Referee one

General comments

The paper presented shows an interesting approach to predict mean monthly flows in data scarce areas. However, the author does not clearly indicate what data was used for the analyses (ie how data scarce is the area and against what data is the model evaluated). The two approaches as mentioned several times should be explained better or referenced to (ie the objective of using each approach and how they complement each other). It is also unclear how the approach can be utilized in different ungauged regions, as in the Semliki it seems that all information is present.

The paper has potential but requires substantial revision, in particular the structure of the paper is lacking. 1) explain how this study is different from previous studies, 2)clear presentation of the methodology, 3) clear presentation of the results (including discussion), 4) conclusion

Specific comments

Introduction

-it is unclear what data set the authors used for the analyses, in the Semliki, where all

the 21 sub-catchments gauged? Page 3603 line 2-4 is not sufficient.

Very limited data are available in the Semliki watershed. Table 1 below presents the landscape attributes, their sources and software that were used to generate data for each of the 21 subcatchments that form the Semliki watershed. These datasets were subsequently used in the regression-based development of models.

The historical monthly flow measurements at the outlet of the Semliki watershed for 1950-1978 were used for the calibration of both the linear and the non- linear models. For each subcatchment, an historical 28-years monthly mean volume was computed proportionally to the subcathment area and was labeled as "control". This approach is supported by the fact that in humid basins like the Semliki as opposed to arid and semi-arid regions, "stream flows increase in the downstream direction, and the spatial distribution of average monthly or seasonal rainfall is more or less the same from one part of the river basin to another, hence the runoff per unit land area is assumed constant over space. In these situations, estimated flows are usually based on the watershed areas, as contributing flow to those sites, and the corresponding streamflows and watershed areas above the nearest or most representative gauge sites" (Loucks et al., 1981 and Loucks and Van Beek, 2005).

These amendments will be included in the revised version of the manuscript.

Table 1

Description	Source	Observation
Landform	SOTERCAF	GIS processing (WINDISP, ARC View3.3, Excel 07-10, Statistica 8.0)
Lithology	SOTERCAF	GIS processing (WINDISP, ARC View3.3, Excel 07-10, Statistica 8.0)
Soils	SOTERCAF	GIS processing (WINDISP, ARC View 3.3, Excel 07-10, Statistica 8.0)
Drainage Density (Dd)	SRTM 90m-DEM, SWAT pre-processor	Subcatchments areas generated from SWAT preprocessor (WINDISP, ARC View3.3, Excel 07-10, Statistica 8.0, SWAT)
Stream Length	SRTRM 90m-DEM, topographical map (1/50,000), SWAT pre-processor	Generated from the SWAT preprocessor and cross validation with traditional map (WINDISP, ARC View3.3, Excel 07-10, Statistica 8.0, SWAT)
Stream slope	SRTM 90m DEM	Remotely sensed acquired (ARC View 3.3, Excel 07-10, Statistica 8.0)
Rainfall	FEWS NOAA / RFE (2001-2007) rain gauge at Beni (1973 to 2008)	Remotely sensed acquired and Locally corrected and calibrated (WINDISP, ARC view3.3, Excel 07-10, Statistica 8.0)
Elevations: Minimum Maximum Area-weighted average	SRTM 90m-DEM, topographical map (1/50,000)	Remotely sensed acquired and cross validation with traditional map. (Arc View 3.3, Excel 07-10, Statistica 8.0)
NDVI	NOAA-AVHRR (1982- 2008)	Remotely sensed acquired and correlated with rainfall. (WINDISP, ARC View3.3, excel 07-10, Statistica 8.0)

-how is this approach different/ better than the ones referred to on page 3601?

No such study (Spatial and temporal resolution) has been reported to date in the Semliki watershed. None of the publications referred to on page 3601 were conducted in equatorial and humid regions, the lack of data in the study area required a rational and creative approach in order to link the landscape attributes to the flows generated within the watershed and that was attempted in this work.

