, the simulated FDC is compared to the uncertain observed FDC, which is based on the estimated uncertainty in the discharge data, at selected evaluation points on the curve. Behavioral simulations are required to be inside the uncertainty in the observed FDC at all these points and a performance measure is calculated by weighting the deviations from the best-estimated discharge within the uncertainty bounds. Similarly to Westerberg et al. [2011b], the evaluation points were calculated by dividing the area under the curve (representing the total volume of water contributed by flows smaller than or equal to a given magnitude) into 20 equal intervals, disregarding the extremes (i.e., 19 evaluation points). If not specified otherwise, the performance measures were calculated with all available discharge data (Table 1). The FDC performance measure ranges from zero to one, where one represents a perfect fit to the best-estimated discharge at all evaluation points and zero means that the simulation falls outside the uncertainty bounds for at least one evaluation point, i.e., it is nonbehavioral.

3.5. Alpha Shapes Calculations

3.5.1. Multimodality Detection

[41] A depth function based on $\alpha$ shapes can give an idea about the median of each mode in a multimodal distribution [Krasnoshchekov and Polishchuk, 2013]. Multimodality can be reflected as disconnected components in the $\alpha$ shape. Since the $\alpha$ shape depends on the Euclidean distance between points, we first scaled the sampling ranges for all parameters (Table 2), to [0;1]. Multimodality was analyzed by examining the $\alpha$ shape with $\alpha$ radius equal to the median of the distances between the connected points of the Delaunay [1934] triangulation of all ($10^6$) MC parameter vectors. The subjective selection of the $\alpha$ radius to detect multimodality was based on two competing factors: (i) the radius should be large enough to avoid excessive fragmentation, and (ii) the radius
3.5.2. Data-Depth Calculations

[42] From a finite set $S$ of points in $\mathbb{R}^d$, where $d$ is the number of dimensions, a finite number of $\alpha$ shapes can be derived and all are subsets of the underlying Delaunay triangulation (Figure 1). The number of $\alpha$ radii for which the $\alpha$ shape experiences topological change is finite, and can be determined from the triangulation. Whereas a convex hull divides space into an interior and an exterior region, an $\alpha$ shape does not necessarily do so (Figure 2b). The boundary $\alpha$ shape (BAS) of $S$ was defined to be, in $\mathbb{R}^3$, the $\alpha$ shape with a single connected component with smallest possible $\alpha$ radius such that the $\alpha$ shape had neither dangling edges (1-simplex) nor faces (2-simplex), thus forming a closed polyhedron (Figure 2c). The BAS was found through a binary search on the ordered set of topologically relevant $\alpha$ radii until a polytope fulfilling the given conditions was identified: a closed polyhedron.

[43] The depth of a single parameter vector with respect to the behavioral set was determined as follows:

1. Let $S$ be the set of behavioral parameter vectors and BAS, the BAS for depth $n$.
2. Find the BAS$_1$ of $S$. Parameter vectors belonging to BAS$_1$ have depths equal to one and all parameter-vectors on or inside BAS$_1$ have a depth $\geq 1$.
3. Eliminate the parameters belonging to the BAS$_1$ from $S$.
4. Increase depth by one.
5. Repeat steps 2-4, until no points are left or BAS$_n$ cannot be found.
6. All parameter vectors outside BAS$_{n-1}$ have zero depth. Iterations were stopped with 200 points remaining. This can be thought of as peeling the layers of an onion, where each successive layer is an $\alpha$ shape and each peel a unit increase in depth.

3.6. Hydrological Robustness and Depth

[50] The hydrological robustness of parameter vectors at different depths was analyzed from four different aspects based on the criteria given by BS08: model performance, sensitivity to parameter-value changes, hydrologically reasonable process representation, and transferability in time. These analyses were performed within an uncertainty estimation framework, accounting for estimated uncertainties in discharge. Where the number of behavioral Monte Carlo simulations was larger than 100,000, the depth computations were limited to the first 100,000 randomly sampled behavioral simulations because of computational limitations related to memory constraints.

[51] First we explored how the performance-measure values varied with depth in the behavioral parameter space, both concerning average performance and variability in performance. Does performance increase when moving from shallow to deep parameter vectors? The performance for the points on the vertices of BAS$_n$ was the baseline. Then we examined the sensitivity of the model performance to a change in the parameter vector. Two aspects were tested:

1. Is the performance sensitive to lateral displacements (as opposed to displacements toward the center)? In $\mathbb{R}^3$ the boundary simplices of the BAS are triangles and this test consisted in comparing the average performance of the vertices of BAS$_1$ to the average performance of the circumcenters of the boundary triangles.

[52] 2. How sensitive is model performance to displacements of parameter vectors near the boundary between the behavioral and the nonbehavioral parameter space? The positive and negative unit normal vectors of the oriented triangles forming BAS$_1$ were found and defined a direction inward and outward the behavioral space. The model performance for parameter vectors at different distances along both normal vectors was evaluated.

[53] The model’s ability to represent the hydrological processes in the catchments was then evaluated by comparing observed and simulated hydrographs for each catchment in a posterior performance analysis. For the Brue and Paso La Ceiba catchments where continuous data were available, six different aspects of the hydrograph were analyzed: base flows, troughs, small peaks, large peaks, rising limbs, and falling limbs. These aspects were identified similarly to Younger et al. [2009] and Westerberg et al. [2011b]. Base flows were defined as flows smaller than 1.7 m$^3$/s at Brue and 5 m$^3$/s at Paso La Ceiba, large peaks defined as peak flows larger than 10 m$^3$/s at Brue and 100 m$^3$/s at Paso La Ceiba, and the time-window parameter for the identification of the other aspects set to 1 day at Brue and 3 days at Paso La Ceiba. The posterior performance analysis consisted of:

1. Calculation of summary-evaluation measures: Reliability (percentage of time the observed and simulated uncertain bounds overlapped) and Precision (the width of the overlapping range as a percentage of the width of the simulated bound and then calculated as an average for the whole time series). The range for both measures is between zero and 100%, and the measures were calculated both for the entire time series and for the different hydrograph aspects. The uncertainty bounds for the simulated discharge were taken to be the 5% and 95% percentiles of the likelihood-weighted distributions of the simulated discharge for every time step for all behavioral parameter vectors. The Reliability and Precision measures were calculated both with a freely varying $A_{et}$ parameter and with a constant $A_{et}$. This provided a basis to assess the impact of a constant $A_{et}$ on the discharge simulations.

[54] 2. Analyses of scaled scores for different depths and different hydrograph aspects [Westerberg et al., 2011b]. The scaled scores describe the deviation between a simulated time series from a single parameter vector and the observed discharge with its uncertainty bounds. A simulated value that is exactly at the best-estimate discharge has a scaled score of zero, one that is at the lower observed uncertainty bound a scaled score of minus one, and one that is at the upper observed uncertainty bound has a scaled score of one. Values inside/outside the bounds are linearly interpolated/extrapolated.

[55] Finally, we analyzed how the temporal transferability of the behavioral parameter vectors varied with depth by performing split-sample tests at Paso La Ceiba and Brue, where sufficient discharge data were at hand (Table 1). At Paso La Ceiba the model parameter values were calibrated in 1980–1988 and evaluated in 1989–1997 and vice versa. At Brue the model parameter values were calibrated in January 1995 to June 1997 and evaluated in July 1997 to June 1998 as well as calibrated in January 1996 to June 1998 and evaluated in January 1995 to December 1995.
3.7. Potential Efficiency of Alpha Shapes

Alpha shapes can be better descriptors of the shape of a set of points than their convex hull. The hyper-volume of an \( \alpha \) shape is, if it exists, at most equal to the hyper-volume of the convex hull. The ROPE (BS08) calibration technique requires iterative MC simulations of points inside a convex hull. The smaller volume of \( \alpha \) shapes therefore has the potential to speed up the ROPE process:

1. Define a hyper-rectangle bounding the parameter space.
2. Perform MC simulations with parameter vectors sampled inside the rectangle.
3. Select well-performing parameter vectors.
4. Find the convex hull of the well-performing parameter vectors.
5. Perform MC simulations with parameter vectors sampled from the inside of the convex hull.
6. Repeat steps 3–5 until the improvement in performance does not change more than expected from observation errors.

Step 4 of the ROPE procedure could be replaced by “select the BAS of the well-performing parameter vectors,” seeking to diminish the search space for the MC simulations by using alpha shapes. Consequently, step 5 would be modified to “Perform MC simulations with parameters inside the BAS.” In order to assess the improvement potential, we compared BAS volumes at different depths to the corresponding convex-hull volumes.

4. Results

4.1. Simulations, Multimodality, and Depth Computations

The number of behavioral parameter vectors differed from station to station and no behavioral parameter vectors were found at all for NSE at one of the stations (Figure 4). It is not possible to directly compare the number of behavioral simulations between NSE and FDC, since the behavioral threshold was subjectively selected to 0.6 for NSE.