Materials and methods

-section is too general, I would have expected a more elaborate explanation about the

PCA and IGA or a reference to the methodology and equations used.

The PCA is a pretty well established technique as a result we did not find fitting to further elaborate this section however the inclusion of a reference is appropriate and will be done.

-" the PCA as Indirect Gradient Analysis in association with clustering analysis wasused: : : Both approaches " What approaches are referred to? **PCA and tree clustering**

Line 8-10 and 11-15 are unclear

Noted, it will be rephrased

Results and discussion

Section 4.1

-Page 3603, line 18-21 is not clear

Not sure I understand the referee's observation; the mean and standard deviation values are quite explicit in Table 2.

-Page 3603, line 22-26 is not clear, should this be part of the methodology?

Noted and will be attended to

-Page 3604, line 1, what is phi and how is it calculated?

The Glaeson – Staelin redundancy statistic (phi) measures the level of interrelation between a group of variables. A zero value of phi indicates zero correlation between the variables whereas a value of one indicates perfect correlation between the variables. The value of phi for this data (0.4), suggests that there is considerable redundancy or complexity in the group of variables which warrants further examination using PCA.

Same for Bartlett's sphericity test, should this be part of methodology?

Bartlett's sphericity test is used to test the null hypothesis that the correlation matrix of the group of variables is a zero identity matrix i.e. none of the variables are correlated. If we obtain a p-value for the Bartlett's test which is greater than 0.05 we should not carry out PCA. The p-value obtained is very low indicating that we can carry out the PCA.

-Page 3604, line 8 What is Kaiser criterion, how are the "principal" components identified

According to the Kaiser criterion when the principal components have been calculated using correlation coefficients is to retain the principal components with an eigenvalue greater than 1. Therefore in this case we would retain the first 3 principal components. These 3 principal components account for 76% of the variation in the data.

-It is unclear, what are the factors, factor loadings, eigenvectors and eigenvalues.

The Eigenvectors are the coefficients that relate the scaled original variables to the derived factors. The scaled original variables are defined as follows:

$$x_i = \frac{X_i - \mu_i}{\sigma_i}$$ **where;**

x_i **= the scaled variable,**

X_i **= the original variable**

μ_i **= the mean of the original variable**

σ_i **= the standard deviation of the original variable**

For instance, the first principal component is:

Factor1 = -0.157886*(Strm_len)* -0.283358*(Drainage)* + 0.241135*(avg_slope)* + 0.468021*(Max_elev)* + 0.268251*(Min_elev)*
+ 0.505941*(avg_elev)* - 0.339627*(monthly__prec)* - 0.417503*(monthly_NDVI)*

The italics are there to emphasise that we are referring to the scaled original variables and not their original values.
Inspection of the eigenvectors shows that the first factor is a contrast of avg_slope, max_elev, min_elev and avg_elev to strm_len, drainage, monthly__prec and monthly_NDVI. This factor explains 44% of the variation in the data. The eigenvectors of the three factors that have been retained are shown in Table 6.

The factor loadings are the correlations between the variables and the factors. For instance, Factor 1 is most highly correlated to the maximum elevation and the average elevation; whereas factor 3 is most highly correlated to the average slope and the minimum elevation (see table 7 of the manuscript).

-Can the tables 5, 6 & 7 be combined?

No, May be Table 6 &7 (columns number to double).

Section 4.2

-Page 3606 line 5 " several normality test were performed: : :" should be explained in the methodology section

Noted and will be corrected

-Page 3606 line 8-9 should be part of the methodology section There is a section on discussion of results missing, the authors only show the results for February and they do not discuss any further possible application of the approach.

Improvement of this section will be undertaken.

Conclusions

The conclusions are very general, and should be more specific. Showing the results of the model, the possible application in other areas (how will the approach work when there is no data available and the area is dissimilar to the Semliki catchment) and the relevance of the output of the model (mean monthly runoff) for management purposes.

Noted and appropriate changes will be made