The behavioral parameter spaces were spatially compact and no signs of multimodality were found for the selected criteria. The space occupied by the behavioral parameter vectors was partly different for the two performance measures (e.g., Figure 5).

The number of depths that could be computed was related to the number of behavioral simulations. All random parameter vectors inside BAS1 were found to be behavioral, i.e., there were no regions with nonbehavioral results inside BAS1, for the given sampling density, thus showing that the behavioral space was well structured in three dimensions.

4.2. Hydrological Robustness and Depth

4.2.1. Model Performance

Average performance generally increased with depth (Figures 6 and 7). The increase in performance was accompanied by a decrease in the standard deviation of the values of the performance measures. Deep parameter vectors did not always include the best-performing ones from the MC simulations. This confirmed the BS08 results that average model performance increases while performance variability decreases going from shallow to deep parameter vectors.
4.2.2. Sensitivity to Parameter-Vector Changes

The sensitivity to parameter-vector changes was generally found to decrease with depth (Figure 7). Changes in performance tended to become smaller with depth and were mostly positive, reflecting a performance increase with depth (Figure 7). Even in the worst cases, the change in performance with depth was not significantly different from zero (Figure 7). Deep parameter vectors tended on average to perform better than shallow ones and the model was also less sensitive to changes in their values.

The change in performance for lateral parameter-value changes showed that some of the circumcenters were not behavioral and the performance at the circumcenters was found to be, on average, different than the performance at the vertices, meaning that $BAS_1$ was only an approximation of the behavioral limit. The performance change was small but not negligible for lateral parameter-vector changes but the small magnitude of these changes in comparison to the change in performance with depth showed that $BAS_1$ was a good approximation of the behavioral boundary (Table 3).

The performance dropped when moving away from $BAS_1$ whereas it increased moving inward (Figure 8). The drop was another indicator of the ability of $BAS_1$ to capture the behavioral limit. In conclusion, deep parameter vectors appeared less sensitive than shallow ones to small parameter-value changes (Figure 7) and parameter vectors at equal depth exhibited a similar degree of sensitivity (Table 3).

4.2.3. Process Representation

The process representation was analyzed for the Brue and Paso La Ceiba catchments where continuous measurements of discharge were available.

The reliability of the simulations varied between 73% and 83% as an aggregated measure, but could span a wider range for specific hydrograph parts (Figure 9). In particular, both performance measures led to parameter vectors for which the model did a poor job in reproducing low flows (Figure 10). This was related to the fixed evaporation parameter. The low-flow simulations were poorer with a fixed parameter than with a varying one for both catchments and objective functions (Figure 9). However, a poorer model performance in the end of the dry season was also observed as a probable model-structural error in a previous study when using a variable evaporation parameter [Westerberg et al., 2011b]. For other parts of the hydrograph (large peaks, small peaks, troughs, rising and falling limbs) the difference between the fixed and the varying evaporation parameter was small. The performance of WASMOD was better for Paso La Ceiba than for Brue and the other Honduran catchments. It is reasonable to believe that the three-parameter model was too simplistic at Brue and that the much larger precipitation and discharge uncertainties for the small Honduran catchments.
catchments compared to Paso La Ceiba caused this difference.

The deviations (measured by the scaled scores) in general decreased with depth for the same parts of the hydrographs (large peaks, small peaks, and falling limbs) for both performance measures, sometimes even under split-sample testing (Figure 11). For other parts of the hydrograph (rising limbs, troughs and base flows), there was no clear decrease in the deviations or instead worse results with depth. This may have been a result of the fixed evaporation parameter. It was therefore difficult to draw conclusions about how the hydrological robustness varies with depth for low flows.

4.2.4. Transferability in Time

Deep parameter vectors transferred better than shallow ones with NSE as performance measure, but not for FDC, except in one case (Figure 12). In two of the tests with FDC as performance measure, the average performance increased with depth initially but ended up decreasing. At Paso La Ceiba, the comparatively poorer results obtained with the FDC performance measure compared to NSE might have been related to generally lower discharge during the 1980–1988 period in combination with the fixed evaporation parameter. Calibrating/evaluating during periods with different characteristics when the low-flow behavior of the model was constrained might explain the drop in performance with depth during evaluation. The FDC performance measure requires adequate simulation of the whole flow range (which NSE does not), which would lead to lower performance if the low-flow parts of the FDCs for the two periods have little overlap.

The different parts of the hydrograph where the deviations increased/decreased with depth were mostly similar between calibration/evaluation periods for both performance measures, although in a few cases there was a decrease in deviations with depth (i.e., better results) in calibration but not in evaluation.

4.3. Delimitation of the Behavioral Parameter Space

Alpha shapes constrained the behavioral parameter space better than the corresponding convex hulls (Figure 13). The reduction in volume between the space delimited by a BAS could be as low as 29% of the volume of the corresponding convex hull. The average volume of the α shapes was around 83% of the convex-hull volume for NSE and 74% for FDC (Figure 14).

5. Discussion and Conclusions

This study supported results from previous studies that hydrological robustness of parameter vectors appears to increase with depth in the behavioral parameter space. Deep parameter vectors for six Honduran and one UK catchment were found to be more robust than shallow parameter vectors with regards to both average model performance and sensitivity of the performance to changes in parameter vectors. The increase in performance with depth, as reflected by the numerical value of the two tested performance measures was undeniably clear, but the posterior performance analysis showed that simulated flows could actually become worse or show little change with depth for some parts of the hydrograph. This was especially the case for low flows, where the conclusions that could be drawn in this study were affected by the fixed evaporation parameter.

To the best of our knowledge only two depth functions have previously been used for parameter estimation in hydrology, (a) the Mahalanobis distance that makes strong assumptions regarding the data distribution, and (b) the half-space depth. Both are convex depth measures that assume unimodality. Here we have introduced a third depth function (Figure 7).

![Figure 7](image-url)

**Figure 7.** Change in performance with depth (n) for calibration using all available data, \( \Delta_{\text{per}}(n) = \frac{\text{per}(\text{BAS}_{n+1}) - \text{per}(\text{BAS}_{n})}{\text{per}(\text{BAS}_{n})} \) where per(\text{BAS}_{n}) is the average performance of the model evaluated on the vertices of \( \text{BAS}_{n} \). The black lines show a linear fit with the slope significantly (95% confidence interval) different from zero (continuous line) or not (dash-dot line). The horizontal gray lines are \( \Delta_{\text{per}}(n) = 0 \).

Table 3. Change in Average Efficiency From Vertices to Circumcenters of the Triangles Forming BAS\(_n\), the Approximate Boundary of the Behavioral Space

<table>
<thead>
<tr>
<th>Station</th>
<th>NSE</th>
<th>FDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paso La Ceiba</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>Concepción</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Guacericque</td>
<td>0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>Rio Del Hombre</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Quiebramontes</td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>Tatumbla</td>
<td>0.03</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Figure 8. Change in average performance near BAS₁ for calibration using all available data. The median length of the edges of the Delaunay triangulation of the Monte Carlo parameter vectors was taken as the unit distance $d$. $V$ was defined as the set of positive normal vectors to the facets of BAS₁ with norm $d$ whose tails are at the circumcenter of the facets. The average performance was evaluated by scaling $V$ and calculating the average performance of the heads of the vectors. The scaling factor multiplier is given in the logarithmic y-axis.

Figure 9. Reliability and precision for the total time series and for different parts of the hydrograph. The different hydrograph aspects are defined in Westerberg et al. [2011b].
measure, which is not affine invariant, but makes no assumptions about the underlying distribution and is better suited to describe the shape of the behavioral parameter space. Affine invariance means that the results are independent of the underlying coordinate system under affine transforms of space, i.e., those preserving linearity.

[81] The study was limited by computational restraints since only a three-parameter model could be used, and the evaporation parameter had to be fixed. This affected model simulations and large deviations were sometimes found between observations and simulations, especially for the Brue catchment that has a more complex rainfall-runoff relationship than the Honduran catchments. Better simulation results for low flows were found in a previous study [Westerberg et al., 2011b] for the Paso La Ceiba catchment using the same model with an unconstrained evaporation parameter, whereas there was less difference for peak flows. The possible model-structural errors imposed by this restriction were therefore significant and hindered conclusions to be drawn about hydrological robustness for low flows.

[82] The temporal transferability was lower when calibrating/evaluating for periods with differing flow characteristics. This was particularly the case when using the FDC performance measure, which, in contrast to the NSE, constrains the whole flow range. The evaporation parameter was fixed to 0.7, likely impacting the performance more negatively when evaluating the model with the FDC criterion than with NSE since the evaporation parameter primarily affected low-flow simulations.

[83] No multimodality was detected within any behavioral parameter space. However, multimodality could be an issue in other cases where different performance measures are used. Further studies should be undertaken to see if multimodality could be an issue for nonparsimonious and complex models with many parameters and/or a markedly nonlinear model structure.

[84] The spatial structure of the behavioral parameter space was shown to be compact and the volume efficiently delimited by the shapes algorithm. Volume reduction was significant between the shapes and the convex-hull methods. The potential improvement in computational efficiency from this volume reduction depends on the benefits of sampling from a smaller space, i.e., having to perform less simulations to attain the same sampling density or, alternatively, sampling the space more densely with the same number of simulations. However, computing an \( \alpha \) shape is more demanding than computing a convex hull, and sampling inside an \( \alpha \) shape also requires more

Figure 10. (top) Observed and simulated discharge at Paso La Ceiba in January 1983 to February 1984. The uncertainty in the observed daily discharge is shown in gray shading. The lower and upper limits for the simulated bounds for the NSE and FDC performance measures (black and blue lines) are the 5% and 95% percentiles of the likelihood-weighted distribution of the simulated discharge for each day for all behavioral parameter vectors. The mean monthly observed discharge (gray line, bottom left, and zoomed for the dry season, bottom right) is the average of the best-estimate daily observed discharge for 1980–1997. The monthly simulated discharge (black and blue lines, bottom plots left and right) is the averaged median discharge from all behavioral simulations.
operations than sampling inside a convex hull. Therefore, the computational advantage will depend on the advantages of sampling inside a smaller space compared to the additional costs. Computation of data depth is increasingly demanding when the parameter-space dimension increases and only brute-force algorithms are presently available for many cases. The operational complexity of geometric methods generally increases faster than exponentially with the number of dimensions. We believe that this will be a limiting factor to the practical application of methods like the shapes and are currently testing a computationally lighter depth function, based on convex hulls, that allows to increase either the dimensionality or the number of sample points.

Besides computational considerations, other factors should be taken into account when selecting a depth function:

1. The ability of the depth function to detect multimodality but not to be affine invariant. Zuo and Serfling [2000] suggest that a choice should be made, regarding modality, before selecting a depth function. If multimodality is not expected then functions such as the half-space depth have added inferential power. The ability to detect multimodality comes at the cost of not fulfilling all conditions desirable in a statistical depth function. Furthermore, the detection of multimodality as disconnected components in the \( \alpha \) shape does not necessarily take into account the performance within disconnected regions and should therefore not be viewed as a method to detect local optima.

2. Not being affine invariant also means that different results might be obtained if the underlying coordinate system is changed, e.g., by a different sampling range for MC simulations.

3. The results presented here were limited to three dimensions, but can be extended to higher dimensional settings. Alpha shape codes are freely available (CGAL-3.9, 2012), up to three dimensions. Alpha shapes are subsets of Delaunay triangulations, for which high-dimensional codes have been developed, which could facilitate further developments of the concept. Half-space depth and Mahalanobis depth algorithms are also freely available and can be found as, e.g., modules to the R language.

BAS1 was only an approximation of the behavioral boundary. It depended on the density of the sampling and the coordinate-system scale. In order to get more consistent estimates, a tactic other than random sampling should be applied. For instance, the parameter space could be evenly sampled through, e.g., a tessellation (a partition of the sampling space in a countable number of closed sets), and the
density of the sampling could be increased near the iteratively refined boundary in order to get a better delimitation of the behavioral space. Deep parameters could then be found through a geometric study of the resulting structure, instead of the random-sampling approach used here. [91] This paper has elucidated the link between data depth in the behavioral space and hydrological model robustness. BS08 have previously demonstrated the existence of this link. The reason behind it remains unclear. Deep parameters are robust in a geometrical sense: the median, which is the deepest point in a given space, is a robust estimator of location. Deep parameter vectors, i.e., close to the median, have again appeared to be more hydrologically robust than shallow parameter vectors. Elucidating the reason for this should be the subject of further research. Since the functions in our model are monotonic, we speculate that parameter vectors on the edge of the behavioral space relate to extremes while deep parameters might provide a more balanced fit.

[92] Acknowledgments. This work was part of the projects Research cooperation in hydrology and geotechnology—subproject Hydrology and water management of the upper Choluteca River basin and Effects of climate change and extreme weather on water resources sustainability in Central America funded by Sida (Swedish International Development Cooperation Agency) under grants 75007349 and SWE-2005-296. The computationally demanding calculations were performed at Uppmax, Uppsala Multidisciplinary Centre for Advanced Computational Science under project p2008030 (flyttad mening). We are grateful to SANAA (Servicio Autónomo Nacional de Acueductos y Alcantarillados) and SERNA (Secretaría de Recursos Naturales) for access to the Honduran data and to Philip Younger for providing the areal precipitation data for the Brue catchment, as well as the figure describing the catchment. Professors Hoshin Gupta and Alberto Montanari and five anonymous reviewers helped to improve the quality of this article. The Editorial Office speeded up publication with their prompt responses.

References


